UNIVERSITAT ROVIRA I VIRGILI

DESENVOLUPAMENT D'UNA EINA PER AL TRACTAMENT I ANÀLISIS DE DADES GENERADES EN EL CAMP BIOTECNOLÒGIC.

Jordi Llatser Torres

TREBALL DE FI DE GRAU DE BIOTECNOLOGIA

Tutor acadèmic: Santi Garcia, grau biotecnologia, departament de bioquímica i biotecnologia, santi.gracia-vallve@urv.cat
En cooperació amb: Dr Carlos Alonso-Villaverde, el grup de recerca de malalties cròniques i d'envelliment de la Xarxa Tecla

Índex

Índex de figures

1 Resum i Paraules Clau

Aquest treball consisteix en desenvolupar un programa informàtic que té com a finalitat ser utilitzat per el tractament i anàlisis de bases de dades, facilitant eines per la visualització de les dades. Per aconseguir aquest resultat, s'utilitza el llenguatge de programació R, un llenguatge altament utilitzat per a l'anàlisi de dades. Alguns dels mètodes d'anàlisis que s'han implementat són correlacions, anàlisis de components principals i models de regressió entre altres. Tot això complementat amb gràfics per visualitzar les dades i facilitar el seu anàlisi i interpretació. Finalment s'ha posat a prova l'eina en una base de dades real per comprovar la seva utilitat en l'exploració de les dades i en la visualització de resultats.

Biotecnologia, Bioinformàtica, Anàlisis de dades, Estadística

2 Introducció

La biotecnologia s'ha convertit en un pilar fonamental de la nostra societat contemporània, revolucionant nombrosos sectors des de la salut i la medicina fins a l'alimentació i protecció del medi ambient. Aquesta disciplina, que és una combinació de la biologia i la tecnologia, ha provocat un desenvolupament exponencial en la generació de coneixement i solucions innovadores.

A mesura que la investigació biotecnològica s'expandeix, es generen volums massius de dades provinents d'experiments, seqüenciació genòmica, anàlisis proteòmics i altres tècniques. Aquestes dades, que contenen informació valuosa sobre la estructura i funció de biomolècules, patrons genètics, característiques de malalties i moltes més, representen un tresor científic que necessita ser processat i analitzat de manera adequada.

En aquest mon, és essencial tenir eines eficients pel processament de les dades generades en la investigació biotecnològica. El processament adequat de les dades biotecnològiques no només accelera el ritme dels descobriments científics, sinó que també ajuda a millorar la qualitat dels resultats i a optimitzar la presa de decisions informades.

Aquest TFG està basat en els resultats de les pràctiques externes fetes en la Xarxa Tecla sota supervisió del Dr. Carlos Alonso. Està fet basat en les necessitats i peticions del grup de recerca dirigit per Dr. Carlos Alonso, per ajudar i facilitar les seves investigacions.

En aquest treball final de grau desenvoluparem una eina adequada pel processament i anàlisis de dades en la investigació biotecnològica, intentant innovar i millorar les eines actuals, i donar nous punts de vista en la visualització de dades.

Mentre que actualment hi ha moltes eines disponibles per l'anàlisi de dades, i moltes tenen molts beneficis, també tenen els seus desavantatges. Ara presentaré les eines més utilitzades pel grup de recerca, i la opinió d'aquestes.

La primera eina és BioConductor (Agrawal et al., 2010), accessible a https://www.bioconductor.org/. Aquesta és un software de codi obert basat en l'anàlisi i comprensió de dades genòmiques. Està basat en el llenguatge de

programació R (Team, 2021) i aporta un munt de llibreries per ajudar al desenvolupament de varies tasques d'anàlisis biològics, des de preprocessament fins a visualització de dades, i es pot aplicar a moltes àrees de la biologia, com a genòmica, transcriptòmica, proteòmica i metabolòmica. Els objectius des de la creació d'aquest projecte ha estat la creació d'un entorn durador i flexible pel desenvolupament que pugui respondre a nous reptes conceptuals, computacions i reptes derivats d'aquests.

Un dels principals avantatges de BioConductor és la seva flexibilitat i la habilitat d'integrar diferents tipus de dades biològiques, permetent als investigadors a enfrontar-se a preguntes complexes. També aporta eines per visualitzar, explorar i interpretar dades.

Aquesta eina s'ha utilitzat molt per la comunitat científica, i molts estudis han demostrat la seva eficàcia en l'anàlisi de dades genòmiques. Un exemple seria l'estudi realitzat per Lun et al., (2016) on s'utilitza BioConductor per analitzar dades de seqüenciació de ARN unicel·lular, permetent identificar nous tipus de cèl·lules i patrons d'expressió gènica. Un altra exemple seria l'estudi realitzat per Huber et al., (2015) on utilitzen BioConductor per realitzar anàlisis genètics d'alt rendiment. Ambdós articles expliquen diferents funcionalitats, i expliquen la preparació de les dades, el seu tractament i visualització parlant de les llibreries més importants pel seu ús en els casos concrets, així com també demostren exemples pràctics del seu ús en escenaris reals.

Com es pot veure en aquests articles, aquesta és una eina molt útil i flexible, però també presenta unes limitacions força importants. Un dels seus principals reptes és l'alta corba d'aprenentatge. BioConductor no és una eina definitiva, sinó un conjunt de llibreries d'R que necessiten ser utilitzades conjuntament programant codi per obtenir el resultat. Si els investigadors no tenen coneixements en el mon de la programació, aquesta tasca acostuma a ser molt dura i quasi impossible, fent que perdi atractiu per certs investigadors.

Una altra eina molt utilitzada és la pàgina web MetaboAnalyst (Pang et al., 2021), una eina per l'anàlisi i interpretació de dades metabòliques. MetaboAnalyst ofereix una interfície *user-friendly* que combina diversos mètodes d'anàlisis de dades,

visualització de dades i l'anàlisi de rutes metabòliques per ajudar als investigadors a extreure resultats importants dels seus experiments.

Els usos principals de MetaboAnalyst són:

- Preprocessament de dades: Permet als usuaris normalitzar, escalar, transformar les dades
- Anàlisis estadístic: MetaboAnalyst ofereix un conjunt d'eines estadístics que inclouen anàlisis univariable i multivariable, classificació, clústering, i anàlisis de regressió.
- Integració de bases de dades biològiques: MetaboAnalyst integra múltiples bases de dades biològiques com per exemple KEGG: Kyoto Encyclopedia of Genes and Genomes(Kanehisa et al., 2017), per donar més context a la informació dels anàlisis de rutes metabòliques.
- Viabilitat i Accessibilitat: Al ser una plataforma web accessible a https://www.metaboanalyst.ca/ elimina la necessitat de ser instal·lat en un ordinador, i és accessible gratuïtament per tota la comunitat d'investigadors.

I mentre aquests poden semblar bons avantatges, també té certs problemes a causa de la seva naturalesa, com serien la necessitat de connexió a internet per poder-ne fer un ús, fet que limita la seva accessibilitat. La falta de privacitat i seguretat de les dades és un altre problema, ja que per poder treballar necessites penjar les dades a la pàgina, i per tan a internet.

Intentar remeiar aquests problemes i aconseguir una eina pràctica i fàcil d'utilitzar així com d'instal·lar és una de les motivacions d'aquest treball.

Al basar-se en aquestes eines, s'aplicaran tècniques d'anàlisi i tractament de dades explicats a continuació:

Scatter-Plot (Wilke, 2019): El scatter plot és una eina valuosa per a l'anàlisi exploratòria de dades i per a la identificació de possibles correlacions o tendències entre les variables. Permet observar la distribució conjunta dels valors de les dues variables i avaluar la seva relació.

Per fer un scatter-plot es crea un pla de coordenades on cada punt representa una observació en el conjunt de dades, amb una variable representada a l'eix horitzontal i l'altra a l'eix vertical.

L'anàlisi de scatter plots permet identificar valors atípics (outliers), agrupaments o agrupacions de punts, i proporciona una visió visual de la distribució conjunta de les variables. També pot ser útil per a la selecció de variables per a models predictius, identificant aquelles que mostren una relació significativa amb la variable objectiu.

Heatmaps (Wilke, 2019): Un heatmap és una representació gràfica utilitzada per visualitzar dades en forma de matriu, representant els valors com colors en una graella. Cada casella correspon a la combinació de dos variables, una en cada eix, i la casella està pintada en un color que representa el valor numèric o la intensitat entre les escales.

Els valors representats ajuden per la visualització de patrons, agrupacions o correlacions entre variables. Una de les dades més representades en aquestes matrius és la correlació de variables, que també es pot veure en l'scatter-plot. El fet de veure-ho conjuntament permet destacar quines variables estan més relacionades amb altres, permetent també veure si una variable es correlaciona més amb altres que la resta.

Anàlisis de Components Principals (PCA) (Bro & Smilde, 2014): Aquesta és una tècnica d'anàlisi estadística que s'utilitza per reduir la dimensionalitat d'un conjunt de dades amb moltes variables, mentre es manté la major part de la seva informació. L'objectiu principal del PCA és trobar les combinacions lineals de les variables originals que expliquin la major variabilitat en les dades. Aquesta és una tècnica important en l'anàlisi de dades multivariable.

PCA té molts propòsits, incloent la visualització de dades, extracció de característiques i compressió de dades. Representar les dades en menys dimensions facilita la visualització i interpretació de bases de dades complexes. Aquesta tècnica d'anàlisis s'utilitza en anàlisis exploratori de dades i també per a construir models predictius.

Models de regressió linear (Kumari & Yadav, 2018): Un model de regressió lineal és una tècnica d'anàlisis que ens permet examinar la relació entre una variable dependent (la variable que volem predir o explicar) i una o més variables independents (les variables que utilitzem per predir o explicar la variable dependent). Això permet explicar com influencien diferents variables com podrien

ser la concentració de ferro, la expressió d'un gen o la presa d'un medicament en un resultat específic.

El resultat d'un model de regressió linear acostuma a ser una fórmula on cada variable independent està multiplicada a un coeficient, i utilitzant aquesta fórmula es pot predir la variable dependent a partir de dades noves.

Corba ROC (Nahm, 2022): És una tècnica d'avaluació de la capacitat predictiva i discriminativa d'un model o classificador en problemes de classificació. La corba ROC proporciona una representació visual de com varia el rendiment del model en funció del llindar de classificació utilitzat per diferenciar les classes.

Per fer un anàlisis de ROC curve es seleccionen variables independents que seran utilitzades per construir un model, i una variable dependent on hi ha la classe que es determinarà. És semblant a un model linear, però aquest serveix per l'anàlisi de variables no contínues.

En una anàlisi de corba ROC, les classes solen ser etiquetades com a "positiu" i "negatiu". Per exemple, en un estudi mèdic, els pacients amb una malaltia serien considerats com a "positius" i els pacients sans com a "negatius". El model o classificador proporciona una puntuació o probabilitat de pertànyer a la classe "positiu" per a cada mostra. La gràfica que ensenya aquesta probabilitat es fa comparant la taxa de positius verdaders (TPR) i la taxa de falsos positius (FPR) a diferents llindars, i això serveix per ajudar a triar el millor llindar.

Aquestes dos últimes tècniques serveixen especialment bé per predir futurs resultats, tant de valors continus utilitzant un model linear o classificar amb malalt o sa utilitzant una ROC curve, sempre que s'utilitzin variables independents suficient rellevants, per això és necessari fer un anàlisis previ utilitzant els altres tipus d'anàlisis.

Per poder fer un anàlisis de qualitat, és necessari aplicar un pretractament a les dades originals. Això és necessari per vàries raons:

1. Neteja de les dades: Les dades reals sovint contenen errors, valors buits o inconsistències. El pretractament de les dades implica la identificació i la correcció d'aquests problemes per assegurar la qualitat i la integritat de les dades abans de l'anàlisi. Això implica eliminar o substituir valors buits, corregir errors de registre i eliminar dades duplicades.

2. Normalització i estandarditzación: Les variables en les dades poden tenir diferents rangs, unitats o distribucions. Mitjançant la normalització i l'estandarditzación de les variables, podem convertir-les a una escala comuna per permetre comparacions significatives. Això assegura que les variables no esbiaixades o amb rangs molt diferents no dominin l'anàlisi.
3. Gestió de valors atípics: Els valors atípics són observacions inusuals o extrems que poden distorsionar els resultats de l'anàlisi. El pretractament de les dades implica identificar i gestionar adequadament aquests valors atípics, ja sigui eliminant-los, substituint-los o aplicant tècniques estadístiques per a la seva mitigació.
4. Resolució de problemes d'inconsistència: En algunes ocasions, les dades poden contenir errors o inconsistències que requereixen un tractament específic. Això pot incloure la reconciliació de dades contradictòries, la gestió de valors perduts o la resolució de conflictes en casos de dades duplicades.

Per a poder reduir i evitar, s'apliquen les següents tècniques de preprocessament, explicades a continuació:

Sample Normalization (Wu & Li, 2016): la normalització de mostres és un procés a vegades oblidat que consisteix en ajustar els senyals adquirits per equalitzar els senyals totals. És un ajust general per a diferències entre mostres, que permet realitzar una quantificació exacta i precisa de la concentració de metabòlits individuals. Aquest pot existir per la manca de precisió de la mostra total, i amb aquest pretractament es pot reduir o eliminar la variació creada per aquesta causa.

Data Transformation (S, 2010):La transformació de dades és un procés utilitzat en l'anàlisi de dades per modificar o ajustar la distribució dels valors d'una variable o conjunt de dades. Aquesta transformació pot ser necessària per satisfer els supòsits d'un determinat model estadístic o per millorar la interpretació i interpretació dels resultats.

Data Scaling: La normalització de dades és un procés que s'utilitza per ajustar els valors de les variables en un rang específic. L'objectiu principal de la normalització de dades és garantir que totes les variables tinguin la mateixa escala i rang de valors, de manera que cap variable tingui un impacte desproporcionat en l'anàlisi o en els resultats. És útil en situacions en què les variables tenen unitats o rangs de

mesura diferents. Aquest procés assegura que totes les variables tinguin una influència equitativa en l'anàlisi i els resultats, independentment de les seves unitats o magnituds originals. També pot millorar la convergència dels algoritmes d'aprenentatge automàtic i facilitar la interpretació dels coeficients en els models de regressió.

Data Cleansing (Calabrese, 2018): La neteja de dades és un procés essencial en l'anàlisi de dades per identificar i corregir errors, inconsistències i anomalies en les dades recopilades o obtingudes. L'objectiu principal de la neteja de dades és garantir la qualitat i la fiabilitat de les dades abans de realitzar anàlisis o aplicar models. Aquesta neteja sol implicar el tractament dels valors que falten, com la eliminació de duplicats.

3 Objectius

L'objectiu principal d'aquest treball de fi de grau és desenvolupar una aplicació d'anàlisi de dades que permeti analitzar bases de dades estructurades, on cada columna representa una categoria i cada fila conté les mostres obtingudes durant la investigació. Aquesta aplicació serà utilitzada en l'àmbit de l'anàlisi proteòmica, transcriptòmica, metabolòmica i altres àmbits de recerca mèdica.

Per assolir aquest objectiu, el treball es planteja tres subobjectius clau:

1. Aprendre el llenguatge de programació R, un llenguatge dedicat a l'anàlisi de dades i un dels més utilitzats en aquest àmbit. Utilitzant aquesta eina es crearà el programa.
2. Dissenyar i programar l'aplicació, incorporant diferents tècniques d'anàlisi de dades i representacions visuals de la informació. Aquest objectiu implica adquirir coneixements sobre les metodologies i les eines més rellevants per a l'anàlisi de dades.
3. Aplicar l'aplicació en un cas d'estudi real per avaluar la seva utilitat i funcionalitat. En aquesta etapa, es posarà a prova l'aplicació amb dades reals i s'avaluarà la seva capacitat per aportar resultats significatius per als investigadors.

Mitjançant aquests subobjectius, aquest treball de fi de grau pretén desenvolupar una aplicació pràctica i eficaç per a l'anàlisi de dades estructurades, contribuint a millorar els processos de recerca en l'àmbit de la biotecnologia i altres àrees relacionades.

4 Metodologia i materials

Per fer aquest treball, s'ha decidit programar amb el llenguatge de programació R. El llenguatge de programació R s'ha consolidat com una eina fonamental en l'anàlisi de dades i la investigació científica. Desenvolupat específicament per a l'estadística i la visualització de dades, R ha guanyat popularitat en diferents camps, incloent-hi la biotecnologia, la medicina, l'economia, machine learning i altres àmbits científics.

R destaca per la seva flexibilitat i potència en el tractament de dades complexes. Amb una àmplia gamma de paquets i llibreries disponibles, permet als investigadors realitzar anàlisis estadístiques sofisticades, models predictius, visualització de dades i moltes altres tasques relacionades amb l'anàlisi de dades.

Un dels beneficis de R és la seva capacitat per a la visualització de dades. Amb l'ajuda de llibreries com ggplot2 i plotly, és possible crear gràfics i representacions visuals d'alta qualitat que permeten explorar i comunicar de manera efectiva les tendències i patrons ocults en les dades.

També té una gran comunitat activa, fet que proporciona una gran quantitat de recursos online, com tutorials i exemples de codi i fòrums útils en l'aprenentatge i resolució de dubtes.

4.1 Llibreries R

Per fer aquest treball s'han utilitzat diferents llibreries del llenguatge de programació R per tal de facilitar i fer més eficient el programa. A continuació es comenten les llibreries utilitzades amb un petit resum del seu us general.

Shiny: Shiny és una llibreria de R que permet crear aplicacions web interactives. Amb Shiny, els usuaris poden visualitzar i explorar dades, realitzar anàlisis i manipular paràmetres en temps real, sense necessitat de coneixements de programació web. Aquesta llibreria s'ha utilitzat per facilitar el posterior us de la eina un cop desenvolupada.

Shinyjs: Shinyjs és una extensió de Shiny que permet controlar i interactuar amb els elements d'una aplicació Shiny mitjançant funcions JavaScript. Això permet afegir interactivitat i personalització avançada a les aplicacions Shiny.

Dplyr: Dplyr és una llibreria de manipulació de dades que ofereix un conjunt d'eines concises i eficients per filtrar, transformar i agrupar les dades en marcs de dades (data frames). Amb dplyr, es pot realitzar tasques com seleccionar columnes, filtrar dades, crear noves variables i resumir les dades de manera senzilla i intuïtiva. Essencial per poder tractar i manipular les dades.

Stats: Stats és una llibreria de R que proporciona una àmplia gamma de funcions estadístiques. Aquesta llibreria inclou des de les estadístiques bàsiques fins a mètodes més avançats, com proves d'hipòtesis, models lineals, anàlisi de variància (ANOVA), regressió i molts altres mètodes estadístics.

ggplot2: ggplot2 és una llibreria de visualització de dades que permet crear gràfics de gran qualitat i flexibilitat.

ggfortify: ggfortify és una extensió de ggplot2 que proporciona funcions per a la visualització de models estadístics. Amb ggfortify, es pot representar gràficament els resultats de models com ara regressions, anàlisis de components principals (PCA), agrupacions (clustering) i altres.

heatmaply: heatmaply és una llibreria per a la visualització de heatmaps o mapes de calor interactius.

Plotly: Plotly és una llibreria que permet crear gràfics interactius i dinàmics. Amb Plotly, es poden generar diversos tipus de gràfics, i també aporta eines per editar-los o unir múltiples gràfics en un sol.

DescTools: DescTools és una llibreria de R que proporciona una gran varietat de funcions útils per a l'anàlisi de dades, com ara càlcul de estadístiques descriptives, manipulació de dades, anàlisi de correlació, proves d'hipòtesis i moltes altres. D'aquí s'utilitzen dos específicament, HuberM i TukeyBiweight, utilitzats per donar estimadors utilitzats als gràfics de correlació.

car: car és una llibreria de R que proporciona funcions per a l'anàlisi de regressió i anàlisi de covariància. Aquesta llibreria ofereix eines per avaluar l'ajustament del model, realitzar anàlisi de residus, calcular intervals de confiança i realitzar proves d'hipòtesis específiques per a models de regressió. S'utilitza per fer anàlisis ANOVA

plotROC: plotROC és una llibreria de R que permet crear gràfics de la corba característica de funcionament del receptor (ROC). Amb plotROC, es pot visualitzar

i avaluar el rendiment de models de classificació, com ara l'àrea sota la corba ROC (AUC), punts d'operació òptims i línies de diagnòstic.

ROCR: ROCR és una llibreria de R que proporciona funcions per a l'avaluació del rendiment de models de classificació. Amb ROCR, es pot generar informació detallada sobre mètriques de classificació com la sensibilitat, especificitat, precisió i altres. Aquesta llibreria també permet crear gràfics ROC i calcular l'àrea sota la corba ROC (AUC). S'utilitza conjuntament amb la llibreria plotROC per poder visualitzar les corbes ROC

missForest: missForest és una llibreria de R per a la determinació de dades perdudes (missing data). Aquesta llibreria ofereix eines per estimar i omplir les dades perdudes a partir d'altres variables observades mitjançant mètodes de random forest.

ggplotify: ggplotify és una llibreria de R que permet convertir gràfics generats amb altres paquets, com ara base, lattice o ggplot2, en objectes ggplot. Això permet unificar la representació gràfica de les dades utilitzant les funcionalitats de ggplot2.

5 Resultats

Els resultats d'aquest treball són prometedors, ja que s'ha aconseguit desenvolupar amb èxit una eina d'anàlisi de dades avançada. En aquesta secció, es proporcionarà una explicació de l'ús d'aquesta eina mitjançant exemples de possibles aplicacions reals.

S'exploraran les diferents funcionalitats de la eina i s'ensenyarà la seva aplicació amb una base de dades de prova, per poder veure la navegació real, així com la experiència del seu us.

5.1 Pantalla inicial

Un cop s'executa l'aplicació, s'obre la pantalla principal (Fig. 1).

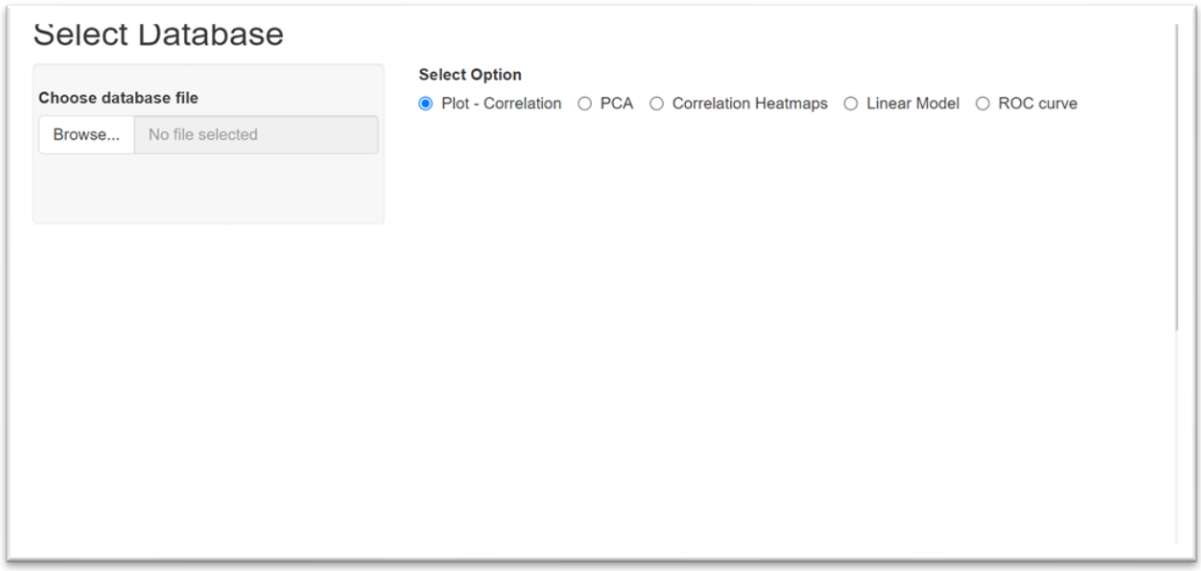

Fig. 1 Menu Principal

En aquest punt, es pot començar a observar l'estructura principal de l'aplicació. A la part superior esquerra, destaca la selecció de la base de dades, mentre que a la seva dreta es troba la principal selecció de categories per a l'anàlisi de dades. Aquesta disposició està feta per proporcionar una navegació intuïtiva i un fàcil accés a les funcionalitats clau de l'eina.

Cada categoria tindrà el seu apartat a la guia, però abans de poder interaccionar amb ells, és necessari fer la selecció de la base de dades.

5.2 Selecció i neteja de la base de dades

En la part superior esquerra hi ha la opció de penjar una base de dades en format .CSV, .XLS i .XLXS. Per fer això, es selecciona el botó *Browse...* i s'obre un menú de selecció de fitxers.

Un cop s'ha penjat el fitxer, sobre un pop-up per netejar la base de dades de valors buits (Fig. 2).

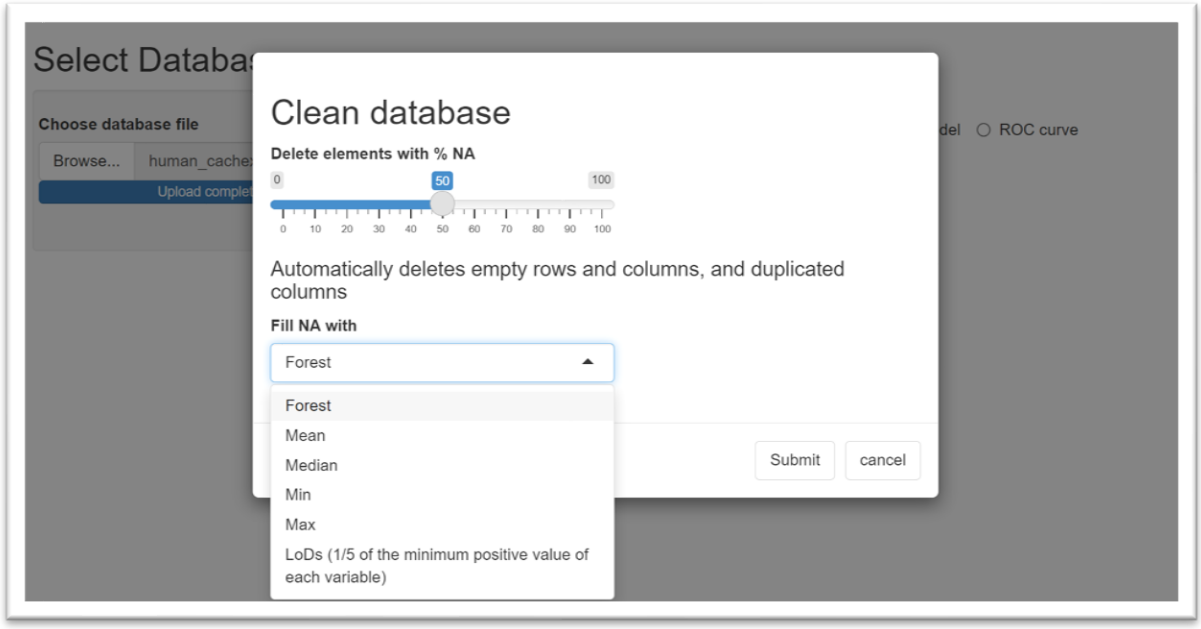

Fig. 2 Neteja de dades

Aquest pop-up ens dona diferents solucions pel primer pretractament essencial en l'anàlisi de dades. Un cop selecciones quins paràmetres vols i acceptes el pop-up, aquest fa 3 operacions:

1. Eliminar totes les columnes i files que estiguin duplicades.
2. Eliminar totes les files i columnes que tinguin un número de caselles buides igual o superior al percentatge indicat al deslliçador. Elimina sempre totes les files i columnes completament buides, ja que no tenen dades per treballar.
3. Reomplir la resta de files i columnes amb elements buits utilitzant una tècnica de la llista
    - Random Forest: És un algoritme que estudiant les altres dades existents genera valors aleatoris per omplir les caselles buides
    - Mean: Omple els valors restants utilitzant la mitjana

- Median: Omple els valors restants utilitzant la mediana
- Min: Omple els valors restants utilitzant el valor mínim
- Max: Omple els valors restants utilitzant el valor màxim
- LoDs: Omple els valors restants utilitzant 1/5 del mínim dels nombres positius

Després de fer aquest pretractament obligatori, es desbloquegen més opcions per continuar treballant.

5.3 Pretractament

En qualsevol de les categories d'anàlisis, sempre hi ha tres opcions compartides que s'explicaran prèviament aquí. Aquestes opcions estan relacionades amb el pretractament de les dades, i es poden veure en el menú de la esquerra (Fig. 3). S'ha decidit posar d'aquesta forma perquè els usuaris tinguin sempre a mà la opció de canviar quin pretractament utilitzen, sense necessitar de reiniciar tot el procés de l'anàlisi de dades que estiguin fent.

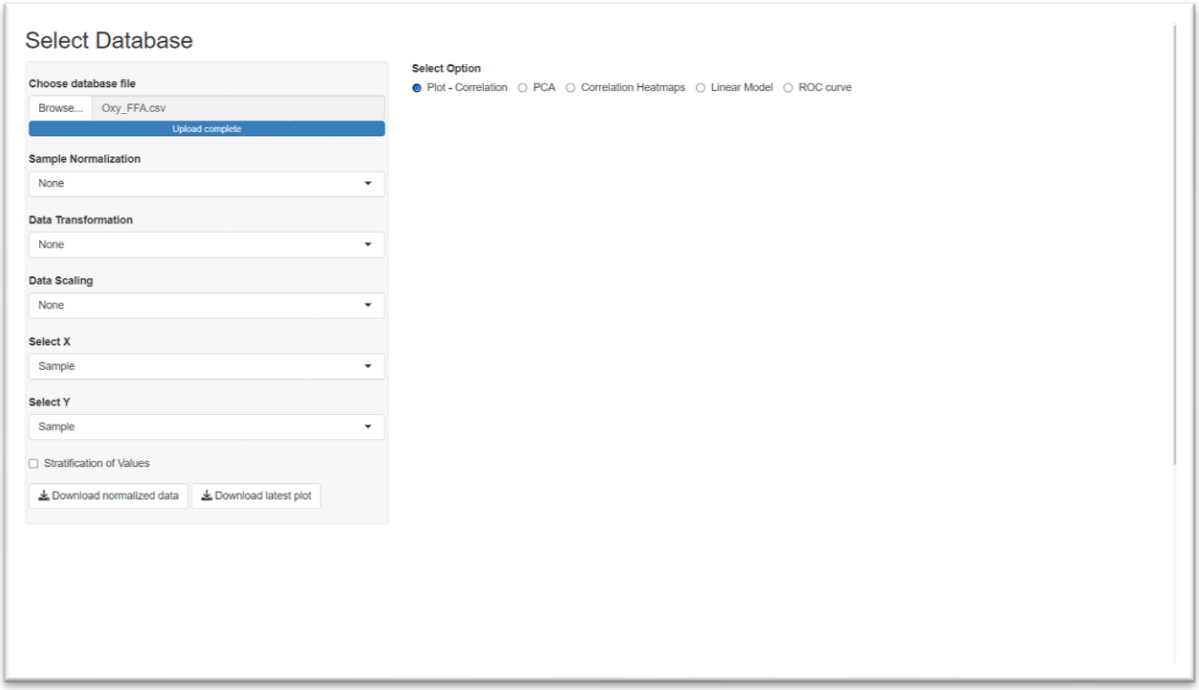

Fig. 3 Menú pretractament

És necessari realitzar aquest pretractament a les bases de dades per tal de poder eliminar diferències sistemàtiques, comparar els valors de manera justa i millorar l'estabilitat i la interpretació dels resultats, com s'ha explicat a la introducció.

Les tres opcions són les següents:

5.3.1 Sample Normalization

Les opcions disponibles per fer *Sample Normalization* disponibles són les següents (Fig. 4):

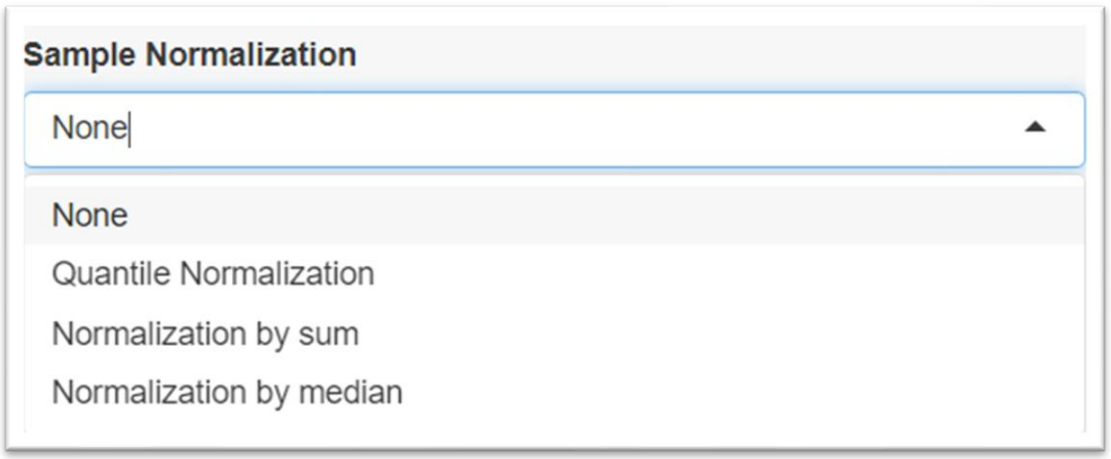

Fig. 4 Opcions Sample Normalization

- None: No aplicar cap normalització a les dades.
- Quantile Normalization: Aquesta tècnica iguala els quantils de les variables per assegurar una mateixa distribució.
- Normalization by sum: La normalització per suma és una tècnica que divideix les observacions de cada variable per la seva suma total, assegurant que la suma de les observacions normalitzades sigui constant i normalment 1.
- Normalization by median: La normalització per mediana és una tècnica semblant a la normalització per suma, però aquest cop divideix les mostres de cada variable per la seva mediana, assegurant que la mediana de les observacions sigui constant i normalment 1.

5.3.2 Data Transformation

Les opcions de *Data Transformation* són les següents (Fig. 5):

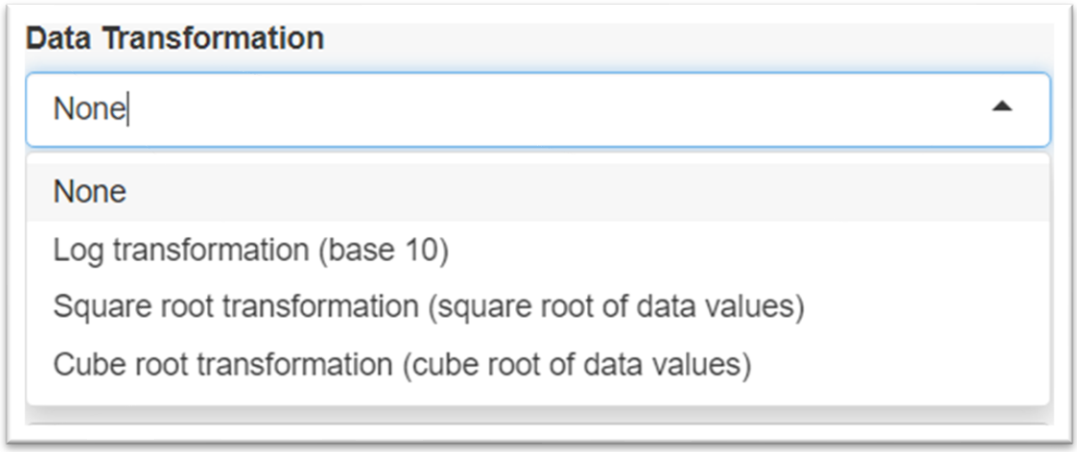

Fig. 5 Opcions Data Transformation

- None: No aplicar cap transformació a les dades.
- Log transformation: Aplicar el logaritme base 10 als valors
- Square root transformation: Aplicar l'arrel quadrada als valors
- Cube root transformation: Aplicar l'arrel cúbica als valors

Aquestes transformacions permeten reduir el nombre de outliers a causa de valors comparadament més grans, o més petits com es pot veure en la següent comparació (Fig. 6).

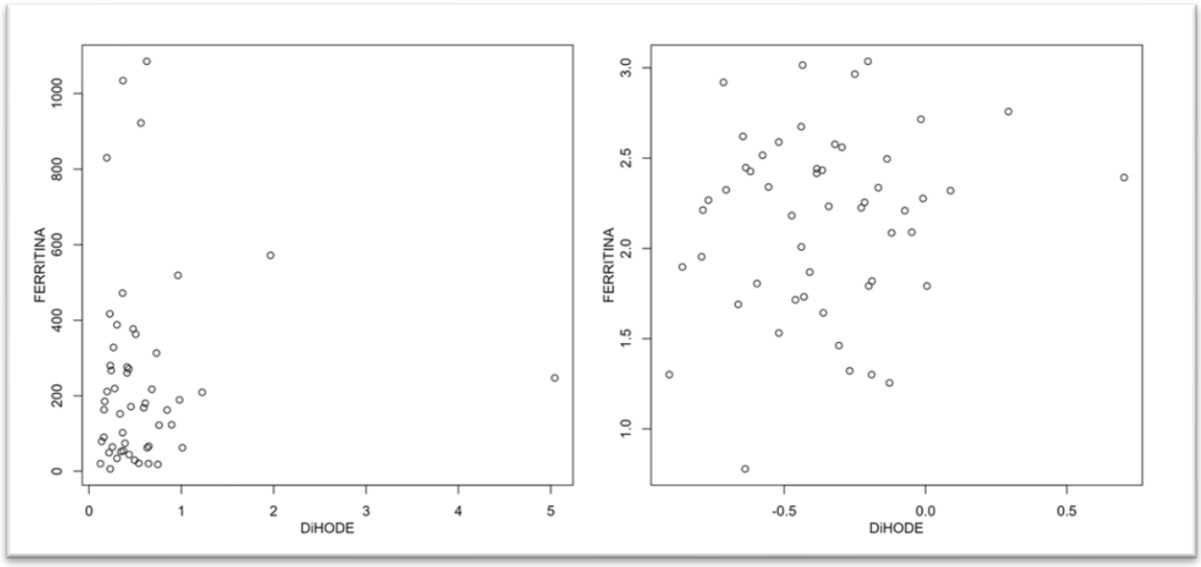

Fig. 6 Sense transformació vs amb log transformation

5.3.3 Data Scaling

L'escalat de dades és un procés en l'anàlisi de dades que consisteix a transformar les variables perquè tinguin una escala similar o estiguin en un rang específic.

Aquesta transformació permet comparar les variables i evitar que una variable amb valors més grans o amb una variança més gran dominin o distorsionin l'anàlisi.

Les opcions de data scaling són les següents:

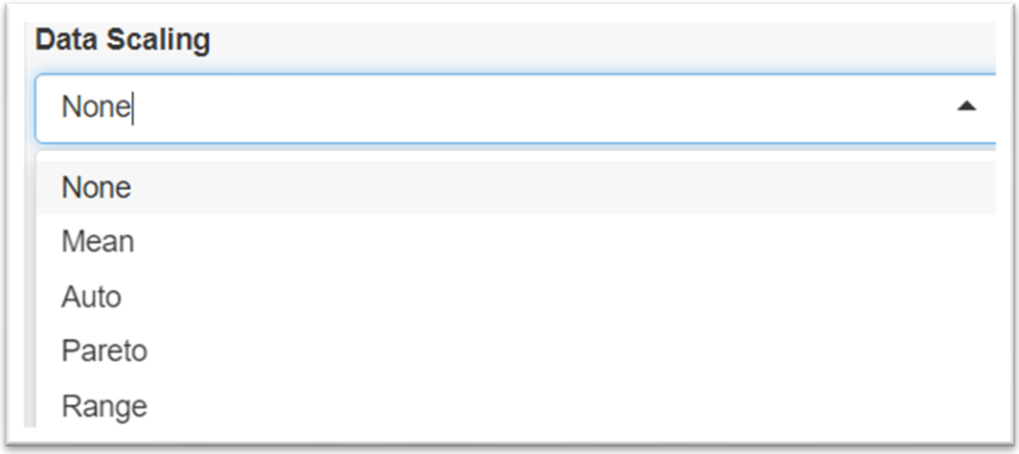

Fig. 7 Opcions Data Scaling

- None: No aplicar cap canvi
- Mean: Aquest mètode consisteix en calcular la mitjana i restar la mitjana a els valors
- Auto: Aquest mètode consisteix en aplicar l'escalat per mitjana i després dividir els valors per la desviació estàndard de cada valor.
- Pareto: Aquest mètode consisteix en aplicar l'escalat per mitjana i després dividir els valors per l'arrel quadrada de la desviació estàndard de cada valor.
- Range: Aquest mètode consisteix en aplicar l'escalat per mitjana i després dividir els valors pel rang de cada valor.

5.4 Descarregues

És possible descarregar-se la base de dades processada tant amb els missing values com amb els canvis fets amb la normalització de mostres, la transformació de dades i l'escalat de dades.

Això es pot fer a través d'un botó disponible sota el menú on s'apliquen aquestes transformacions, i es veu així:

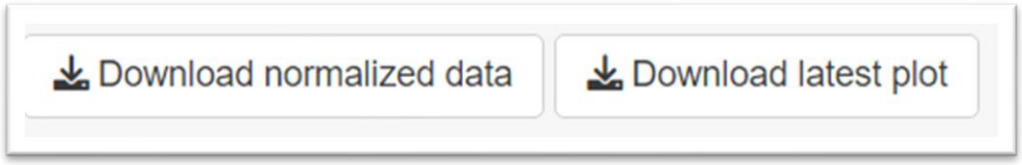

Fig. 8 Botons descarregues

Al costat té un altre botó de descàrrega, i aquest serveix per baixar l'últim gràfic generat. El text en el botó pot canviar segons amb quina categoria d'anàlisis s'estigui treballant.

Ambdós opcions són útils per tal de posteriorment poder continuar amb l'anàlisi, o adjuntar les fotografies per a documentació, com s'està utilitzant per aquest treball.

5.5 Plot-Correlation

Plot-Correlation és el primer anàlisis que hi ha al menú superior, i és el que està preseleccionat. Aquest anàlisis és un dels més senzills disponibles. Consisteix en fer un scatter-plot per poder visualitzar les dades i poder comparar diferents columnes de la base de dades.

Per utilitzar-se s'ha de seleccionar quina variable es vol veure en l'eix X, i quina variable es vol veure en l'eix Y. Això es pot fer en el menú desplegable que hi ha sota el pretractament. Allà es poden seleccionar qualsevol de les columnes de la base de dades, i quan dos diferents estan seleccionades, es mostra el gràfic (Fig. 9).

Aquesta funcionalitat és útil per poder veure quins valors té cada pacient, per veure la correlació entre dos variables, i també per veure si el pretractament aplicat és satisfactori.

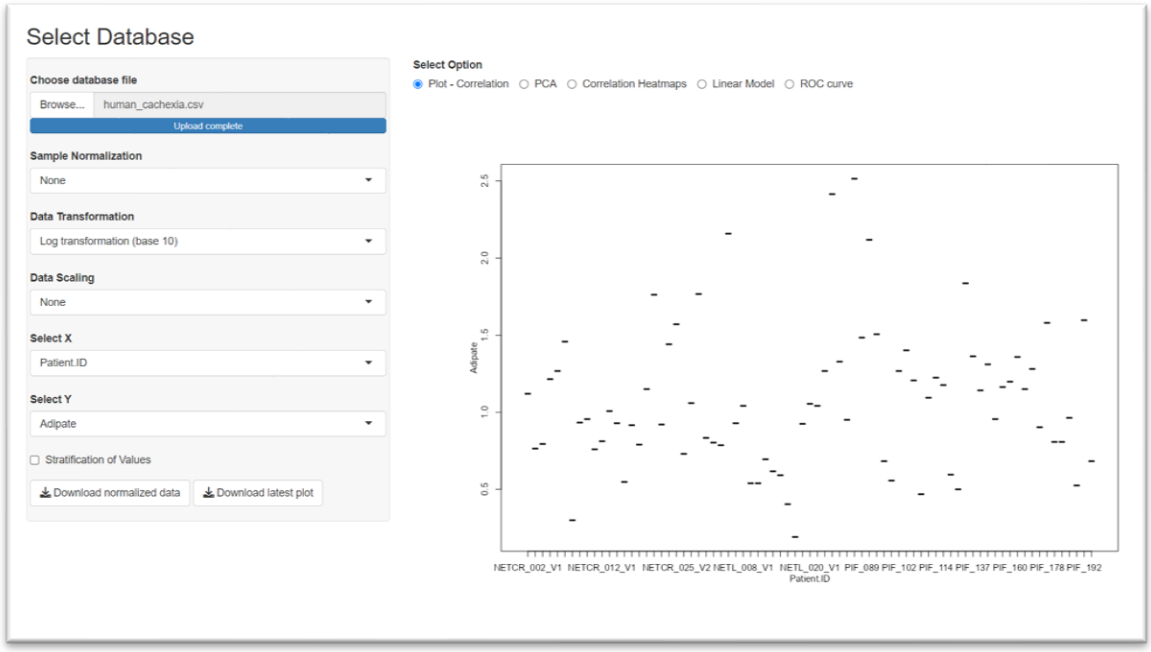

Fig. 9 Exemple us Plot - Correlation

Aquesta és la funcionalitat bàsica, però també hi ha disponible un botó extra, que indica *Stratification of values*. Un cop seleccionat aquest botó, es desplega un petit menú extra on ens permet seleccionar respecte quin paràmetre volem separar els valors. I com els paràmetres seleccionats pot tenir valors numèrics o text, hi ha dos menús diferents:

5.5.1 Estratificació amb variables numèriques

En cas de seleccionar una variable que tingui valors numèrics, apareix un espai a on posar el número que faci de límit per separar les variables, així com quin valor mínim i màxim pots posar. També es proporcionen altres valors estadístics com són la mitjana i la mediana, i dos estimadors com són *Huber's M-estimator* i *Tukey's biweight*, que poden ser útils per decidir quin valor límit utilitzar (Fig. 10). Utilitzant això, el gràfic canvia i passa a estar separats per colors per poder diferenciar els que superen el límit i els que no (Fig. 11).

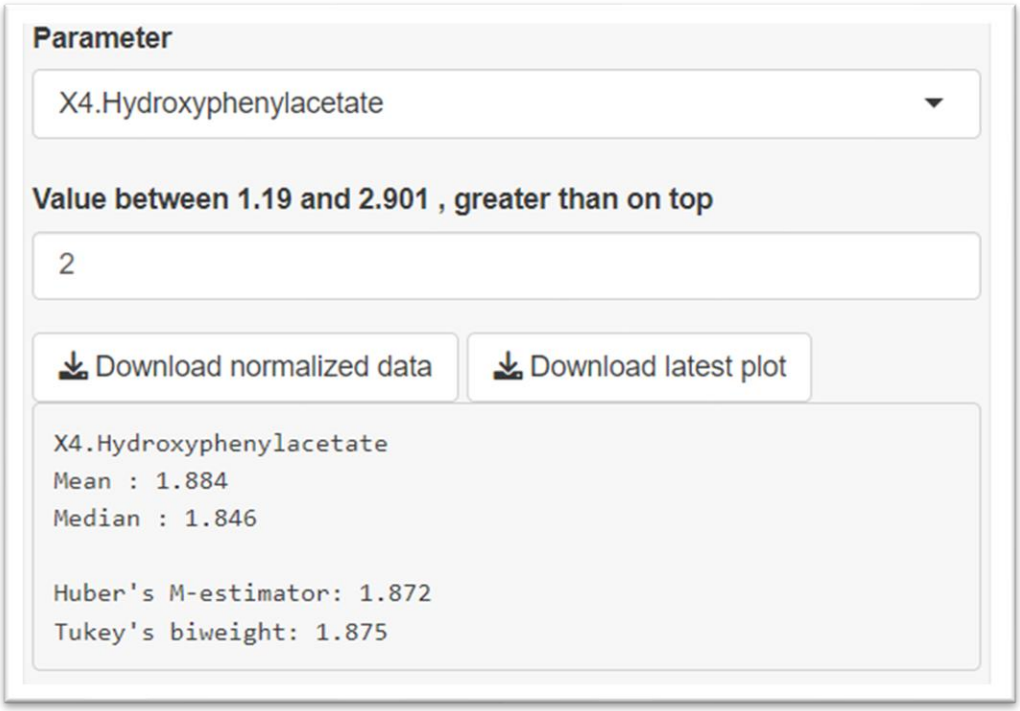

Fig. 10 Estratificació amb variables numèriques

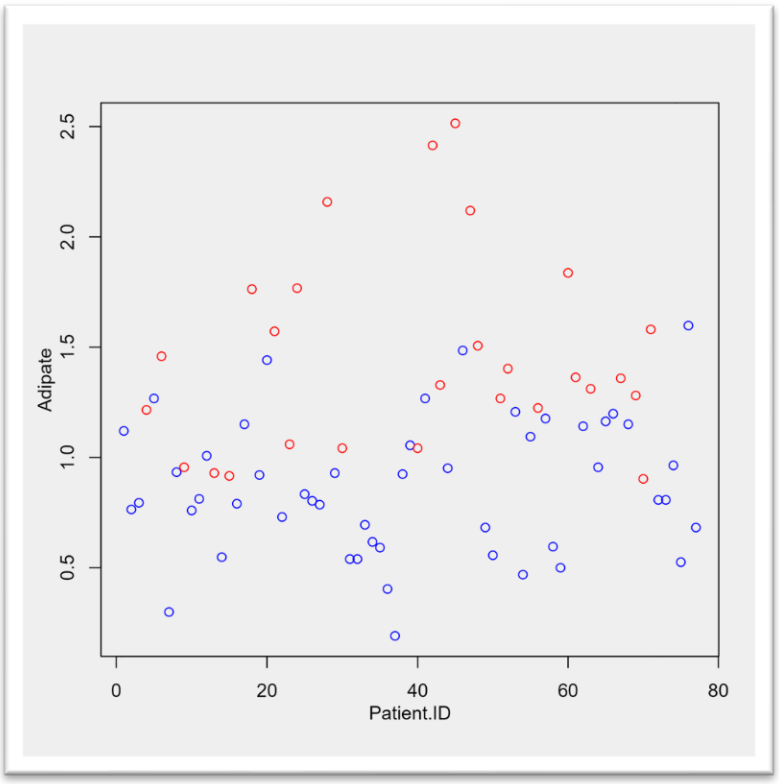

Fig. 11 Exemple estratificació numèrica en scatter-plot

5.5.2 Estratificació amb variables no numèriques

En les variables no numèriques, com podria ser el cas del nom de les mostres o la classificació de les mostres com a malaltes/no malaltes o segons el sexe, es mostra un menú desplegable que permet seleccionar la opció desitjada per a la separació (Fig. 12). I igual que amb la estratificació amb variables numèriques, el gràfic torna a canviar separant per color les noves variables (Fig. 13).

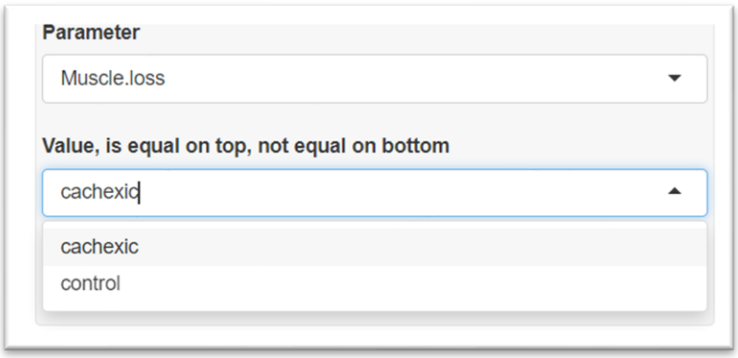

Fig. 12 Estratificació amb variables no numèriques

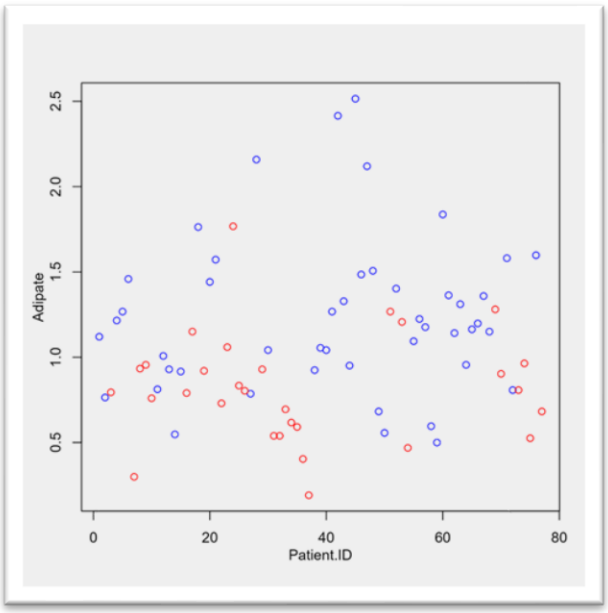

Fig. 13 Exemple estratificació no numèrica en scatter-plot

5.6 PCA

Anàlisis de components principals o PCA, és una tècnica d'anàlisi estadística que s'utilitza per reduir la dimensionalitat d'un conjunt de dades amb moltes variables, mentre es manté la major part de la seva informació. L'objectiu principal del PCA

és trobar les combinacions lineals de les variables originals que expliquin la major variabilitat en les dades.

Per tal de poder utilitzar aquesta funcionalitat, un cop seleccionada aquesta opció en el menú superior, apareix un lliscador i dos llistes desplegables. En les llistes desplegables selecciones quins components vols comparar, i el lliscador és per comoditat, ja que a vegades pot haver-hi fins a més de 300 components, i el lliscador limita els components visibles en les llistes desplegables (Fig. 14).

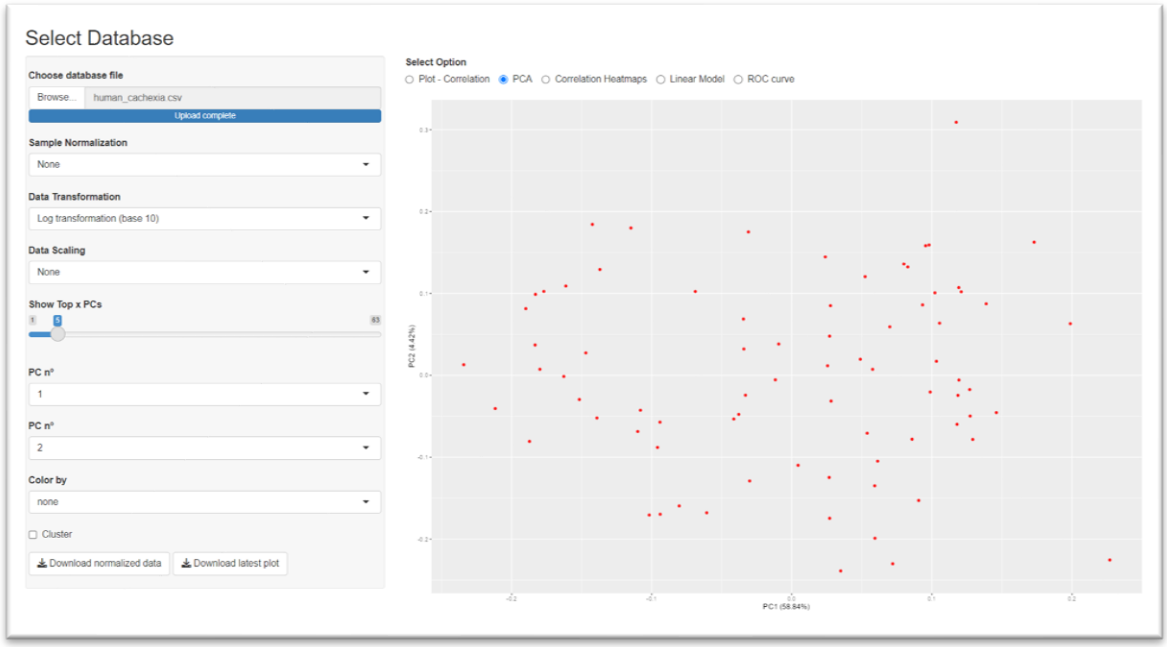

Fig. 14 Menú Principal Compontent Analysis

En aquest menú es pot apreciar dos possibles seleccions extra, que son donar color i fer clústers.

Donar color serveix per millorar la visualització de dades, poder comparar de millor forma i poder destacar subgrups o patrons específics. Per donar color s'utilitza una de les variables originals (Fig. 15).

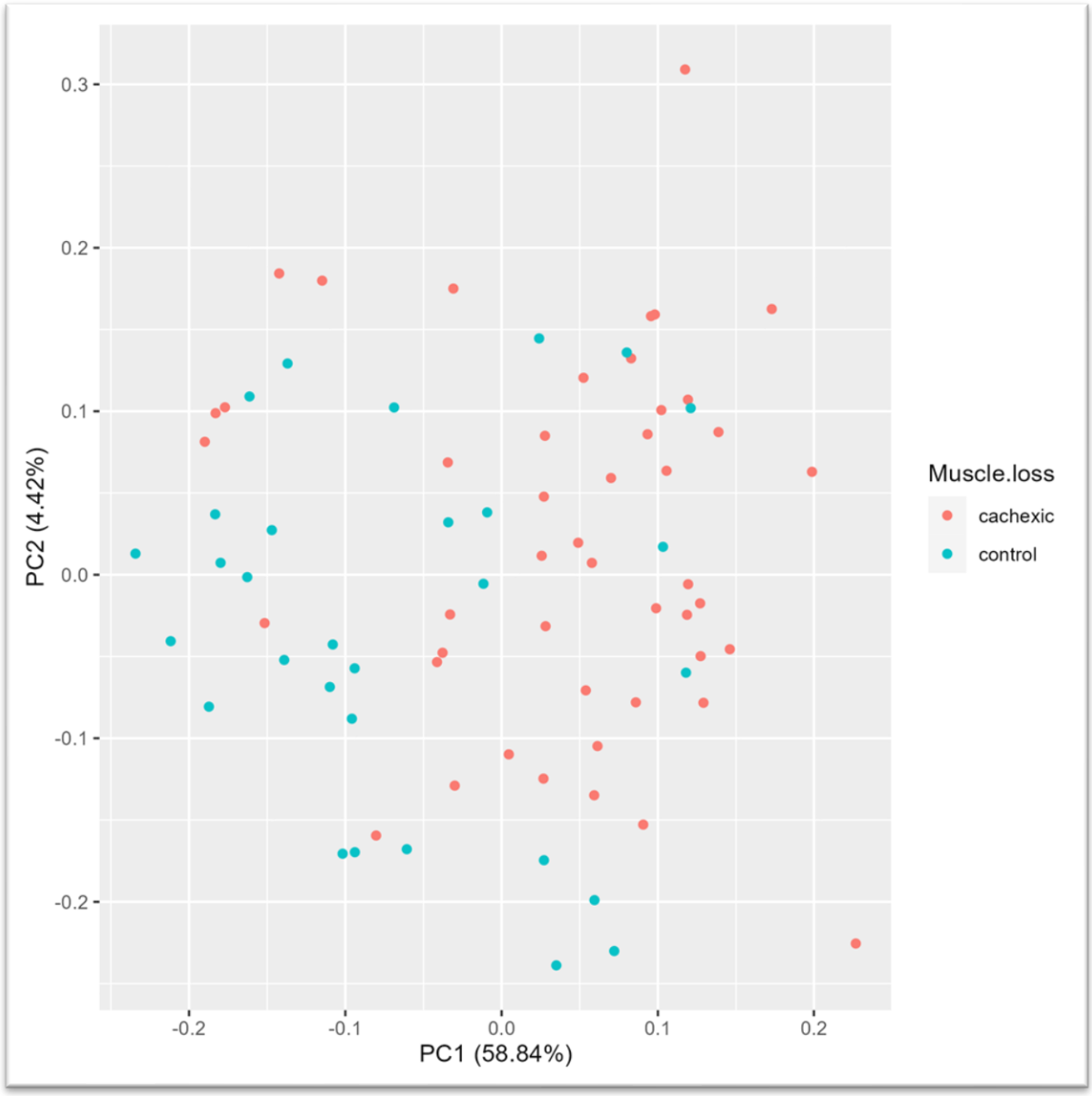

Fig. 15 Exemple PCA amb color

Clústers és una continuació de la utilitat de donar color, on segons el seleccionat es dibuixa una el·lipse que marca la zona a on poden estar els valors. No sempre es pot fer clústers, depèn de la data, però és una eina útil a tenir a disposició per els casos on es pot (Fig. 16).

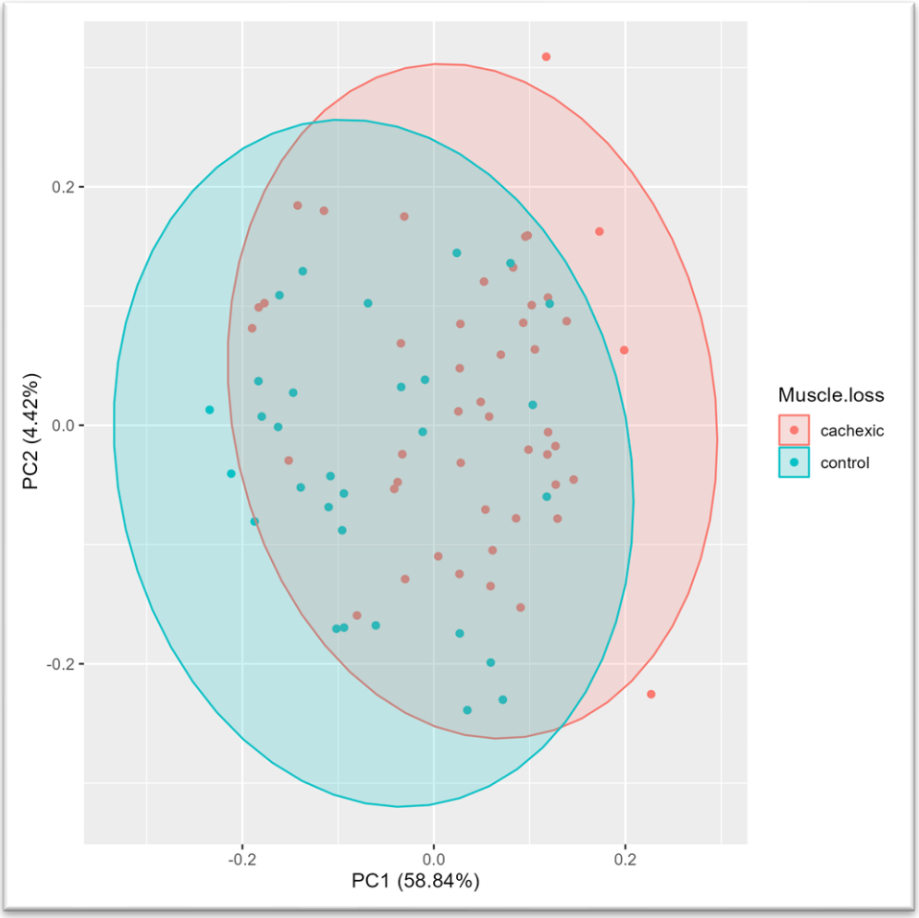

Fig. 16 Exemple PCA amb clústers

5.7 Correlation Heatmap

Un correlation heatmap, o mapa de correlació, és una representació visual de les correlacions entre les variables d'un conjunt de dades. Aquesta representació utilitza colors per indicar la força i la direcció de les correlacions entre parelles de variables.

En un correlation heatmap, les variables són representades en els eixos vertical i horitzontal, i en cada intersecció de la matriu es mostra la correlació entre les variables corresponents. El color de cada cèl·lula reflecteix el valor de la correlació: els colors vermells indiquen correlació més alta, mentre que els tons blaus indiquen menys correlació.

Els heatmaps que s'obtenen són interaccionables. Es pot fer zoom, seleccionar veure un tros de forma més precisa, així com passar el ratolí per sobre per poder veure concretament els valors representats en el heatmap, com quins dos valors s'estan comparant.

Aquí també es pot fer clúster de files o columnes, que consisteix en reordenar les dades per tal de que les files o columnes amb valors més semblants estiguin juntes. Aquesta funcionalitat és útil per poder destacar i veure quins valors poden ser més importants, o quins afecten més als resultats.

També hi ha un botó que s'ha de clicar per dibuixar els heatmaps, ja que aquesta tasca té un cost computacional elevat, i així es pot assegurar que

Els diferents tipus de heatmaps que es poden fer són els següents:

5.7.1 Correlation i P-Value

La correlació i p-value són els dos tests estadístics més utilitzats per establir una relació entre variables. La correlació ens indica si dos variables estan relacionades, i el p-value ens indica si aquesta relació és estadísticament significant.

Hi ha diferents formes de calcular la correlació, i les tres principals es poden seleccionar (Pearson, Kendall, Spearman). Els valors de correlació es queden entre -1 i 1, on els extrems implica una correlació força, i el més a prop a 0 indica no correlació (Fig. 17).

Fig. 17 Rang de correlació

També està disponible un slider per indicar quins valors es volen veure, i així poder treure valors negatius si no interessen, o per exemple treure la franja diagonal on es comparen un valor amb ell mateix, que sempre dona correlació 1 però P-value de 0 indicant que no és significant (Fig. 18).

Fig. 18 Menú Correlation Heatmap - Correlation

La diferència entre seleccionar Correlation o P-value és els valors que surten representats en el gràfic, els dos valors es poden veure igualment passant el ratolí per a sobre de les cel·les concretes.

Aquí també hi ha la opció d'estratificar per valors, explicada prèviament en el plot – correlation. Es pot utilitzar per separar els valors representants en el gràfic, i poder visualitzar millor les diferències entre sexes, què passa quan el ferro té valors per sobre la mitjana entre altres (Fig. 19). Això permet comparar de forma més informada, i veure si hi ha canvis significatius.

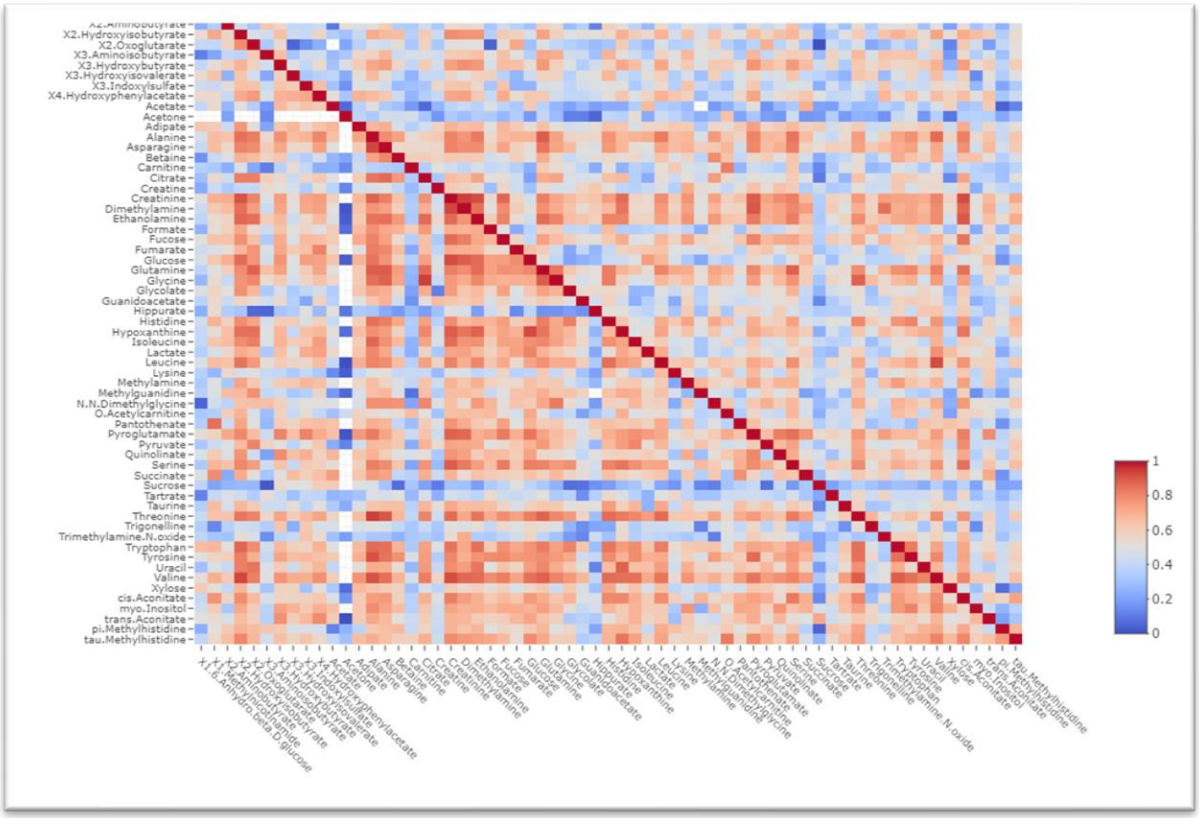

Fig. 19 Correlation heatmap separat per cachexic/control

5.7.2 Collineality

La col·linealitat és una situació en què dues o més variables en un conjunt de dades estan altament correlacionades entre elles. Aquesta correlarció elevada indica que les variables proporcionen informació similar o redundant. En altres paraules, hi ha una relació lineal forta entre les variables, de manera que una variable pot ser pràcticament descrita a partir de l'altra, o aplicant una combinació lineal.

Això pot portar problemes a l'hora d'analitzar les dades o al fer models de predicció, i per això es dona la oportunitat de poder fer aquest anàlisis.

En el següent exemple (Fig. 20) es pot veure l'exemple d'us, on es veu també amb l'ajuda del clúster per columnes que les dades amb més col·linealitat és la columna de Creatinina.

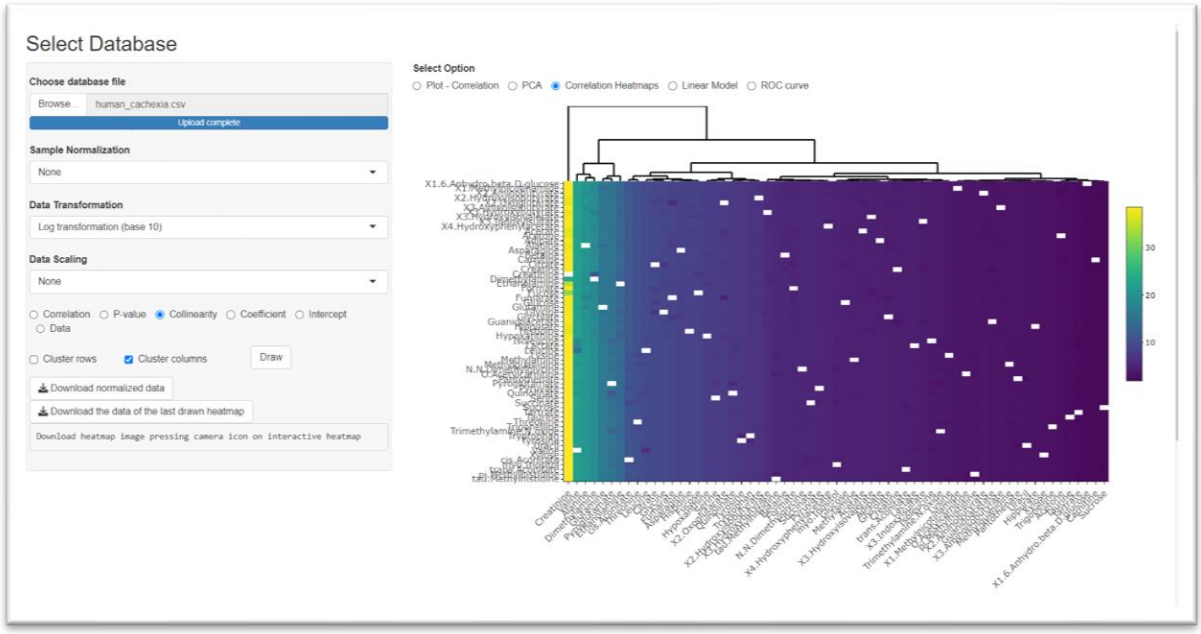

Fig. 20 Exemple col·linealitat

5.7.3 Coefficient i Intercept

El Coefficient i l'intercept són conceptes relacionats amb models lineals. Un model lineal consisteix assumir una relació lineal entre diferents variables. Això es fa utilitzant un o múltiples elements per descriure'n un altre utilitzant els coneixements obtinguts prèviament de la correlació, P-value i col·linealitat. Un dels menús disponibles serveix per construir models lineals, i per tal de tenir més facilitat en aquella tasca s'ha decidit fer aquests heatmaps.

En aquests heatmaps es pot veure respectivament el pendent i l'ordenada a l'origen respectivament. En cada casella del heatmap es pot veure el valor corresponent al model lineal de la fila respecte la columna (Fig. 21). Aquests models són d'una variable respecte només una i és poca informació, però és útil per construir els models lineals.

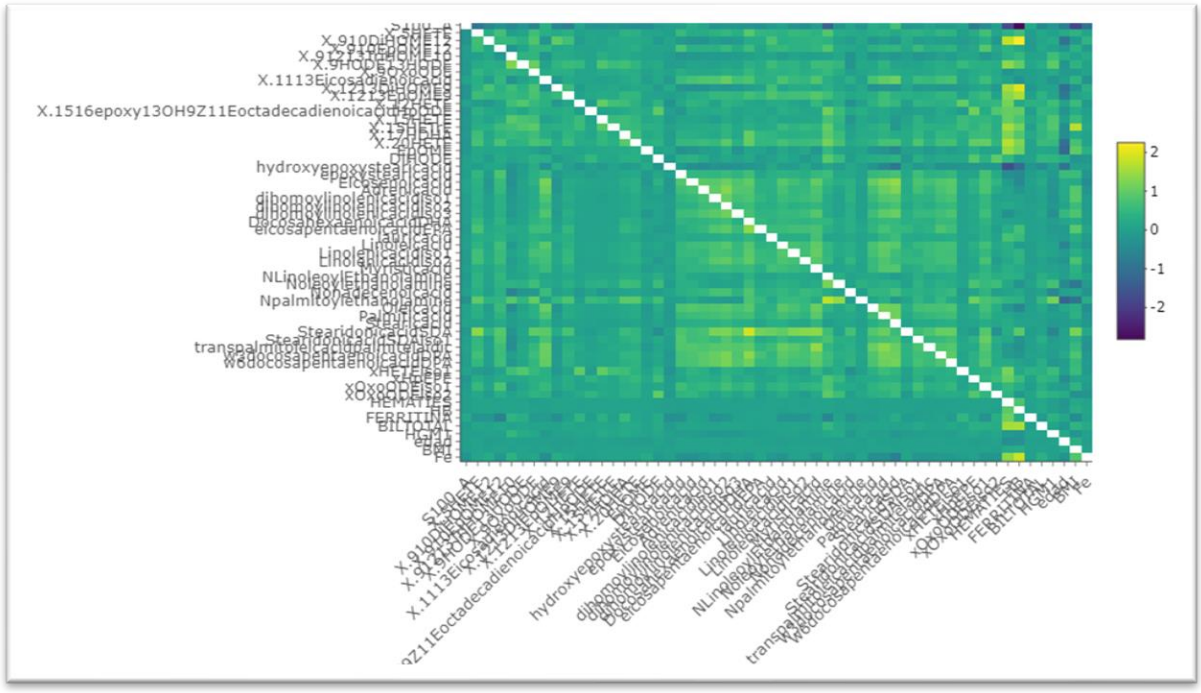

Fig. 21 Exemple Coeficient

5.7.4 Data

També hi ha l'opció de visualitzar la base de dades directament. Aquesta opció té un límit, i amb bases de dades amb masses mostres pot no funcionar, però és una bona forma de veure visualment les dades, i també serveix per veure si el pretractament de dades ha estat adequat.

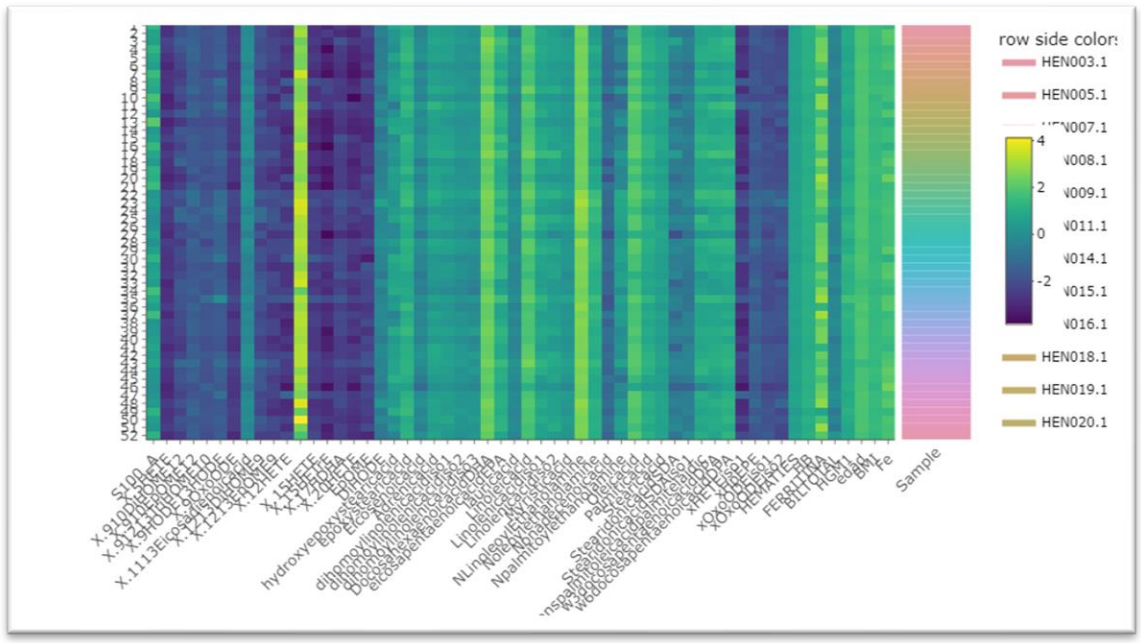

Fig. 22 Exemple Data

5.8 Linear Model

Un model lineal és un tipus de model estadístic que assumeix una relació lineal entre una variable dependent i una o més variables independents. També es coneix com a regressió lineal. Aquest model és amplament utilitzat per predir o explicar la relació entre les variables en base a un conjunt de dades observades.

En un model lineal, la variable dependent s'intenta definir com a combinació lineal de les variables independents, més un valor extra. Es representen de la següent forma

$$Y = m_0x_0 + m_1x_1 + m_2x_2 + m_3x_3 + \cdots + m_nx_n + n$$

On:

- *Y* és la variable dependent que volem predir o explicar
- m_n és el pendent o coeficient relacionat amb la variable independent x_n
- *n* és l'ordenada a l'origen, encara que a vegades també es diu que és l'error.

Per aconseguir ajustar un model lineal s'utilitza intel·ligència artificial per aconseguir els valors òptims a partir de les variables seleccionades. Per treballar ens hem de fixar en el menú (Fig. 23), on es pot veure un heatmap, que és el heatmap dels coeficients. Aquest heatmap és el mateix que obtindríem en el menú anterior, i està aquí per facilitar la selecció de variables, que és l'altre element interactiu en aquest menú.

Fig. 23 Menú Linear Model

Per poder fer el model lineal, primer s'ha de seleccionar la variable a estudiar amb el menú desplegable, i després s'han de seleccionar les variables independents (Fig. 24). Les variables independents es seleccionen clicant a la llista i es poden deseleccionar clicant en elles i prement la tecla d'eliminar.

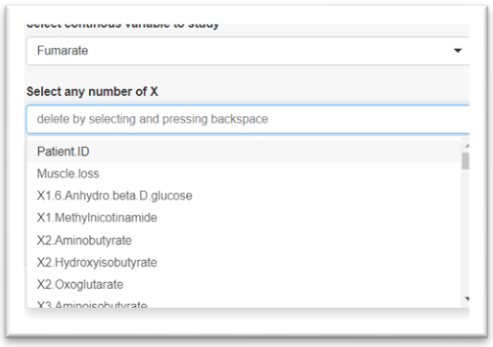

Fig. 24 Selecció variables model lineal

Un cop les variables amb les que es vol fer l'estudi han estat seleccionades, es clica en el botó de Calculate Linear Model i s'obtenen els resultats (Fig 25).

```
Call:
"Fumarate ~ Glutamine + Lactate + Glycolate + Acetate"

Residuals:
    Min      1Q  Median      3Q     Max
-0.5250 -0.1744 -0.0207  0.1107  0.8451

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.88044    0.16188  -5.439 7.00e-07 ***
Glutamine    0.55902    0.11259   4.965 4.46e-06 ***
Lactate      0.38688    0.10230   3.782 0.000319 ***
Glycolate   -0.18736    0.08630  -2.171 0.033235 *
Acetate     -0.04358    0.08857  -0.492 0.624217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2694 on 72 degrees of freedom
Multiple R-squared:  0.6207,    Adjusted R-squared:  0.5996
F-statistic: 29.46 on 4 and 72 DF,  p-value: 1.626e-14
```

Fig. 25 Resultats model lineal

En els resultats es poden veure molts valors importants. De dalt a baix es veu primer quins son els valors utilitzats, en aquest cas hem calculat Fumarat a partir de Glutamina, Lactat, Glicolat i Acetat.

Després hi ha les residuals, que representen les discrepàncies entre els valors observats i els valors predits. En aquest cast veiem que el residual mínim és de –0.52, el residual del primer quartil (25%) és -0.1744, la mitjana és -0.0207, el tercer quartil (75%) és 0.1107, i el màxim residual és 0.8451.

Els següents valors són els coeficients pels diferents elements, i també l'origen de coordenades. Es dona els seus valors, com el seu error estàndard, t-value i p-value per veure la seva significança. En aquest cas es pot veure com la Glutamina i el Lacat són altament significants, el Glicolat també es significant però menys i en aquest cas que l'acetat és insignificant, per lo que si es vol predir el Fumarat és bona idea utilitzar els primers valors i descartar l'últim.

Per últim queden uns altres valors estadístics útils per avaluar la validesa del model linear

5.9 ROC Curve

Una ROC curve (Receiver Operating Characteristic curve) és una representació gràfica utilitzada en l'anàlisi de classificació binària per avaluar i comparar el rendiment d'un model predictiu.

La ROC curve mostra com varia la capacitat de discriminació d'un model a mesura que es modifiquen els punts de tall o thresholds per a la classificació. Cada punt de tall representa un límit a partir del qual les prediccions del model es consideren com a positius o negatius. Movent el punt de tall, es pot ajustar el balanç entre la sensibilitat i l'especificitat del model.

Per poder utilitzar el menú per fer ROC curve, es treballa igual que amb el linear model.

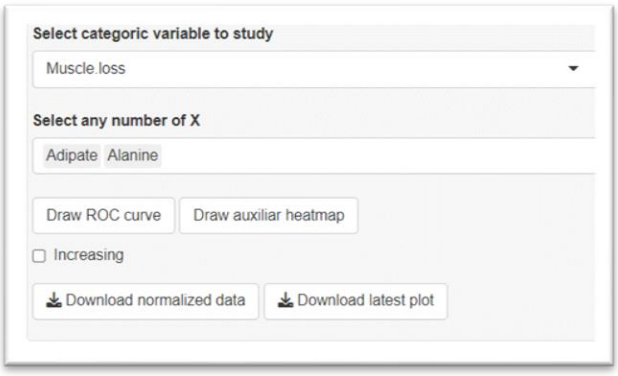

Fig. 26 Selecció variables ROC Curve

Primer es selecciona la variable a estudiar, tenint en compte que ha de ser una variable binària. Després es seleccionen les variables que es volen utilitzar per construir la ROC curve, i un cop fet es selecciona dibuixar (Fig. 26). Un cop fet això, es dibuixa la corba ROC (Fig. 27), i es proporcionen valors com l'àrea sota la corba (AUC) o el límit òptim per fer de classificador, i així poder avaluar si la corba és adequada o si aniria millor continuant fent proves.

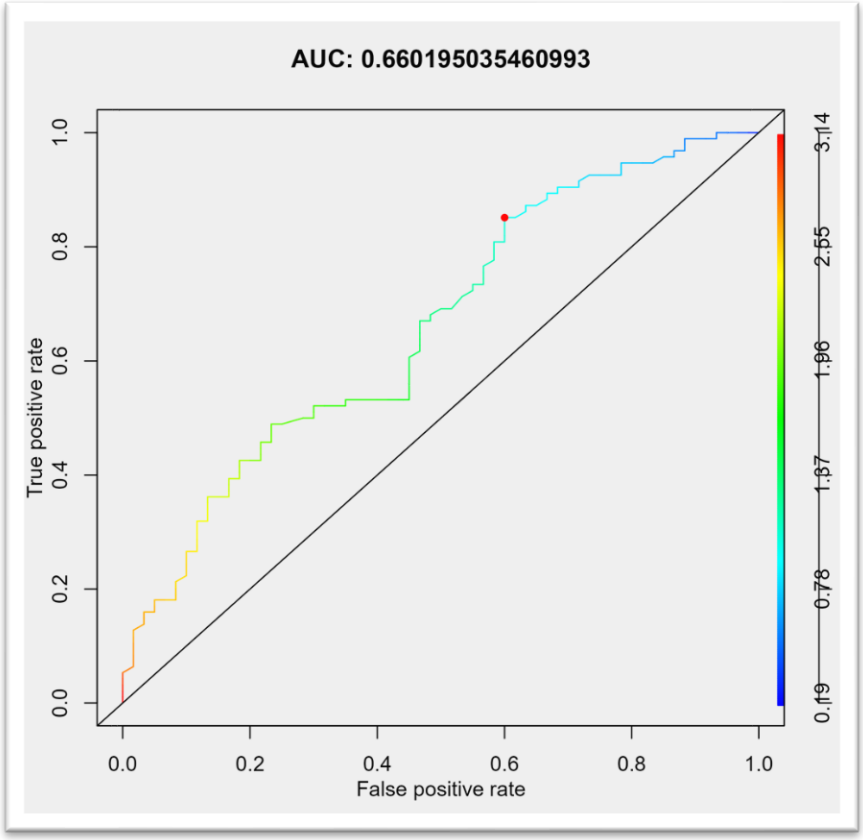

Fig. 27 Corba ROC

I per ajudar a la selecció de les variables a utilitzar per construir la corba ROC, també hi ha la opció de dibuixar un gràfic d'ajuda que indica quina àrea sota la corba tindria la corba ROC que estudiï la variable seleccionada per estudiar i cada altre variable si s'utilitzés individualment per construir la corba ROC.

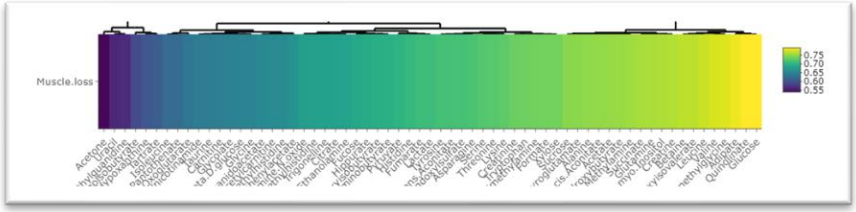

Fig. 28 Gràfic Auxiliar AUC ROC

6 Aplicació Real

Per provar aquesta eina, s'utilitzarà una base de dades amb concentracions de metabòlits de mostres d'orina de pacients amb càncer, obtinguda d'un article (Eisner et al., 2011). Aquest article parla de la predicció de la pèrdua de múscul a causa del càncer, fet que afecta a la funcionalitat i a l'esperança de vida dels pacients. I com els mètodes per detectar la pèrdua de múscul impliquen tècniques d'anàlisis d'imatge com la tomografia computada (CT scan) que poden ser costosos, es proposa com alternativa la detecció a partir d'un anàlisis de metabòlits a partir d'una mostra d'orina.

En aquesta base de dades hi ha mostres d'orina, que s'han pres basat en la hipòtesis de que es poden trobar metabòlits relacionats amb la pèrdua de múscul, i que a partir d'això es podria detectar prèviament una caquèxia, que d'altre forma pot passar per inadvertida en les seves fases inicials. La caquèxia és un símptoma metabòlic complex associat amb malalties que és caracteritzat per la pèrdua de múscul amb o sense pèrdua de grassa. Els metabòlits que s'espera poder utilitzar per detectar la caquèxia serien metabòlits del múscul (creatina, creatinina, 3-OH-isovalerat), aminoàcids (Leu, Ile, Val, Ala, Thr, Tyr, Gln, Ser) i altres metabòlits intermediaris.

Hi ha un total de 77 mostres d'orina, de les quals n'hi ha 37 del grup de pacients caquèctics, i 30 del grup control. S'han pres sense tenir en compte ni l'hora del dia ni la dieta per poder facilitar una presa de mostres futures.

Per començar el pretractament ens hem de fixar en la pròpia base de dades. En el nostre cas està 100% completa, no hi ha cap valor faltant, per lo que no és important quina opció de neteja de base de dades relacionat amb valors buits seleccionem.

Un cop hem carregat la base de dades, s'ha de mirar si és necessari fer un pretractament a les dades abans de començar l'anàlisi. Utilitzant el menú plot-correlation podem observar i decidir si ens fa falta algun pretractament. I si ens fixem en aquestes dos figures de mostra (Fig. 28)(Fig. 29), ens podem adonar que hi ha valors anormals i que és preferible utilitzar pretractament.

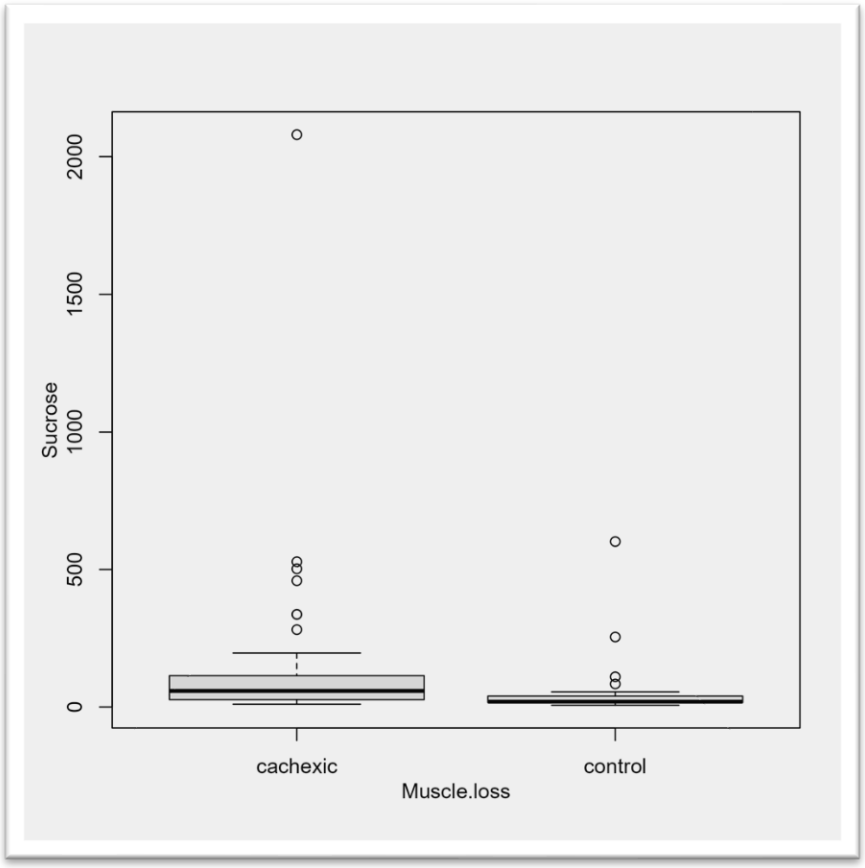

Fig. 29 Muscle loss vs Sucrose sense preprocessament

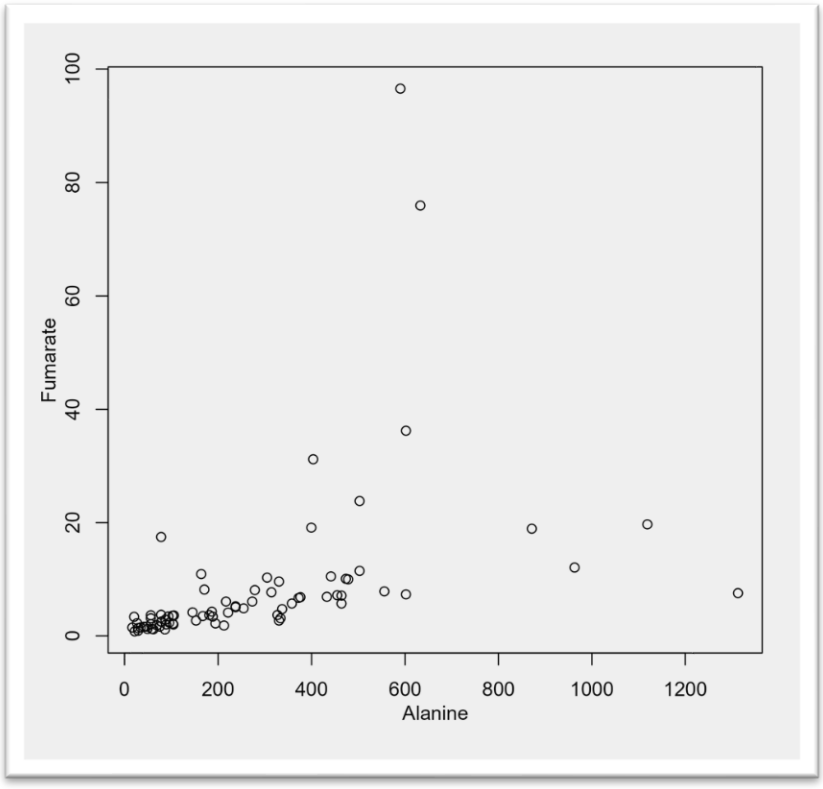

Fig. 30 Alanine vs Fumarate sense preprocessament

Aquests valors es poden solucionar fàcilment aplicant un pretractament, i per la forma que tenen, aplicar el logaritme base10 com a data transformation aconsegueix un bon resultat, on excepcionalment poden quedar algun valor anormal però dintre d'un rang més còmode.

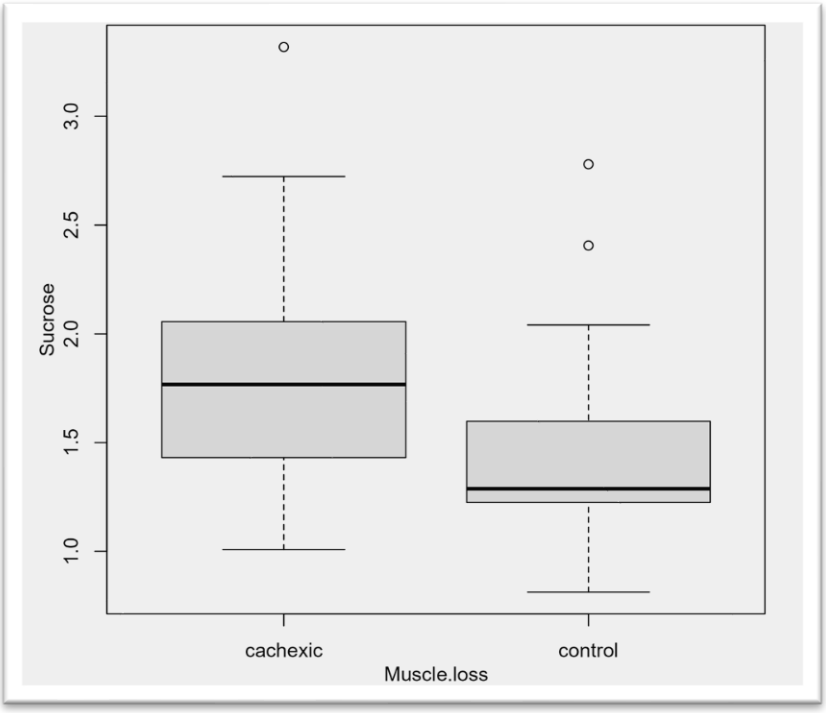

Fig. 31 Muscle loss vs Sucrose amb preprocessament

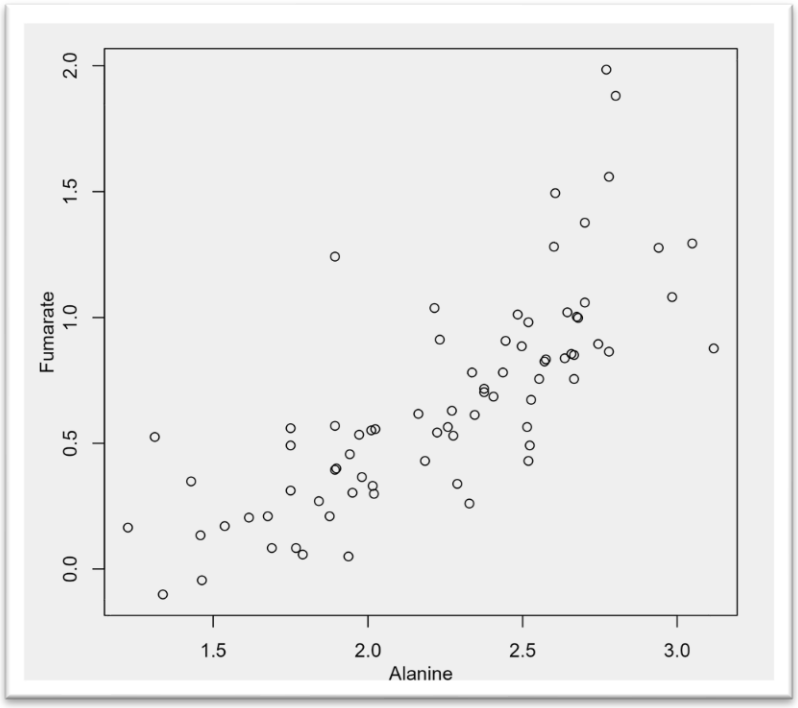

Fig. 32 Alanine vs Fumarate amb preprocessament

Ja que ens interessa poder classificar mostres de futurs pacients com a caquèctics o no caquèctics, haurem de fer un model de predicció, que en aquest cas serà una corba ROC.

Per començar l'anàlisi anem al menú ROC, i podem utilitzar el gràfic d'ajuda per començar a treballar (Fig. 33). En aquest podem veure ja quines variables individuals ens poden ajudar a construir una bona corba ROC. Fins i tot es pot veure en la llegenda que ja hi ha valors amb una àrea sota la corba amb valors superiors al 0.75, valor completament acceptables.

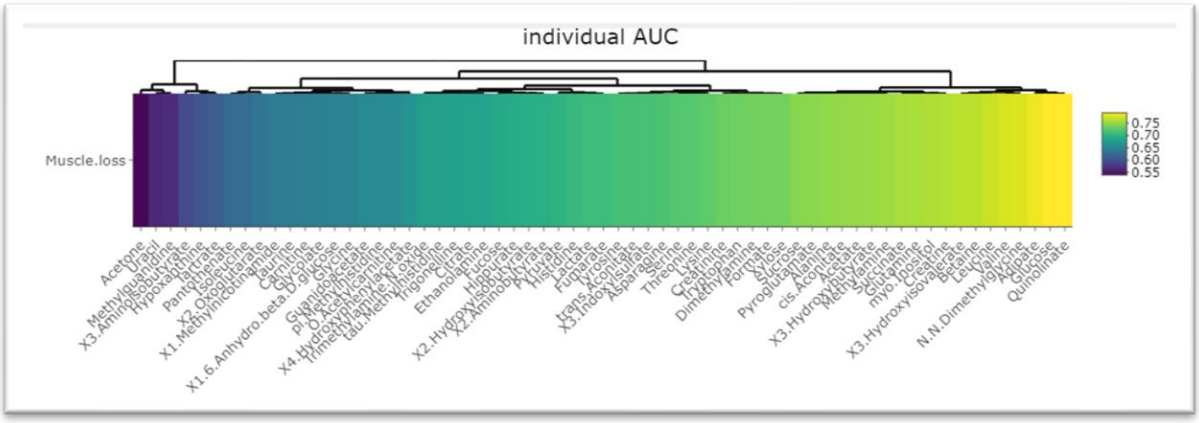

Fig. 33 AUC de les variables per l'anàlisi

Agafant la variable amb una àrea sota la corba més gran que té individualment, que en aquest cas és la Quinolinate (Fig. 34), ja obtindríem uns valors adequats per poder fer una classificació, utilitzant com indica el límit de tall de 1.537, que indica que classifica tots els que tinguin un valor de Quinolinate superior a 1.537µM (havent aplicat abans el logaritme base 10 per normalitzar) com a caquèctics.

El problema que hi ha utilitzar només la Quinolinate, és que per la poca quantitat de dades que tenim és poc fiable, per lo que és preferible utilitzar més variables per fer la corba ROC i tenir més confiança. Quan s'utilitzen més variables, es fan servir més punts per a la construcció de la corba ROC. És possible que això pugui reduir l'àrea sota la corba, però s'obté un avantatge en termes de quantitat de dades utilitzades, la qual cosa augmenta les possibilitats que el classificador obtingut sigui més efectiu en la classificació correcta de les mostres.

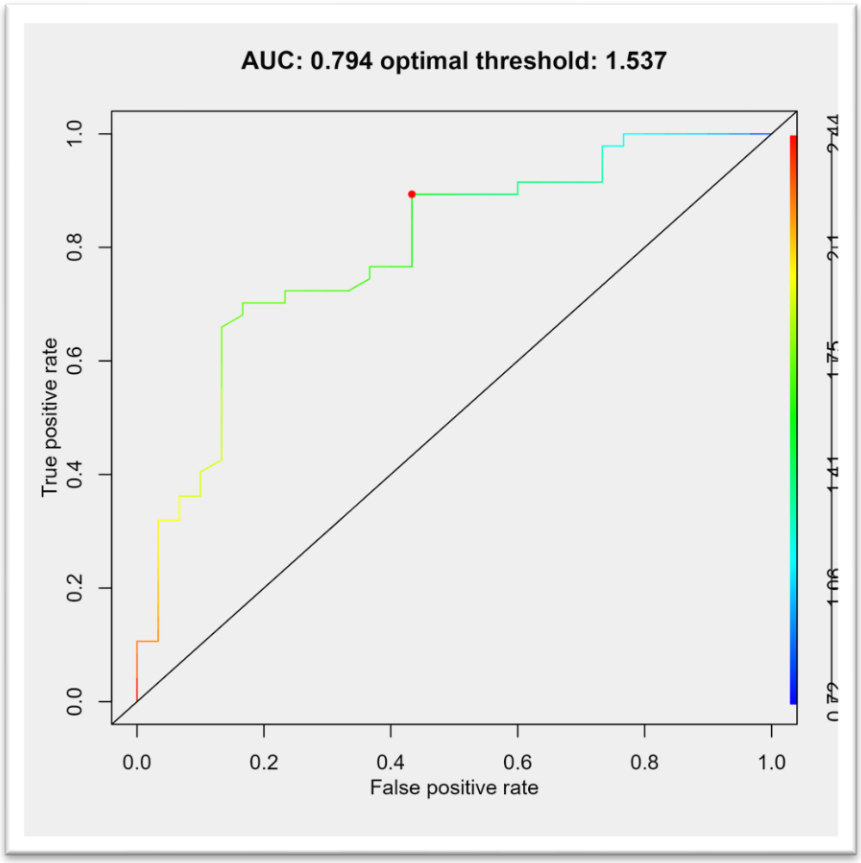

Fig. 34 ROC curve de Quinolinate

Després d'utilitzar diferents combinacions de variables, s'han aconseguit diferents classificadors amb dos variables (Fig. 35), tres variables (Fig. 36) i quatre variables (Fig. 37)(Fig. 38).

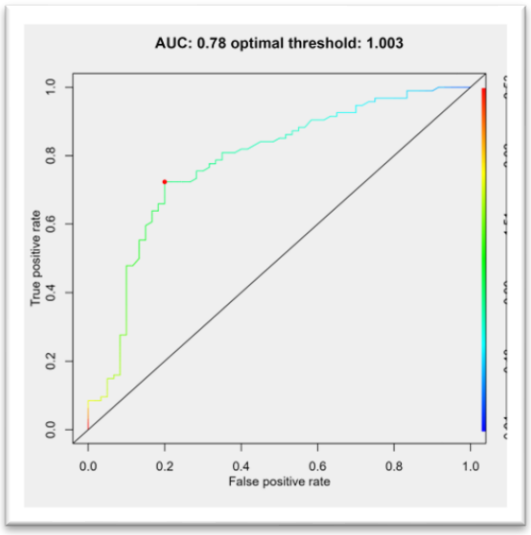

Fig. 35 ROC curve de adipate, x3.hydroxyisovalerate

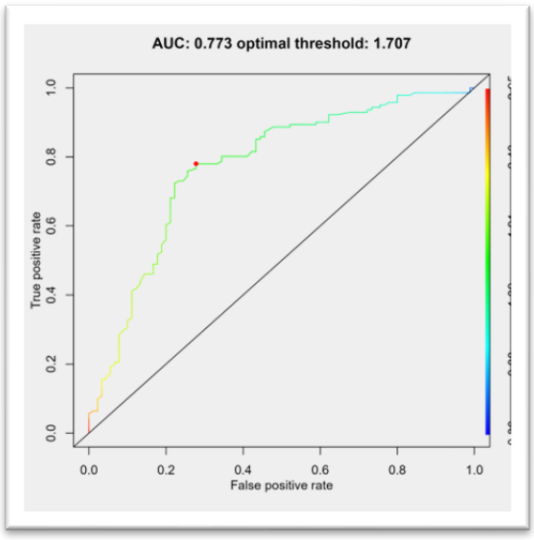

Fig. 36 Roc curve de Quinolinate, Betaine, myo.Inositol

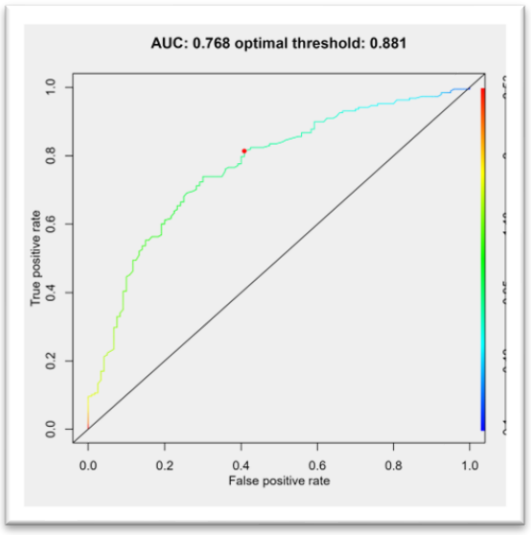

Fig. 37 Roc curve de Adipate, N.N.Dimethylgycine, X3.hydroxybutyrate, X3.Hydroxyisovalerate

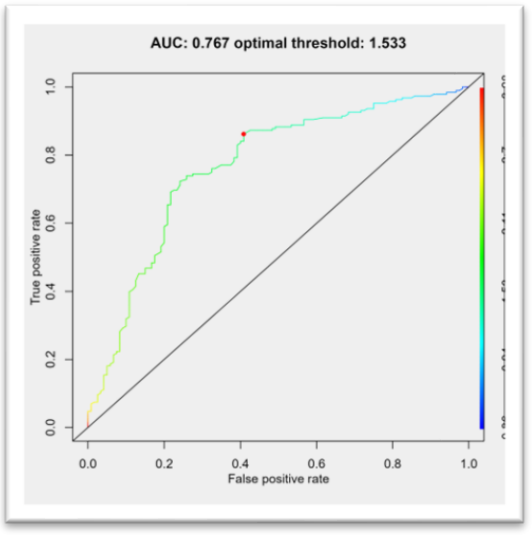

Fig. 38 Roc curve de Betaine, Creatine, myo.Inositol, Quinolinate

Si es comparen totes les corbes roc (Fig. 34-37)(Pepe et al., 2009) es pot veure com amb més variables aconseguim una corba amb menys àrea, encara que ja hagi triat les corbes que tenen una àrea més gran. Però això no implica que siguin pitjors, ja que el que representa una corba ROC és els verdaders positius contra els falsos positius d'un classificador, i seleccionant un límit adequat que classifiqui de millor forma. Per tant, tenint això en compte, la millor classificació que s'ha obtingut és la representada a la Fig. 38, on s'ha predit la pèrdua de múscul utilitzant Betaine, Creatine, myo.Inositol i Quinolinate. Aquesta corba, mentre no té la millor AUC, és

el que té un punt òptim amb millor sensibilitat. Utilitzant el límit de 1.533, o 34.12 si no es normalitza les dades, classifica correctament com a verdaders un 86% de les mostres, que és el més elevat entre totes les corbes roc mostrades. També classifica incorrectament com a verdaderes mostres un 40% de les vegades, que es podria classificar com a moltes vegades per lo que no és massa òptim en aquesta banda, per lo que es podria canviar de classificador en cas de que es volgués trobar un millor equilibri, però en aquest cas considero que és més important classificar correctament els pacients amb caquèxia encara que se'n classifiquin d'extres.

Com a conclusió d'aquesta aplicació en un cas real, l'aplicació ha estat útil i ha ajudat a fer un classificador adequat i senzill per poder predir si pacients amb càncer poden tenir caquèxia a partir d'una mostra d'orina.

7 Conclusions

Les conclusions d'aquest treball de fi de grau basats en els objectius establerts són els següents:

1. S'ha après el llenguatge de programació R, que ha estat necessari per poder desenvolupar l'aplicació.
2. S'ha dissenyat i programat una aplicació que incorpora diverses tècniques d'anàlisi de dades biotecnològiques, i que ajuden als usuaris a fer anàlisi i obtenir visualment informació rellevant sobre el seu conjunt de dades.
3. S'ha aplicat l'aplicació en un cas d'estudi real per avaluar la seva utilitat i funcionalitat. Mitjançant l'ús de dades reals, s'ha posat a prova l'aplicació i s'han obtingut bons resultats.

En resum, tots els subobjectius establerts inicialment han estat complerts amb èxit, conduint a l'acompliment de l'objectiu principal. Aquest treball de fi de grau ha contribuït al desenvolupament d'una aplicació d'anàlisi de dades funcional i útil en àmbits de recerca mèdica, obrint noves oportunitats per a l'avanç del coneixement i la comprensió en aquests àmbits.

8 Referències

Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., Auman, J. T., Balasundaram, M., Balu, S., Baylin, S. B., Behera, M., Bernard, B., Beroukhim, R., Bishop, J. A., Black, A. D., Bodenheimer, T., Boice, L., Bootwalla, M. S., Bowen, J., … Xing, M. (2010). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *11*(3).

Bro, R., & Smilde, A. K. (2014). Principal component analysis. In *Analytical Methods* (Vol. 6, Issue 9). https://doi.org/10.1039/c3ay41907j

Calabrese, B. (2018). Data cleaning. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3). https://doi.org/10.1016/B978-0-12-809633-8.20458-5

Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S., Greiner, R., Wishart, D. S., & Baracos, V. E. (2011). Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics*, *7*(1). https://doi.org/10.1007/s11306-010-0232-9

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MaCdonald, J., Obenchain, V., Oleš, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2). https://doi.org/10.1038/nmeth.3252

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1). https://doi.org/10.1093/nar/gkw1092

Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, *4*(1). https://doi.org/10.4103/jpcs.jpcs_8_18

Lun, A. T. L., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, *5*. https://doi.org/10.12688/f1000research.9501.2

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1). https://doi.org/10.4097/kja.21209

Pang, Z., Chong, J., Zhou, G., De Lima Morais, D. A., Chang, L., Barrette, M., Gauthier, C., Jacques, P. É., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research*, *49*(W1). https://doi.org/10.1093/nar/gkab382

Pepe, M. S., Longton, G., & Janes, H. (2009). Estimation and comparison of receiver operating characteristic curves. *Stata Journal*, *9*(1). https://doi.org/10.1177/1536867x0900900101

S, M. (2010). Data transformation. *Journal of Pharmacology and Pharmacotherapeutics*, *1*(2). https://doi.org/10.4103/0976-500x.72373

Team, R. C. (2021). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing*.

Wilke, C. (2019). Fundamentals of Data Visualization. *Journal of Chemical Information and Modeling*, *53*(9).

Wu, Y., & Li, L. (2016). Sample normalization methods in quantitative metabolomics. In *Journal of Chromatography A* (Vol. 1430). https://doi.org/10.1016/j.chroma.2015.12.007