UNIVERSITAT DE BARCELONA

Generació i anàlisi d'un model per relacionar el microbioma humà i dades clíniques amb malalties autoimmunitàries

Joan Canet Carbó
Màster Universitari en Bioinformàtica i Bioestadística
TFM – Bioinformàtica i Bioestadística Àrea 3

Andreu Paytuví Gallart
Ferran Prados Carrasco

3 de gener de 2020

FITXA DEL TREBALL FINAL

Títol del treball:	*Generació i anàlisi d'un model per relacionar el microbioma humà i dades clíniques amb malalties autoimmunitàries*
Nom de l'autor:	*Joan Canet Carbó*
Nom del consultor/a:	*Andreu Paytuví Gallart*
Nom del PRA:	*Ferran Prados Carrasco*
Data de lliurament (mm/aaaa):	*01/2020*
Titulació o programa:	*Màster Universitari en Bioinformàtica i Bioestadística*
Àrea del Treball Final:	*TFM – Bioinformàtica i Bioestadística Àrea 3*
Idioma del treball:	*Català*
Paraules clau	*Microbioma humà, Machine Learning, malalties autoimmunitàries*

Resum del Treball:

Diferents composicions taxonòmiques al microbioma intestinal han estat relacionades a algunes malalties com la diabetis o la malaltia de Crohn. En aquest projecte s'han descrit la composició microbiològica de mostres fecals i variables clíniques –com l'edat o l'índex de massa corporal– associades a un gran nombre d'individus. S'han generat diferents models amb algoritmes de *Machine Learning*, com *Random Forest*, *Support Vector Machine* i XGBoost, per predir si un subjecte ha desenvolupat, o no, alguna malaltia autoimmunitària, a partir de la seua composició taxonòmica intestinal i unes variables clíniques.

Els resultats obtinguts a la descripció taxonòmica no mostren uns enterotips clarament diferenciats entre les mostres. La majoria de les variables clíniques categòriques no estan segueixen una distribució equilibrada de les classes. Les distribució de les variables clíniques numèriques analitzades sí que s'aproximen a una distribució normal. El millor model de classificació s'ha obtingut utilitzant un mètode de mostratge anomenat SMOTE per a generar les dades d'entrenament i emprant l'algoritme XGBoost, obtenint un valor de l'estadístic Kappa de 0.6612. Aquest valor es considera que té una adequació substancial a les dades reals. El gènere *Bifidobacterium* ha sigut el que més ha contribuït en el rendiment del model.

A tall de cloenda, no s'han pogut classificar les mostres en enterotips clarament diferenciats; no obstant això, s'ha pogut generar un model per predir si una persona ha desenvolupat, o no, una malaltia autoimmunitària, utilitzant dades del microbioma intestinal i variables clíniques amb una adequació substancial a les dades reals.

Abstract:

Different taxonomic compositions of the gut microbiome have been related to some diseases, such as diabetes or Crohn's disease. In this project, the microbiological composition of fecal samples and clinical variables –such as age or body mass index– associated to a big number of subjects have been described. Different models have been generated using Machine Learning algorithms, such as Random Forest, Support Vector Machine and XGBoost, to predict whether a subject has developed, or not, any autoimmune disease, using its gut taxonomic composition and some clinical variables.

The obtained results in the taxonomic description do not show very differentiated enterotypes between the samples. Most of the categorical clinical variables do not follow a balanced distribution of their levels. The analyzed numerical clinical variables' distribution does follow approximately a normal distribution. The best classifier model has been obtained using a sampling method called SMOTE to generate the training set and using the XGBoost algorithm, obtaining a Kappa statistic value of 0.6612. This value is considered to have a substantial adequacy to the real data. The *Bifidobacterium* genus has been the one that has contributed the most to the model performance.

In conclusion, the samples could not be classified into very differentiated enterotypes; however, a model to predict whether a subject has developed, or not, an autoimmune disease has been generated, using gut microbiome data and clinical variables, giving a substantial adequacy to the real data.

Índex

Llista de figures

Llista de taules

1. Introducció

1.1 Context i justificació del Treball

La metagenòmica és l'estudi de la col·lecció de genomes dels microorganismes recollits a partir d'una sèrie de mostres mitjançant la seqüenciació i l'anàlisi del seu material genètic. El microbioma és el conjunt dels genomes dels microorganismes presents en un nínxol determinat, com pot ser el cos humà o els intestins (Ley *et al*., 2006). S'ha comprovat que segons la regió del cos d'on provinguen aquestes mostres (pell, intestins, òrgans sexuals, etc.) es mostren uns perfils completament diferents. A més a més, se sap que aquesta diversitat pot implicar un paper important en la susceptibilitat de desenvolupar algunes malalties, com l'obesitat, la malaltia de Crohn (Kho & Lal, 2018), l'esclerosi múltiple, la diabetis, al·lèrgies, asma, autisme ,i inclús, càncer (Lloyd-Price *et al.*, 2016; Mezouar *et al.*, 2018). Per aquestos motius, és molt important fer recerca sobre el microbioma i altres factors d'importància clínica, com poden ser l'edat o l'índex de massa corporal, per trobar relacions entre malalties i aquest tipus de dades.

Les tècniques de *Machine Learning* es poden utilitzar per predir probabilitats fent servir l'entrenament d'un model mitjançant diferents tècniques a partir d'un conjunt de dades. Aquestes tècniques són molt útils a l'hora de realitzar classificacions de pacients i donar les probabilitats que aquestos resultats es complisquen (Yatsunenko *et al.*, *2012*). En un estudi s'han utilitzat tècniques de *Machine Learning* per classificar pacients que tenen malalties inflamatòries intestinals a partir de la comparació de les famílies de proteïnes ortòlogues presents en el microbioma intestinal de pacients amb aquest tipus de malalties i pacients sans (Yazdani *et al*., 2016). En un altre article, es van utilitzar tècniques de *Machine Learning* per classificar pacients amb el síndrome del colon irritable tipus C, tipus U i individus sans utilitzant dades del microbioma intestinal (Riehle *et al*., 2012). L'equip de Rahman i col·laboradors (2018) va demostrar que el *Machine Learning* pot utilitzar dades metagenòmiques per identificar gens clau de certs microorganismes per sobreviure al tractament d'antibiòtics i predir com respondrà el microbioma intestinal a l'administració d'aquest tipus de fàrmacs.

Per a l'elaboració d'aquest treball s'ha disposat de més de 16.000 comptatges normalitzats de perfils taxonòmics dels microorganismes presents en l'intestí de diferents pacients, a més de moltes altres dades (tant clíniques com no clíniques) –també anomenades metadades en aquest text– que podrien estar relacionades amb malalties. Com que s'ha disposat de dades respectives a malalties que podrien incloure's dins de les patologies del sistema immunitari, com les autoimmunitàries, s'ha intentat predir la probabilitat de haver-les desenvolupat, o no, segons totes les dades que s'acaben de mencionar.

És molt important fer recerca sobre el microbioma i aquestos factors ja que, d'una banda, s'obtindria un coneixement de les dades que sí que estan relacionades amb algunes malalties i les que no, motiu que ens permetria seguir amb una línia concreta per poder millorar els models existents (si és que n'hi ha) o, inclús, desenvolupar-ne de nous (Yatsunenko *et al.*, 2012), com és el cas d'aquest treball. D'altra banda, amb els resultats d'aquest projecte es podria crear un model que ens podria ser útil en un futur a l'hora d'elaborar el diagnòstic d'un pacient d'una manera més precoç, ja que es podria seguir un pacient d'una manera més precisa i dirigida si se sap que té una probabilitat relativament alta d'haver desenvolupat alguna malaltia autoimmunitària.

1.2 Objectius del Treball

Aquest treball consisteix en la generació i anàlisi d'un model realitzat utilitzant tècniques de *Machine Learning* per relacionar el microbioma humà intestinal i dades susceptibles de ser d'importància clínica per poder predir probabilitats d'haver desenvolupat alguna malaltia autoimmunitària. A continuació, s'exposen els objectius generals del projecte, desglossats amb els seus objectius específics:

a) Descripció i anàlisi de la composició taxonòmica dels microorganismes presents al microbioma intestinal humà i de metadades.
    - Descriure els gèneres i/o espècies presents al microbioma intestinal dels pacients.
    - Elegir els factors que seran utilitzats per la creació d'un model de classificació.
    - Descriure les metadades.
b) Generació d'un model de *Machine Learning* per predir la probabilitat de haver desenvolupat alguna malaltia autoimmunitària a partir de dades taxonòmiques i clíniques.
    - Entrenar un model de classificació utilitzant l'algoritme de *Random Forest*.
    - Avaluar del rendiment del model de classificació.
    - Millorar el model de classificació amb altres característiques d'entrenament o altres algoritmes de classificació.
    - Analitzar l'impacte que tenen les diferents variables predictores amb el funcionament del model.

1.3 Enfocament i mètode seguit

En primer lloc, s'han carregat i preparat les dades per a tindre-les en un format estàndard i fàcil d'utilitzar, com arxius CSV, a partir d'arxius en format JSON amb les metadades corresponents als subjectes del projecte PRJEB11419 de l'*American Gut Project* (McDonald *et al.*, 2018). La transformació de les dades a aquest format permet que siga possible utilitzar-les emprant la majoria de llenguatges de programació amb relativa facilitat. Les dades amb la informació de la taxonomia del microbioma intestinal s'han carregat a partir d'un arxiu TSV. Una vegada s'ha disposat de les dades en aquest format concret s'han realitzat diverses taules i

figures amb estadística descriptiva. En aquest pas, s'han analitzat els perfils taxonòmics del microbioma intestinal de una gran quantitat de pacients, i també les metadades d'aquestos subjectes. Amb tota aquesta informació i la present a la bibliografia s'han elegit una sèrie de factors que s'han utilitzat posteriorment per entrenar un model de classificació.

Seguidament, s'ha entrenat un model classificador *Random Forest* per a poder calcular les probabilitats que els pacients tinguen alguna malaltia autoimmunitària. Aquest classificador ha utilitzat com a dades d'entrenament els perfils taxonòmics del microbioma intestinal, les dades clíniques i altres tipus de dades d'aquestos pacients. El classificador *Random Forest* té els avantatges, entre altres, que pot utilitzar variables amb soroll i també les característiques numèriques i categòriques; a més, selecciona únicament els factors més importants (Lantz, 2015). Després d'haver entrenat aquest model, s'ha avaluat el seu rendiment i s'ha intentat millorar utilitzant altres característiques d'entrenament, com per exemple el "número d'arbres" utilitzats per al seu entrenament i altres mètodes de mostratge. Com no s'ha obtingut un model amb un bon valor de Kappa, s'ha intentat obtindre un millor model utilitzant altres algoritmes de *Machine Learning*, com el *Support Vector Machine* i el XGBoost. Finalment, una vegada s'ha obtingut el millor model per realitzar les classificacions utilitzant l'algoritme de XGBoost, s'ha analitzat l'impacte que tenen les diferents variables predictores amb el funcionament del model.

Per a realitzar aquest projecte s'ha utilitzat el llenguatge de programació **R** (R Core Team, 2019) i la seua interfície d'usuari **RStudio** (RStudio Team, 2016). Encara que hi ha altres llenguatges de programació adequats per a utilitzar tècniques de *Machine Learning*, s'han desenvolupat diferents paquets per a **R** que ens permeten crear i analitzar models de classificació *Random Forest* d'una manera reproduïble i relativament senzilla, com "randomForest" (Liaw & Wiener, 2002) i "caret" (Kuhn *et al.*, 2019).

1.4 Planificació del Treball

A continuació, es presenten les tasques que s'assignen al projecte en qüestió i també el calendari amb el qual s'ha planificat per elaborar cadascuna d'aquestes. També s'especifiquen les fites del TFM.

1) Preparació de les dades
    - Del 15 al 21 d'octubre de 2019
2) Descripció dels gèneres i/o espècies presents al microbioma intestinal dels pacients:
    - Del 22 al 28 d'octubre de 2019
3) Elecció dels factors que seran utilitzats per crear un model de classificació
    - Del 29 d'octubre al 4 de novembre de 2019
4) Descripció de les metadades
    - Del 5 a l'11 de novembre de 2019

5) Entrenament del model de classificació de *Random Forest* per a la predicció de la probabilitat d'haver desenvolupat alguna malaltia autoimmunitària
    - Del 12 al 18 de novembre de 2019

FITA 1: Presentació del codi de preparació de les dades i l'utilitzat per a realitzar el primer objectiu a més d'un informe de seguiment del projecte. Inici de la discussió amb el director del TFM sobre els resultats obtinguts als primers entrenaments del model de *Random Forest*.

6) Avaluació del rendiment del model de classificació
    - Del 19 al 25 de novembre de 2019
7) Millora del model de classificació
    - Del 26 de novembre al 9 de desembre de 2019
8) Anàlisi de l'impacte de les diferents variables predictores del model de classificació
    - Del 10 al 16 de desembre de 2019

FITA 2: Presentació de les dades i resultats finals provisionals, a més d'un informe de seguiment del treball.

9) Memòria Final del TFM
    - Del 2 de desembre de 2019 al 6 de gener de 2020
10) Presentació Virtual del TFM
    - Del 17 de desembre de 2019 al 12 de gener de 2020

FITA 3: Presentació de la memòria final i elaboració de la presentació virtual del TFM.

A continuació, s'inclou un diagrama de Gantt a la *Figura 1* amb un resum de la temporalització per tasca i fita. El codi utilitzat es pot trobar a la secció "2. Temporalització del TFM – Diagrama de Gantt" del codi indicat a l'annex "6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM"; és una adaptació realitzada a partir d'un codi elaborat per Marcel Ramos (2015).

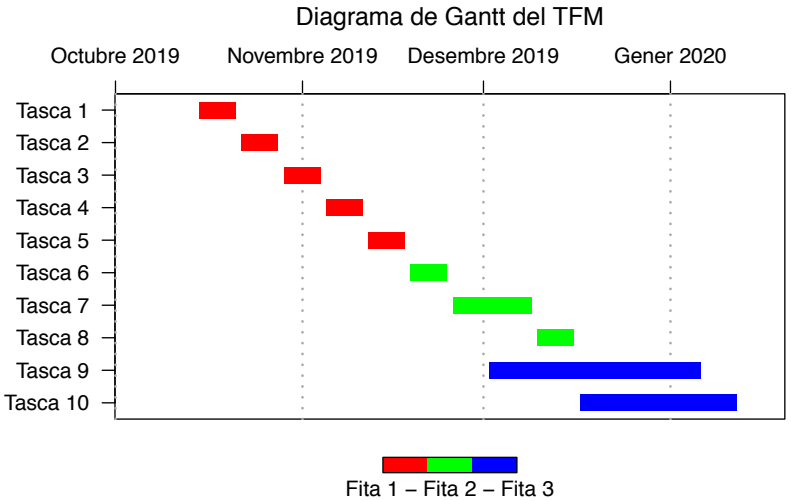


Figura 1. Diagrama de Gantt del TFM

1.5 Breu sumari de productes obtinguts

Els productes obtinguts a aquest Treball de Fi de Màster es descriuen, a continuació:

- Memòria del Treball de Fi de Màster: és el document present, on es descriu tot el procés realitzat durant el projecte i els resultats obtinguts.
- Codi utilitzat per dur a terme les diferents tasques plantejades i assolir els objectius del treball.
- Models de *Machine Learning* obtinguts utilitzant els algoritmes de *Random Forest*, *Support Vector Machine* i XGBoost per predir l'autoimmunitat dels pacients en funció del perfil taxonòmic del microbioma intestinal i altres metadades.

1.6 Breu descripció dels altres capítols de la memòria

Al **segon** capítol d'aquesta memòria es troba la descripció de la metodologia seguida en aquest projecte. En ell s'explica com s'han obtingut i preparat les dades, tant les taxonòmiques com les metadades. Després s'indica com s'han estandarditzat per analitzar-les posteriorment i poder ser utilitzades pels diferents algoritmes de *Machine Learning* per generar els models de classificació. També s'indiquen les variables clíniques seleccionades per la generació del model i per què s'han elegit. Seguidament, es detalla el procés d'entrenament i avaluació dels models de classificació utilitzant els algoritmes *Random Forest*, *Support Vector Machine* i XGBoost. Finalment, s'explica com s'ha analitzat l'impacte de les diferents variables predictores del model obtingut.

Al **tercer** capítol es mostren tots els resultats obtinguts durant els diferents passos d'aquest projecte i s'han discutit els valors obtinguts. En primer lloc, s'ha analitzat la taxonomia del microbioma intestinal. Seguidament, s'ha mostrat la distribució de les diferents metadades seleccionades per generar el model. Després, s'han mostrat, avaluat i analitzat els resultats obtinguts a tots els models generats utilitzant els algoritmes *Random Forest*, *Support Vector Machine* i XGBoost. Finalment, s'han analitzat els resultats de l'impacte que han tingut les diferents variables predictores del millor model de classificació obtingut.

Al **quart** capítol s'exposen les conclusions d'aquest treball, on també s'indica vies alternatives i futures per seguir investigant mitjançant els resultats obtinguts.

Al **cinqué** capítol del treball s'hi pot trobar un glossari amb la definició dels termes més rellevants utilitzats durant el desenvolupament del projecte.

Al **sisé** capítol es troba la bibliografia consultada en aquest treball i les referències a les diferents llibreries utilitzades durant la programació amb **R**.

Al **seté** capítol es poden trobar els annexos. Aquestos consten de: les llibreries d'**R** utilitzades; les figures de la descripció totes les metadades seleccionades; totes les matrius de confusió i estadístics d'avaluació dels classificadors SVM generats; els resultats de l'avaluació del rendiment dels models de classificació XGBoost utilitzant diferents mètodes de mostratge i paràmetres d'entrenament; i els enllaços referents a la ubicació del codi utilitzat en aquest projecte.

2. Metodologia

Els passos i mètodes seguits per assolir els objectius d'aquest projecte s'expliquen en els següents apartats. Per a realitzar aquest projecte s'ha utilitzat el llenguatge de programació **R** (R Core Team, 2019) i la seua interfície d'usuari **RStudio** (RStudio Team, 2016). Les llibreries d'**R** utilitzades durant tot el treball es troben en l'annex "6.1 Llibreries d'**R**".

2.1 Obtenció de les dades

Les dades amb informació sobre el perfil taxonòmic dels bacteris presents en l'intestí humà utilitzats en aquest projecte fan referència a mostres fecals obtingues en l'estudi PRJEB11419 de l'*American Gut Project* (McDonald *et al.*, 2018). D'aquestes mostres metagenòmiques s'han obtingut seqüències genètiques. Mitjançant un processament d'aquestes seqüències i mapant-les amb una base de dades personalitzada amb les seqüències de la subunitat 16 de l'ARN ribosòmic, Nana Teukam, en un treball de fi de grau (Nana Teukam, 2019) es va obtindre una taula OTU (*Operational Taxonomic Units*) que conté la composició microbiològica en percentatges de cada espècie per mostra. Les subunitats menudes 16S de l'ARN ribosòmic són marcadors genètics estàndards per identificar bacteris, ja que hi ha una gran quantitat d'eines i bases de dades de referència molt completes (Callahan *et al.*, 2016). Aquesta taula OTU esmentada està en format TSV (*Tab-Separated Values*) i el seu contingut es troba en un arxiu anomenat "genus.perc.txt". En la part superior de cada columna separada per un tabulador es troben els identificadors de les mostres mentre que el primer identificador de cada fila és el nom de l'espècie (o classificació taxonòmica) a la qual fan referència les dades de la mateixa fila.

Les dades cíniques o d'interés clínic, també anomenades metadades, estan emmagatzemades en arxius en format JSON (*JavaScript Object Notation*) que contenen informació sobre les variables clíniques de més de 16.000 pacients que pertanyen al mateix estudi que s'ha anomenat prèviament. Entre aquestes dades es troba informació com és l'edat dels pacients, índex de massa corporal, sexe, entre d'altres característiques. Aquestes dades s'han obtingut del mateix treball fi de grau (Nana Teukam, 2019), que a la seua vegada l'autor les va descarregar de la pàgina *BioSamples* de l'*European Bioinformatics Institute* (EBI) (Madeira *et al.*, 2019), fent servir l'API per descarregar exactament les mateixes mostres de què ja es disposava de dades taxonòmiques.

Per poder fer un filtre de les dades per número de *reads* de cada mostra, s'ha procedit a descarregar un arxiu de la base de dades d'EMBL-EBI del projecte esmentat anteriorment (amb accés el 13 de novembre de 2019; https://www.ebi.ac.uk/ena/data/view/PRJEB11419). En aquesta pàgina web s'ha clicat en "Select columns" i després s'ha marcat "Run accession" i "Read count". Per descarregar l'arxiu de mapatge, s'ha clicat en "TEXT" i s'ha guardat com a "PRJEB11419_Read_Count.tsv".

2.2 Preprocessament de les metadades

Les metadades, en primer lloc, s'han passat a format CSV (*Comma-Separated Values*). Aquest és un tipus de format d'emmagatzemament de dades estàndard i molt fàcil d'utilitzar. La transformació de les dades a aquest format permet que siga possible utilitzar-les emprant la majoria de llenguatges de programació amb relativa facilitat. Per realitzar aquesta transformació, s'ha obtingut una llista amb tots els arxius que contenen metadades; la informació d'aquestos arxius s'ha ajuntat utilitzant el paquet "rjson" de **R** (Couture-Beil, 2018) i s'ha guardat en format CSV en un arxiu anomenat "allmetadata.csv". En aquest arxiu, els noms de les files contenen informació dels diferents factors de les metadades mentre que els noms de les columnes són els identificadors de les mostres, que eren els noms dels arxius originals en format JSON. El codi utilitzat es pot trobar a la secció "3.1. Reestructuració de les metadades" del codi indicat a l'annex "6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM".

Una vegada s'ha disposat de les metadades en format CSV s'ha vist que moltes d'aquestes variables fan referència a la mostra en si (enllaços d'internet, diferents identificadors de la mostra, dates, organisme, etc.), o que contenen pràcticament la mateixa informació (per exemple: edat categòrica, edat categòrica en termes d'ontologia, edat corregida, etc.), o que contenen informació sobre la dieta dels pacients. Degut a que moltes d'aquestes variables no seran utilitzades de cap manera, s'han eliminat de la base de dades en el aquest pas de la preparació. Seguidament, s'han escurçat els noms de les diferents metadades (eliminant paraules com "characteristics" o "text"). Finalment, s'han guardat les metadades filtrades en un nou arxiu CSV anomenat "filteredmetadata.csv". El codi utilitzat per filtrar les metadades es pot trobar a la secció "3.2. Preparació de les metadades" del codi indicat a l'annex "6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM". A continuació, es poden veure els noms de les 30 primeres variables ordenades alfabèticament:

- acid_reflux
- acne_medication
- add_adhd
- age_corrected
- alcohol_consumption
- allergic_to_other
- allergic_to_peanuts
- allergic_to_shellfish
- allergic_to_tree_nuts
- allergic_to_unspecified
- altitude
- alzheimers
- animal_age
- animal_free

- animal_gender
- animal_origin
- animal_type
- antibiotic_history
- appendix_removed
- artificial_sweeteners
- asd
- assigned_from_geo
- autoimmune
- birth_year
- bmi_corrected
- bowel_movement_frequency
- bowel_movement_quality
- breastmilk_formula_ensure
- cancer
- cancer_treatment

2.3 Descripció de la taxonomia del microbioma intestinal

En 2011 es va publicar un article per part del consorci MetaHIT en què s'exposaven els diferents enterotips de microbioma intestinal humà. Els enterotips són estrats que representen els tipus de microbioma intestinal humà diferenciats per la seua composició taxonòmica utilitzant, per exemple, una Anàlisi Principal de Coordinades (PCoA). En aquest article es van utilitzar 33 mostres fecals i es van trobar 3 enterotips diferents. Aquests enterotips es caracteritzaven per una alta presència relativa de comptatges dels següents tres gèneres bacterians: *Bacteroides*, *Prevotella* i *Ruminococcus* (Arumugam *et al.*, 2011).

En aquest treball, s'ha realitzat una metodologia molt semblant a aquella utilitzada en l'article d'Arumugam *et al.* (2011) per observar si les mostres de què disposem presenten els mateixos enterotips.

En primer lloc, degut a la diversitat de la qualitat de les mostres, s'ha procedit a filtrar per utilitzar només aquelles mostres que tenen, almenys, 35.000 lectures de seqüenciació. Per extraure el número de mostres, s'ha utilitzat l'arxiu "PRJEB11419_Read_Count.tsv" que conté el número de lectures de cada mostra. S'ha filtrat els OTUs amb un nombre major de 35.000 lectures.

En segon lloc, s'han eliminat aquells gèneres microbiològics amb poca abundància; és a dir, s'han eliminat aquells gèneres que tenen una abundància mitjana menor del 0.01% en el total de les mostres disponibles.

Degut a la gran quantitat de dades que encara estaven disponibles després d'haver aplicat aquests dos filtres explicats anteriorment, s'ha obtingut una mostra aleatòria de la població amb la meitat de mostres.

Una vegada s'han preparat les dades taxonòmiques per ser analitzades, s'ha procedit a realitzar el PCoA. Per realitzar la clusterització de les mostres s'ha utilitzat una mesura de la distància de la distribució de la probabilitat relacionada amb les distàncies Jensen-Shannon *divergence* (JSD). Les distàncies JSD, en aquest context, s'empren per saber com de semblants són les diferents mostres entre sí (Endres & Schindelin, 2003).

L'algoritme de clusterització utilitzat és el *Partitioning Around Medoids* (PAM), que s'ha generat a partir de l'algoritme de les *k*-mitjanes amb l'avantatge de que és més robust que aquest últim i, a més, permet la inclusió de mesures de distància arbitràries. Aquest algoritme utilitza un número predeterminat de clústers (*k*) com a input per poder realitzar la partició de les dades. En aquest treball, per començar ens hem basat en el número de 3 clústers, tal i com s'ha realitzat a l'article d'Arumugam *et al.* (2011). Seguidament, per calcular el número òptim de clústers s'ha utilitzat l'índex *Caliński-Harabasz* (CH), que indica el bon rendiment a l'hora de realitzar diferents números de clústers (Caliński & Harabasz, 1974).

S'han realitzat dos gràfics amb els PCoAs: un amb els 3 enterotips (3 clústers) tal i com s'havia realitzat a l'article d'Arumugam *et al.* (2011) i un altre amb el número d'enterotips indicats pel millor número de clústers (*k*) mostrat per la figura dels resultats de l'índex CH, que s'ha calculat per a unes *k*s des de 2 fins 15.

Finalment, per possibilitar una comparació amb els resultats obtinguts a la bibliografia, s'han realitzat uns diagrames de caixes amb el percentatge d'aparició en les mostres dels gèneres més abundants als 3 enterotips obtinguts (utilitzant $k = 3$). Algunes funcions i instruccions d'**R** s'han obtingut del codi publicat per Bork (2011) a l'enllaç https://enterotype.embl.de/enterotypes.html, utilitzat a la publicació d'Arumugam *et al.* (2011). El codi utilitzat per descriure la taxonomia del microbioma intestinal es pot trobar a la secció "4.1. Descripció de la taxonomia del microbioma intestinal" del codi indicat a l'annex "6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM".

2.4 Estandardització de les metadades

Una vegada s'han filtrat les metadades tal i com s'ha explicat a la secció 2.2, i s'ha realitzat una descripció de la taxonomia del microbioma intestinal, s'ha procedit a visualitzar per primera vegada el contingut de les metadades. Amb aquesta visualització s'ha pogut comprovar que les variables en què es troben les metadades no han estat estandarditzades prèviament. Per aquest motiu, s'ha procedit a modificar aquesta base de dades amb les variables clíniques.

En primer lloc, s'ha creat una funció que modifica les variables que són factors, i per tant categòriques, estandarditzant els valors d'algunes

variables dicotòmiques, de freqüència, traduint alguns valors al català i posant com a valors faltants algunes respostes indefinides o sense criteri científic. A continuació, es disposen els valors originals i la codificació que s'ha realitzat per estandarditzar-los.

- Variables dicotòmiques:
  - "true": codificat com a "Sí".
  - "false": codificat com a "No".
- Variables de freqüència:
  - "Daily": codificat com a "1. Molt alta".
  - "Regularly (3-5 times/week)": codificat com a "2. Alta".
  - "Occasionally (1-2 times/week)": codificat com a "3. Mitjana".
  - "Rarely (a few times/month)": codificat com a "4. Baixa".
  - "Rarely (less than once/week)": codificat com a "4. Baixa".
  - "Never": codificat com a "5. Molt baixa".
- Traduccions:
  - "Yes": traduït com a "Sí".
  - "Diagnosed by a medical professional (doctor, physician assistant)": traduït com a "Diagnosticat".
  - "I do not have this condition": traduït com a "Sense aquesta condició".
  - "African American": traduït com a "Afroamericà".
  - "Asian or Pacific Islander": traduït com a "Asiàtic".
  - "Caucasian": traduït com a "Caucàsic".
  - "Hispanic": traduït com a "Llatinoamericà".
  - "female": traduït com a "Dona".
  - "male": traduït com a "Home".
- Respostes indefinides o sense criteri científic, que s'han codificat com a "NA" (un valor nul).
  - "Other" i "other".
  - "Not provided".
  - "Unspecified" i "unspecified".
  - "Diagnosed by an alternative medicine practitioner".
  - "Self-diagnosed".
  - "Not sure".
  - "Not applicable".
  - "Unknown".

Aquesta funció s'ha aplicat a les variables categòriques de la base de dades amb les variables clíniques.

En segon lloc, a l'hora d'importar les dades de les variables amb números a partir de l'arxiu CSV, **R** els ha considerat com a factors. Per aquest motiu, aquestes variables s'han passat de format categòric amb factors a numèric. El codi utilitzat es pot trobar a la secció "4.2.1. Estandardització de les variables d'interés clínic" del codi indicat a l'annex "6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM".

2.5 Selecció de les metadades per entrenar el model de classificació

La selecció de les metadades per entrenar el model de classificació ha sigut necessària ja que no seria viable entrenar un model de classificació utilitzant totes les variables presents en les metadades. Per seleccionar unes variables específiques, s'ha realitzat una recerca bibliogràfica. A continuació, s'expliquen els motius de per què s'han seleccionat cadascuna de les variables en concret de les metadadades per desenvolupar el model de classificació:

- Malaltia autoimmunitària:
  Aquesta variable és considerada com a resposta del model de classificació que es genera en aquest projecte i, per tant, serà analitzada. Aquesta variable conté informació de si una persona ha sigut, o no, diagnosticada amb alguna malaltia autoimmunitària per part d'un professional mèdic.
- Edat:
  Se sap que l'edat és un important risc per a l'autoimmunitat, ja que moltes de les malalties d'aquest tipus es desenvolupen predominantment en la segona meitat de l'etapa adulta, quan la competència immune ha disminuït i la generació de limfòcits T ha finalitzat (Goronzy & Weyand, 2012).
- Sexe:
  La majoria de les malalties autoimmunitàries afecten predominantment a les dones, la qual cosa indica un biaix donat pel sexe. Alguns factors que el provoquen podrien ser les hormones sexuals, la presència o l'absència del segon cromosoma X i part del microbioma específic del sexe (Rubtsova *et al.*, 2015).
- Índex de massa corporal:
  En un estudi s'han obtingut uns resultats que mostren un augment del risc de patir malalties autoimmunitàries en dones amb un índex de massa corporal alt, respecte a les dones que en tenen un de normal (Harpsøe *et al.*, 2014).
- Raça:
  En un estudi publicat en 2011 es va vore que el gen *FMR1*, que es troba en una àrea del cromosoma X que està bastant associada amb l'autoimmunitat i la reserva ovàrica. Es van analitzar dones de diferents races (caucàsiques, africanes i asiàtiques) per saber si hi havia associacions entre les oportunitats de ser fertilitzades *in vitro*, el genotip del gen *FMR1*, la raça i l'autoimmunitat. Es va concloure que si s'utilitzava la raça es podia explicar millor el resultat d'una fecundació *in vitro*, segons el genotip *FMR1* (Gleicher *et al.*, 2011).
- Medicaments per a l'acne:
  Alguns medicaments contra l'acne, com la minociclina, s'han relacionat amb l'aparició d'esdeveniments adversos que han

donat lloc a reaccions autoimmunitària, com per exemple l'hepatitis autoimmunitària (Eichenfield, 1999).

- Al·lèrgia als cacaus, al marisc, estacional i malalties pulmonars:
  La família dels gens *TIM* està associada al desenvolupament d'al·lèrgies alimentàries, rinitis al·lèrgica (símptoma molt comú quan se sofreix al·lèrgia estacional), asma i malalties autoimmunitàries (Li *et al.*, 2013; Angiari & Constantin, 2014).
- Al·lèrgia a vegetals com l'heura o el roure, a la llum intensa del Sol i malalties de la pell:
  En un article publicat per Sharma i Bayry (2015) es nomena que, a més de que els basòfils tinguen un paper important en malalties al·lèrgiques de la pell i l'asma, també pareix que estan relacionats amb la patogènesi de malalties autoimmunitàries.
- Convivència amb gats i gossos, i al·lèrgia al pèl animal:
  En 2014, es van publicar els resultats d'un estudi realitzat per saber si el contacte amb animals durant la infantesa s'associava amb el desenvolupament de la diabetis tipus 1, tant clínica com preclínica. Els resultats van mostrar que una convivència amb gossos reduïa la probabilitat de desenvolupar aquesta malaltia (Virtanen *et al.*, 2014).
- Amígdales eliminades:
  Segons un estudi publicat en 2016, la incidència d'algunes malalties autoimmunitàries va ser major en els pacients a qui se'ls havia eliminat les amígdales (Ji *et al.*, 2016).
- Freqüència de fumar i consumició d'alcohol:
  Segons un article publicat per D'hooghe i altres investigadors (2012), tant fumar tabac com la consumició d'alcohol (i d'altres substàncies) i la progressió de l'esclerosi múltiple en els pacients està relacionada.

Amb tota aquesta informació, a continuació, es presenta el llistat de les variables que s'han utilitzat per entrenar el model de classificació:

- Malaltia autoimmunitària (variable "autoimmune")
- Edat (variable "age_corrected")
- Sexe (variable "sex")
- Índex de massa corporal (variable "bmi_corrected")
- Raça (variable "race")
- Medicació utilitzada per a l'acne (variable "acne_medication")
- Al·lèrgia als cacaus (variable "allergic_to_peanuts")
- Al·lèrgia al marisc (variable "allergic_to_shellfish")
- Al·lèrgia estacional (variable "seasonal_allergies")
- Malaltia pulmonar (variable "lung_disease")
- Al·lèrgia als vegetals com l'heura o el roure (variable "non_food_allergies_poison_ivyoak")
- Al·lèrgia a la llum intensa del Sol (variable "non_food_allergies_sun")
- Malalties de la pell (variable "skin_condition")
- Convivència amb algun gat (variable "cat")

- Convivència amb algun gos (variable “dog”)
- Al·lèrgia al pèl animal (variable “non_food_allergies_pet_dander”)
- Amígdales eliminades (variable “tonsils_removed”)
- Freqüència de fumar (variable “smoking_frequency”)
- Consumició d’alcohol (variable “alcohol_consumption”)

Una vegada s’ha obtingut aquest llistat amb les variables seleccionades, s’ha generat un arxiu en format CSV que conté només les metadades presents en aquestes variables, anomenat “selectedmetadata.csv”. El codi utilitzat es pot trobar a la secció “4.2.2. Selecció i descripció de les variables d’interés clínic” del codi indicat a l’annex “6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM”.

2.6 Descripció de les variables seleccionades

La descripció de les metadades seleccionades ha consistit en realitzar una sèrie de figures: diagrames de caixes i histogrames per les variables numèriques; diagrames de barres per les variables categòriques. El codi utilitzat es pot trobar a la secció “4.2.2. Selecció i descripció de les variables d’interés clínic” del codi indicat a l’annex “6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM”.

2.7 Mètodes de mostratge

Les classes disponibles a la variable resposta “autoimmune”, en principi, haurien d’estar equilibrades a l’hora d’entrenar un model. Per obtindre un número igualat de les diferents classes s’han utilitzat tres mètodes de mostratge: *up-sampling*, *down-sampling* i SMOTE.

- *Up-sampling*: Realitza una mostra aleatòria (amb reemplaçament) de la classe minoritària perquè aquesta siga igual de freqüent que la classe majoritària.
- *Down-sampling*: Realitza mostres aleatòries de les dues classes perquè tinguen la mateixa freqüència, obtenint un nombre menor de mostres d’ambdues classes.
- SMOTE (de l’anglés, *Synthetic Minority Over-sampling Technique*): És un mètode híbrid que el que fa és obtindre una mostra aleatòria de la classe majoritària i genera noves mostres a partir de la classe minoritària (Kuhn, 2019).

2.8 Generació de models de classificació de *Random Forest*

La generació d’un model de classificació mitjançant el *Machine Learning* es pot realitzar utilitzant diferents tipus d’algoritmes. Els arbres de decisió, per exemple, són un tipus d’algoritme que es basa en anar classificant la variable resposta segons xicotetes decisions, seguint una estructura d’un arbre, preses a partir de les característiques introduïdes a les dades de l’*input*. Un exemple d’arbre de decisió seria aquell que prediu si es deuria

acceptar una oferta de feina a partir de diferents paràmetres; es comença a partir d'un node arrel, que passa als nodes de decisió que requereixen eleccions basades en les característiques de l'oferta. Finalment, es prén una decisió, que acaba en els nodes terminals, que denoten l'acció a prendre com a resultat de la sèrie de decisions. A continuació, a la *Figura 2* es mostra aquest exemple d'arbre de decisió extreta del llibre de Lantz (2015):

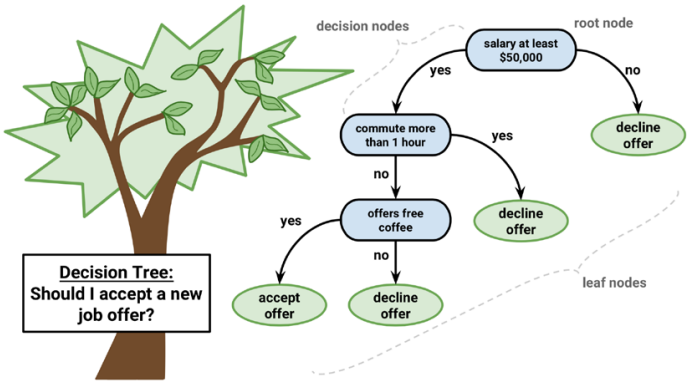


Figura 2. Exemple d'un arbre de decisió. Extret de Lantz (2015).

Aquest algoritme dels arbres de decisió es construeix dividint les dades en *subsets* basant-se en la variable més predictora de la classe *objectiu*. Aquest *subset*, a la seua vegada es divideix en més *subsets* utilitzant altres variables que divideixen al màxim possible les dades segons les seues característiques. L'algoritme va seguint aquest procés fins que determina que els *subsets* són el suficientment homogenis, o fins que es complisca algun criteri de parada (Lantz, 2015).

L'algoritme de *Random Forest* es basa en la unió de molts arbres de decisió. El gran avantatge d'aquest algoritme és que no sobreajusta tant el model a les dades com els arbres de decisió. A més, aquest tipus d'algoritme pot utilitzar variables amb soroll i tant característiques numèriques com categòriques, seleccionant els factors més importants (Lantz, 2015). En aquest punt del projecte, s'ha entrenat un model classificador *Random Forest* per a poder calcular les probabilitats que els pacients hagen desenvolupat alguna malaltia autoimmunitària. Aquest classificador utilitza com a dades d'entrenament els perfils taxonòmics del microbioma intestinal, les dades clíniques i altres tipus de dades d'aquestos pacients.

En primer lloc, s'han realitzat els mateixos filtratges que els utilitzats a la secció 2.3 d'aquest treball: s'han eliminat aquelles mostres que tenen menys de 35.000 lectures de seqüenciació i s'han eliminat aquells gèneres microbiològics amb poca abundància.

En segon lloc, s'han ajuntat les dades taxonòmiques i les metadades utilitzant els identificadors de les mostres. Seguidament, s'han eliminat aquelles mostres que tenen algun valor faltant en alguna de les variables clíniques seleccionades o en algun OTU. D'aquesta manera han quedat 225 variables i 1809 mostres per poder general el model. S'han desat totes aquestes dades combinades i sense valors faltants a un arxiu anomenat

“OTU_Metadata_processed.csv”. El codi utilitzat es pot trobar a la secció “2.1. Adequació de les dades de la taula OTU i metadades seleccionades” del codi indicat a l’annex “6.6 Codi de les tasques corresponents al segon objectiu del TFM”.

El següent pas ha estat l’obtenció dels data sets d’entrenament i de test. Per això, primer s’han normalitzat els valors de l’edat i de l’índex de massa corporal. La normalització utilitzada és assignar-li un valor de 0 al valor mínim de cada variable i un 100 als respectius valors màxims, sent els valors entremitjos números entre 0 i 100 relacionats proporcionalment al valor original. Després, s’han generat una partició de dades d’un 67% per al data set d’entrenament i un 33% per al data set de test, utilitzant la funció “createDataPartition()” de la llibreria “caret” (Kuhn *et al.*, 2019). A aquesta funció se li ha indicat que la variable resposta de la classificació serà “autoimmune”, pel que té en compte que els dos data sets d’eixida tinguen un nombre proporcional i equilibrat de valors de les diferents categories d’aquesta variable. No obstant això, en un principi no s’ha utilitzat cap mètode de mostratge. El codi utilitzat es pot trobar a la secció “2.2. Obtenció dels datasets d’entrenament i de test” del codi indicat a l’annex “6.6 Codi de les tasques corresponents al segon objectiu del TFM”.

El següent pas és l’entrenament del primer model de *Random Forest*. S’ha utilitzat la funció “randomForest()” del paquet “randomForest” (Liaw & Wiener, 2002). El codi utilitzat es pot trobar a la secció “2.3. Entrenament del model” del codi indicat a l’annex “6.6 Codi de les tasques corresponents al segon objectiu del TFM”.

Seguidament, s’ha realitzat una avaluació del model generat. Per a això, utilitzant la funció “predict()” de la llibreria “stats”, s’ha realitzat una predicció de les classes del data set de test a partir de les variables predictores d’aquest mateix data set. Després, s’ha generat una matriu de confusió utilitzant la funció “confusionMatrix()” del paquet “caret” (Kuhn *et al.*, 2019). El codi utilitzat es pot trobar a la secció “2.4. Avaluació de l’algoritme” del codi indicat a l’annex “6.6 Codi de les tasques corresponents al segon objectiu del TFM”. Una matriu de confusió és una taula que categoritza les prediccions, segons si han predit correctament, o no, el valor real. La classe positiva és aquella que ens interessa, és a dir, en aquest cas seria que un pacient ha estat diagnosticat amb una malaltia autoimmunitària.

Una vegada s’ha generat el primer model amb les opcions per defecte, s’ha començat amb la millora del mateix. Un dels paràmetres d’entrada que té la funció per generar el model de *Random Forest* és el número d’arbres (paràmetre “ntree”). Per defecte, aquesta funció utilitza “ntree = 500”; en primer lloc, per millorar el model s’han provat, també, les opcions “ntree = 1000” i “ntree = 10000”.

Una altra manera d’intentar millorar l’entrenament del model és generar uns data sets d’entrenament i test amb els valors de la variable resposta

més equilibrat. Per això, s'han utilitzat els diferents mètodes de mostratge indicats a la secció anterior: *up-sampling*, *down-sampling* i SMOTE.

El primer mètode de mostratge utilitzat ha sigut el SMOTE (amb la funció "SMOTE" de la llibreria "DMwR" (Torgo, 2011)). Després s'han utilitzat, també, els mètodes *up-sampling* i *down-sampling* (amb les funcions "upSample" i "downSample" de la llibreria "caret" (Kuhn *et al.*, 2019)). A més d'utilitzar els tres mètodes de mostratge, també s'han fet servir els tres valors anteriorment mencionats per al paràmetre "ntree" (500, 1000 i 10000). El codi utilitzat es pot trobar a la secció "2.5. Millora del model" del codi indicat a l'annex "6.6 Codi de les tasques corresponents al segon objectiu del TFM".

2.9 Generació de models de classificació *Support Vector Machine*

El següent pas per obtindre un millor model que els obtinguts utilitzant l'algoritme de *Random Forest* és provar amb l'algoritme *Support Vector Machine* (SVM). Per realitzar l'entrenament d'aquest model, cal generar abans unes variables *dummy* d'aquelles metadades que són categòriques presents en la base de dades, excepte les variables de freqüència, que s'han transformat a numèriques. La variable resposta s'ha transformat en una variable binària. Una de les llibreries utilitzades ha sigut la de "fastDummies" (Kaplan, 2019).

Un SVM és un algoritme que genera un objecte semblant a una superfície que crea uns límits entre els punts de les dades representats multidimensionalment que separa els seus valors utilitzant les diferents característiques o variables presents a la base de dades. L'objectiu d'un SVM és crear un hiperplà que divideix l'espai per crear particions homogènies a cada costat. Normalment hi ha més d'un hiperplà o línia que separe els diferents grups. Per aquest motiu, s'utilitza l'hiperplà de marge màxim (en anglés: *Maximum Margin Hyperplane*, MMH) que crea la separació més gran entre les dues classes de la variable resposta. Els *support vectors* són els punts dels que cada classe estan més prop dels MMH. Per aquest motiu, cada classe ha de tindre, almenys, un *support vector*. A continuació, en la *Figura 3*, es mostra un exemple de diferents línies que poden separar les dades, el marge màxim i els *support vectors*, una imatge extreta de Lantz (2015).

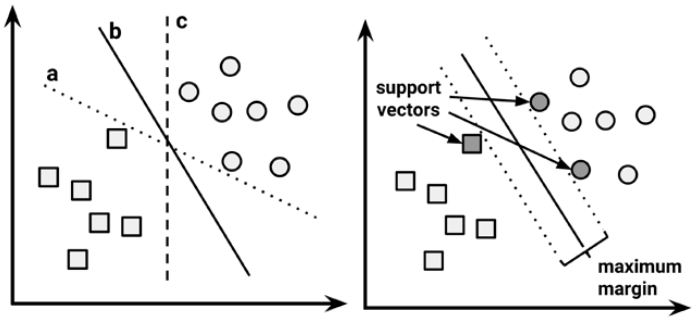


Figura 3. Línies separadores de dades, línia amb un marge màxim i support vectors. Extret de Lantz (2015).

La majoria d'aplicacions reals tenen una relació entre les variables que no és lineal. Els SVMs poden mapar el problema en un espai de moltes dimensions, utilitzant un procés anomenat *kernel trick*. Amb això, una relació no-lineal es pot transformar en una lineal. A continuació, en la *Figura 4*, s'observa un exemple d'aquest tipus de transformació, extret de Lantz (2015):

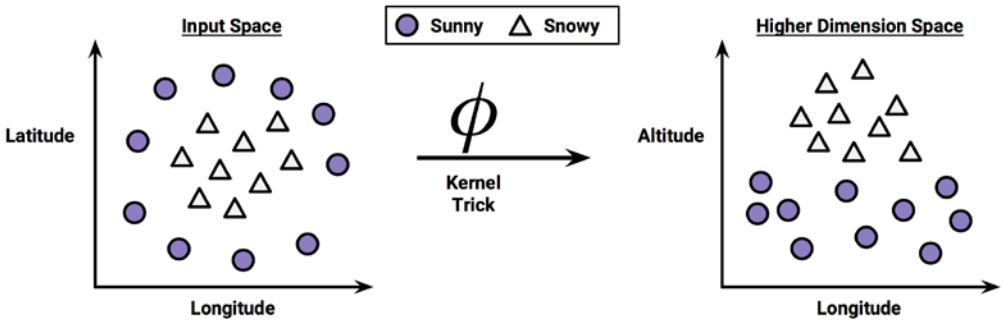


Figura 4. Representació de dades relacionades amb característiques d'una forma no-lineal utilitzant el "kernel trick" per poder representar la relació d'una forma lineal. Extret de Lantz (2015).

Els SVMs utilitzant *kernels* no-lineals són classificadors extremadament potents: es poden utilitzar en problemes de classificació o de predicció numèrica, pot treballar utilitzant dades amb soroll sense sobre ajustar. No obstant, és molt complicat trobar un bon model, ja que s'han de provar molts *kernels* i diferents combinacions de paràmetres d'entrada per a l'entrenament del model, són relativament lents a l'hora de ser entrenats.

Els *kernels* utilitzats en aquest projecte són els lineals i els gaussians. Aquestos primers no transformen les dades. Els *kernels* gaussians sí que transformen les dades i solen tindre un bon rendiment en molts tipus de dades; per això s'utilitzen molt com a punt de partida en els projectes (Lantz, 2015). A més d'utilitzar aquest dos tipus de *kernels*, s'empren també els mètode de mostratge SMOTE per generar els data sets d'entrenament i test, com també les dades sense utilitzar cap mètode dels esmentats anteriorment. El codi utilitzat es pot trobar a la secció "3. Generació del model de classificació Support Vector Machine" del codi indicat a l'annex "6.6 Codi de les tasques corresponents al segon objectiu del TFM".

2.10 Generació de models de classificació de XGBoost

L'últim algoritme de *Machine Learning* utilitzat en aquest treball ha estat el XGBoost. Aquest nom fa referència a *Extreme Gradient Boosting*, en nglés. Aquest algoritme utilitza un modelatge anomenat *gradient boosting*. El *boosting* és un procés seqüencial: els arbres van generant-se utilitzant els arbres anteriors, un rere un altre, en les diferents iteracions. Aquest procés va aprenent de les dades i intenta millorar la predicció realitzada pel model en les iteracions subseqüents. En la *Figura 5*, extreta de Saraswat (2019), s'hi representa la idea general dels algoritmes de *boosting*.

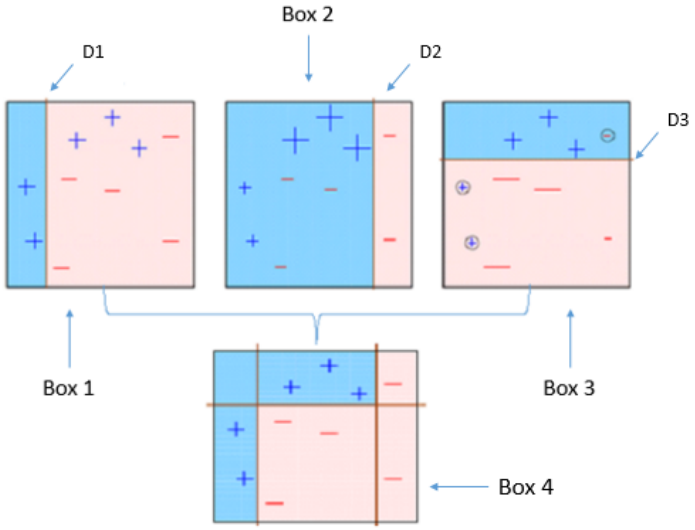


Figura 5. Algoritmes de boosting. Mitjançant la combinació de diferents classificadors de signes positius i negatius s'ha obtingut un model que classifica totes les dades perfectament. Extret de Saraswat (2019).

L'algoritme de XGBoost utilitza la regularització per evitar el sobreajustament dels models a les dades d'entrenament. A més, utilitza el *gradient descent*, que és un mètode que comprimeix un vector de pesos (o coeficients) per minimitzar l'error de predicció. Un altre dels avantatges que té és que pot utilitzar la computació paral·lela, és a dir, utilitzar diferents *cores* a la vegada. Aquest algoritme, també és un algoritme basat en arbres, com el *Random Forest* (Saraswat, 2019).

Els paràmetres utilitzats per generar els models de classificació en aquest projecte són "nrounds" i "max.depth":

- "nrounds": controla el número màxim d'iteracions.
- "max.depth": controla la profunditat de l'arbre. Una major profunditat implica un model més complex, implicant una possibilitat major de sobreajustament del mateix (Saraswat, 2019; XGBoost developers, 2019).

Per poder introduir la base de dades en la funció "xgboost" de la llibreria "xgboost" (Chen *et al.*, 2019) per generar el model de classificació, primer s'han transformat els data sets, sense utilitzar cap mètode de mostratge i utilitzant el de SMOTE, a matrius esparses mitjançant una funció generada per Ishihara (2019) anomenada "data.frame.2.sparseMatrix()". Aquesta funció, també transforma els factors a variables *dummy*. Una vegada s'ha modificat la base de dades, s'han generat molts models de classificació XGBoost amb els següents valors per als paràmetres d'entrada:

- "nrounds": 10, 50, 150, 300, 400, 500, 600, 1000.
- "max.depth": 1, 2, 3, 4, 5, 6, 7, 8.

El codi utilitzat es pot trobar a la secció “4. Generació del model de classificació XGBOOST” del codi indicat a l’annex “6.6 Codi de les tasques corresponents al segon objectiu del TFM”.

2.11 Avaluació dels models

Els resultats obtinguts a partir del codi utilitzat per desenvolupar els models, tal i com s’ha detallat anteriorment, genera matrius de confusió i també obté una sèrie d’estadístics, entre ells, el valor Kappa. Per avaluar el rendiment dels models obtinguts prèviament, cal tindre en compte diferents aspectes. D’una banda, cal observar les matrius de confusió on es categoritza les prediccions per indicar si s’han predit correctament els valors de les dades reals. En una matriu de confusió 2x2 es pot saber si alguna de les prediccions pertany a una de les quatre categories:

- Vertader positiu: ha estat correctament classificat en la classe d’interés.
- Vertader negatiu: ha estat correctament classificat en la classe que no és d’interés.
- Fals positiu: ha estat incorrectament classificat en la classe d’interés.
- Fals negatiu: ha estat incorrectament classificat en la classe que no és d’interés.

Sabent aquestes quatre categories es poden calcular diferents paràmetres per avaluar el rendiment d’un model de classificació, com poden ser l’exactitud (*accuracy* en anglés), taxa d’error, sensibilitat, especificitat, entre d’altres. Aquest tipus de valors s’obtenen també en els resultats mostrats per la funció “confusionMatrix()” que s’acaba de mencionar (Lantz, 2015).

D’altra banda, l’estadístic Kappa ajusta l’exactitud tenint en compte les diferents proporcions dels les classes. És molt important, sobre tot quan es disposa de bases de dades amb moltes classes en diferents proporcions, ja que un classificador pot obtindre molta exactitud simplement elegint la classe més freqüent. L’estadístic Kappa premiarà a aquells classificadors si són correctes més vegades de les que ho serien simplement per estratègia de freqüències. Els valors d’aquest estadístic van de 0 a 1, on 0 vol dir que les prediccions no es corresponen amb els valors reals i 1 significa que són iguals (Lantz, 2015). Els valors de Kappa es poden interpretar de la següent manera: de 0.01 a 0.20 no hi ha quasi gens d’adequació entre les dades originals i les classificacions obtingudes, de 0.21 a 0.40 l’adequació és justa, de 0.41 a 0.60 l’adequació és moderada, de 0.61 a 0.80 l’adequació és substancial i de 0.81 a 1.00 és quasi perfecta (McHugh, 2012). Aquestos valors són molt subjectius ja que, un valor de 0.6-0.8 pot ser un bon valor si es vol predir, per exemple, el sabor d’un menjar a partir d’uns ingredients i unes proporcions, mentre que un valor de 0.8-0.99 pot no ser suficient si es vol predir algun defecte de naixement amb un test genètic. Aquest valor també el calcula la funció “confusionMatrix()” automàticament (Lantz, 2015). Per tots aquestos motius, l’estadístic Kappa és el que s’utilitza per elegir el millor model.

2.12 Obtenció dels resultats de tots els models de classificació

En aquest apartat, en primer lloc, s'ha generat una taula que mostra les dimensions dels *data sets* generats utilitzant els diferents mètodes de mostratge.

Durant la generació de tots els models, s'ha anat avaluant a la vegada el rendiment d'aquests i s'ha anat guardant tota aquesta informació en diferents variables. En aquest punt del treball, en segon lloc, s'ha ajuntat tota aquesta informació. S'han generat taules per a cada algoritme de *Machine Learning* utilitzat amb els resultats respectius, on s'hi mostra la informació del mètode de mostratge, els paràmetres d'entrenament, la exactitud (*accuracy* en anglés), els valors Kappa, la sensibilitat i l'especificitat dels models. Després, s'ha mostrat, també, la matriu de confusió i tots els estadístics calculats del millor algoritme de *Random Forest*, de tots els SVMs generats i del millor model generat amb XGBoost.

El codi utilitzat es pot trobar a la secció "5. Resum dels resultats dels models de classificació" del codi indicat a l'annex "6.6 Codi de les tasques corresponents al segon objectiu del TFM".

2.13 Càlcul de la importància relativa de les variables predictores

L'última tasca en els objectius d'aquest projecte ha estat l'anàlisi de l'impacte de les diferents característiques predictores en el model de classificació. Amb la finalitat de mostrar les variables que més importància relativa han tingut en el model, s'ha generat una figura amb les 25 variables més importants, mitjançant un diagrama de barres. Aquestes variables consten dels diferents gèneres analitzats i les metadades més importants utilitzades en el model. La matriu d'importància s'ha desat en un arxiu en format CSV anomenat "matriu_importancia.csv".

El codi utilitzat es pot trobar a la secció "6. Anàlisi de l'impacte de les diferents característiques predictores" del codi indicat a l'annex "6.6 Codi de les tasques corresponents al segon objectiu del TFM".

3. Resultats i Discussió

3.1 Taxonomia del microbioma intestinal

Tal i com es va realitzar a l'article d'Arumugam *et al.* (2011), en la *Figura 6* es mostren els resultats del PCoA utilitzant *k* = 3 en l'algoritme PAM. En aquesta representació, s'observa que els tres clústers estan relativament separats. No obstant això, no hi ha una separació clara entre els tres enterotips.

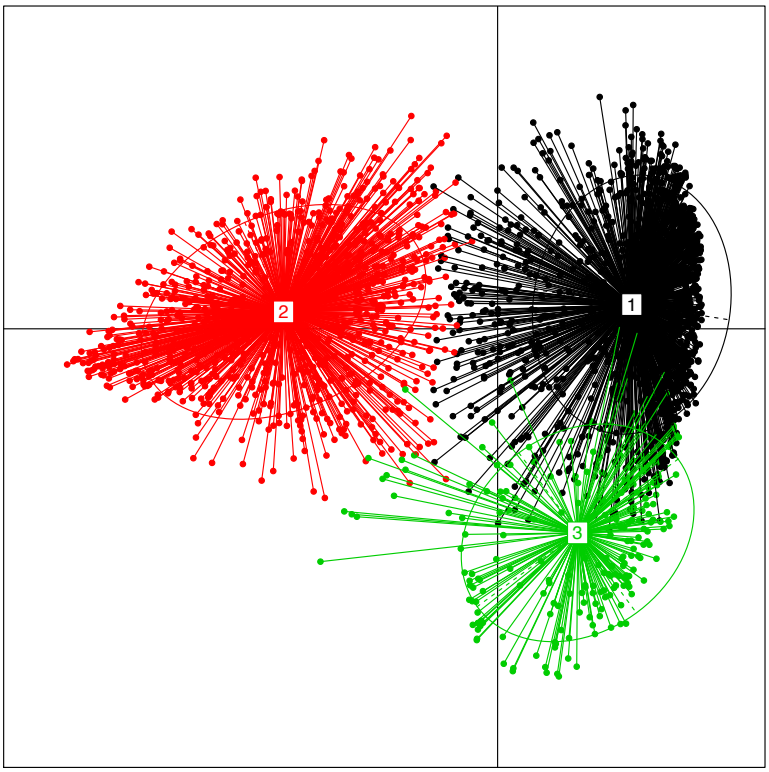


Figura 6. PCoA de les dades taxonòmiques del microbioma intestinal utilitzant k = 3 en l'algoritme PAM.

Els índex CH calculats són els representats en la *Figura 7*. Els resultats mostren un número òptim de clústers de *k* = 5, ja que és la *k* amb un índex CH calculat més alt.

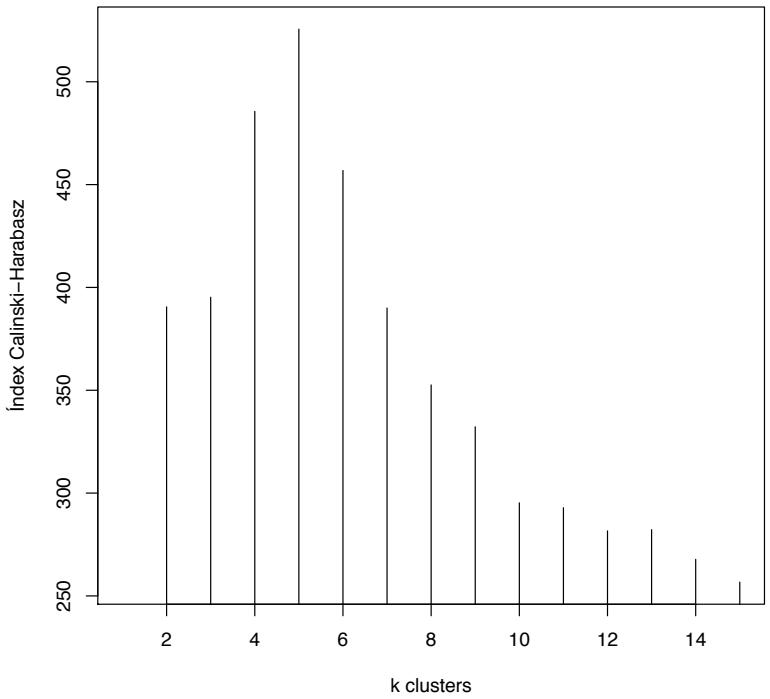


Figura 7. Índex Caliński-Harabasz per a k clústers, sent k un número enter des de 2 fins 15.

A la *Figura 8* es mostra el PCoA obtinguts a l'hora d'indicar-li a l'algoritme PAM el número de clústers òptim segons l'índex CH, és a dir, *k* = 5. En aquest gràfic, amb *k* = 5, s'observa més clarament que no hi ha una separació vertadera entre tots els enterotips, encara que els diferents índex CH calculats indicaven que el millor *k*, en aquest cas, era 5.

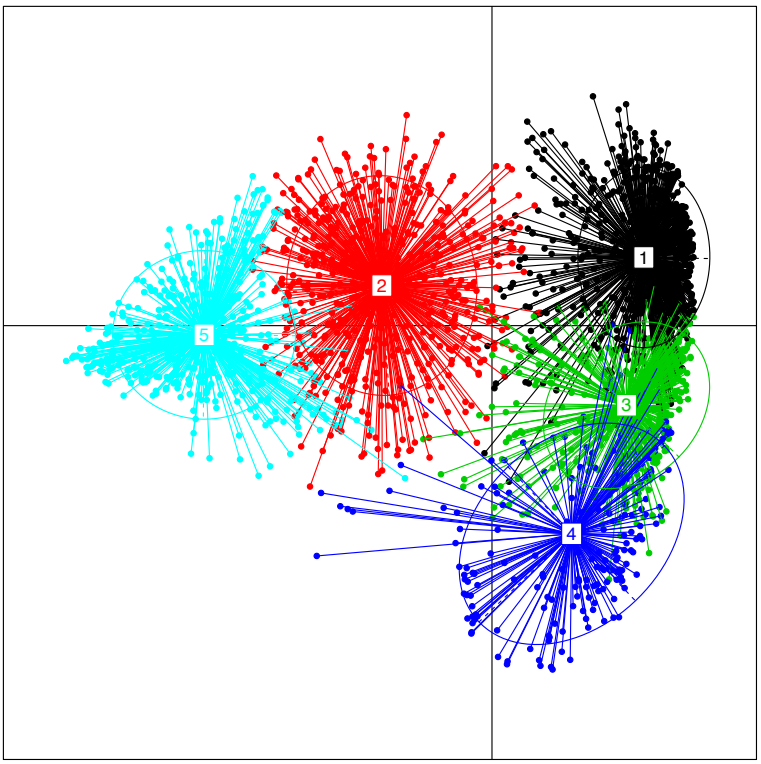


Figura 8. PCoA de les dades taxonòmiques del microbioma intestinal utilitzant k = 5 en l'algoritme PAM.

Per poder comparar els resultats d'aquest projecte amb els obtinguts a l'article d'Arumugam *et al.* (2011), i també perquè pareix que s'han obtingut uns millors resultats amb una clusterització de les mostres utilitzant k = 3, es mostren els percentatges d'abundància dels gèneres més abundants en els tres enterotips obtinguts en la *Figura 9*, la *Figura 10* i la *Figura 11*.

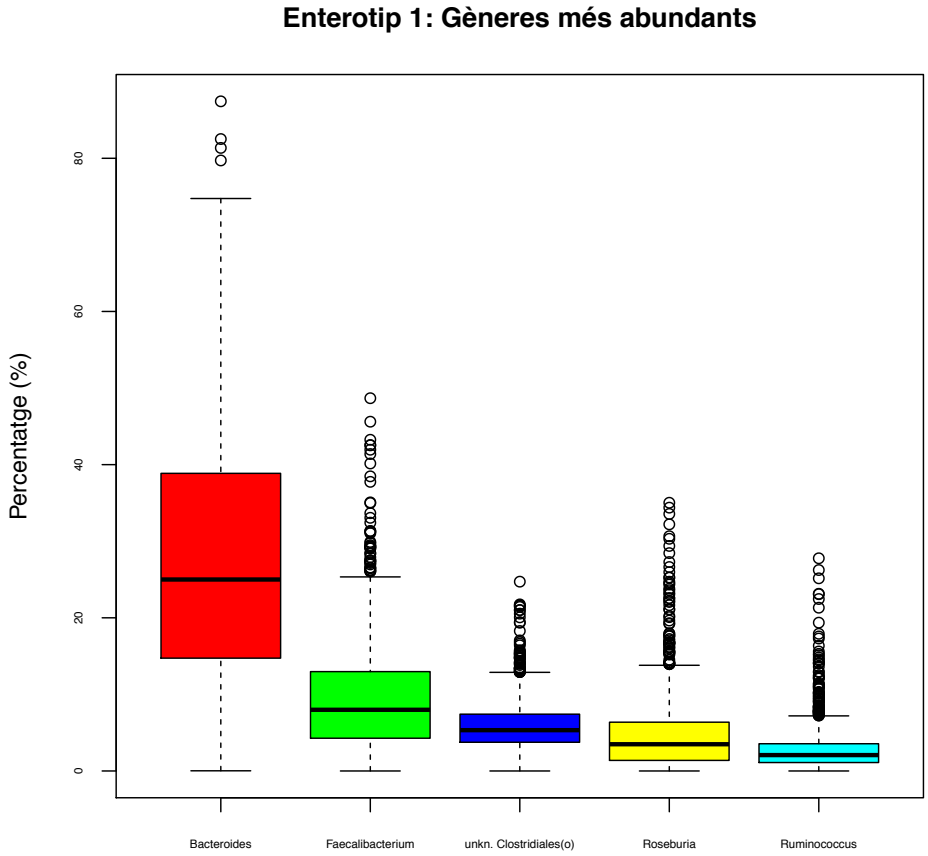


Figura 9. Gèneres més abundants a l'enterotip 1 del microbioma intestinal.

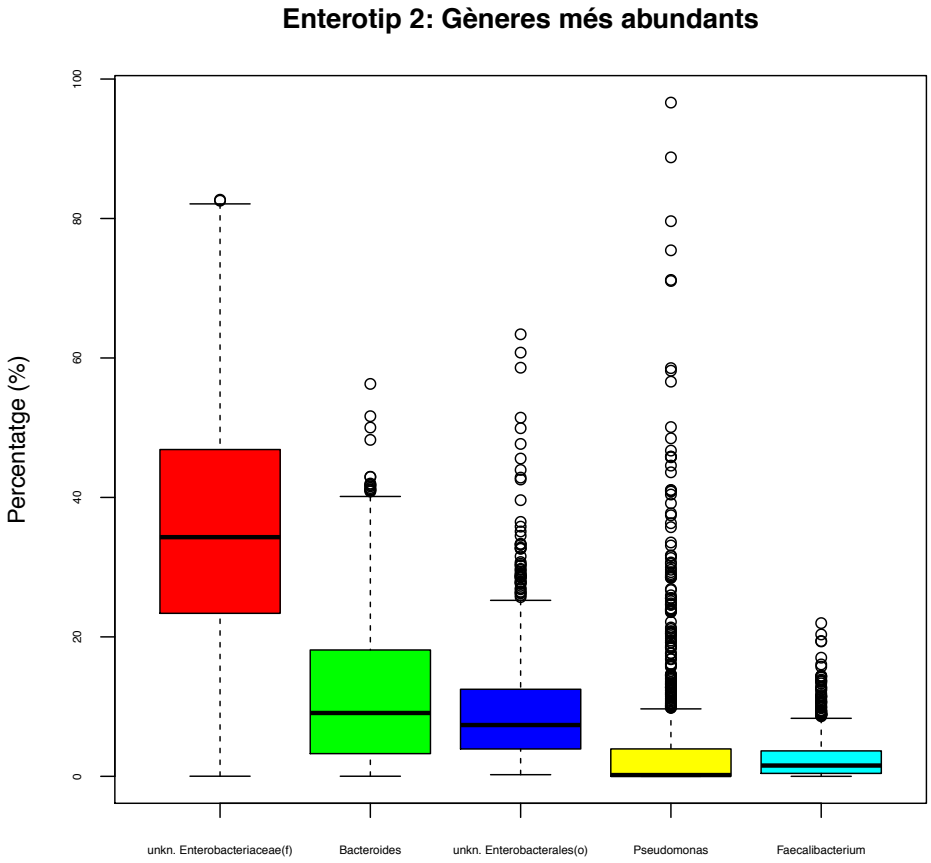


Figura 10. Gèneres més abundants a l'enterotip 2 del microbioma intestinal.

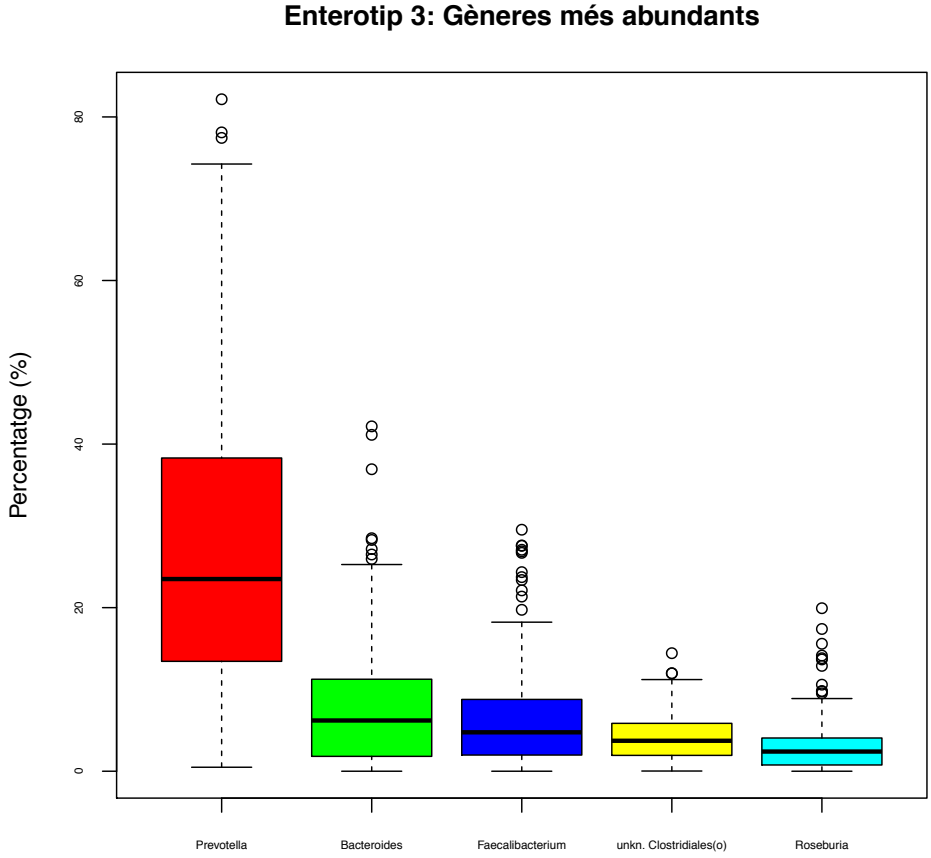


Figura 11. Gèneres més abundants a l'enterotip 3 del microbioma intestinal.

A la publicació d'Arumugam *et al.* (2011) es van obtindre tres enterotips caracteritzats per l'abundància de tres gèneres bacterians concrets: *Bacteroides*, *Prevotella* i *Ruminococcus*. Els resultats obtinguts amb les mostres utilitzades en aquest projecte, quan s'han analitzat 3 enterotips, s'han obtingut els gèneres *Bacteroides* i *Prevotella*, i la família dels *Enterobacteriaceae* com a característics dels enterotips obtinguts.

En publicacions posteriors a la d'Arumugam *et al.* (2011) s'han obtingut resultats diferents a aquest primer paper en què es proposa l'existència dels enterotips. A una publicació realitzada per l'equip compost per Wu i col·laboradors (2011) s'associen dos enterotips amb dues dietes a llarg termini: l'enterotip *Bacteroides* s'associa a una dieta enriquida en proteïnes i greixos animals, mentre que l'enterotip *Prevotella* s'associa a una dieta composta principalment per carbohidrats. En aquesta publicació es posa en dubte l'existència de l'enterotip *Ruminococcus*, el qual es podria explicar fent pertànyer els subjectes a l'enterotip *Bacteroides* amb un lleuger augment del comptatge del gènere *Ruminococcus* (Wu *et al*., 2011).

A l'article de Liang *et al.* (2016) es van trobar també 3 enterotips que es corresponen als resultats d'aquest treball, és a dir, es caracteritzen per la presència dels gèneres *Bacteroides*, *Prevotella* i la família *Enterobacteriaceae*. En aquesta publicació es proposa l'existència d'aquest últim enterotip esmentat (Liang *et al*., 2016).

En una revisió publicada per Cheng i Ning en 2019 es posa en dubte l'existència d'uns pocs enterotips. En aquesta publicació se li dona èmfasi a la complexitat extrema del microbioma intestinal humà i s'ha vist que amb l'augment del número de mostres s'ha observat, en tot cas, un gradient continu de diferències lleugeres en la composició microbiològica de l'intestí humà, ja que els enterotips analitzats en les últimes anàlisis realitzades per diferents grups d'investigació es veuen molts solapaments de mostres (Cheng & Ning, 2019).

Als resultats obtinguts en aquest projecte, encara que es poden separar les mostres en tres enterotips, no es deixa d'observar els solapaments mencionats per Cheng i Ning (2019).

3.2 Descripció de les metadades

Els resultats de la descripció de les metadades s'han representat utilitzant histogrames, diagrames de caixes i de barres. A la *Figura 12*, es mostren les distribucions de les primeres variables seleccionades, segons el llistat de la secció 2.5 del treball. Les figures corresponents a totes les metadades utilitzades per entrenar el model de classificació es poden trobar a l'annex "6.2 Figures de la descripció de les metadades". Aquestos resultats mostren que les variables clíniques i presenten diferents distribucions i algunes no estan equilibrades, com per exemple, la de la variable resposta "Malaltia autoimmunitària". Això s'ha de tindre en

compte a l'hora de generar qualsevol model de classificació per no obtindre uns resultats esbiaixats a favor de la categoria majoritària.

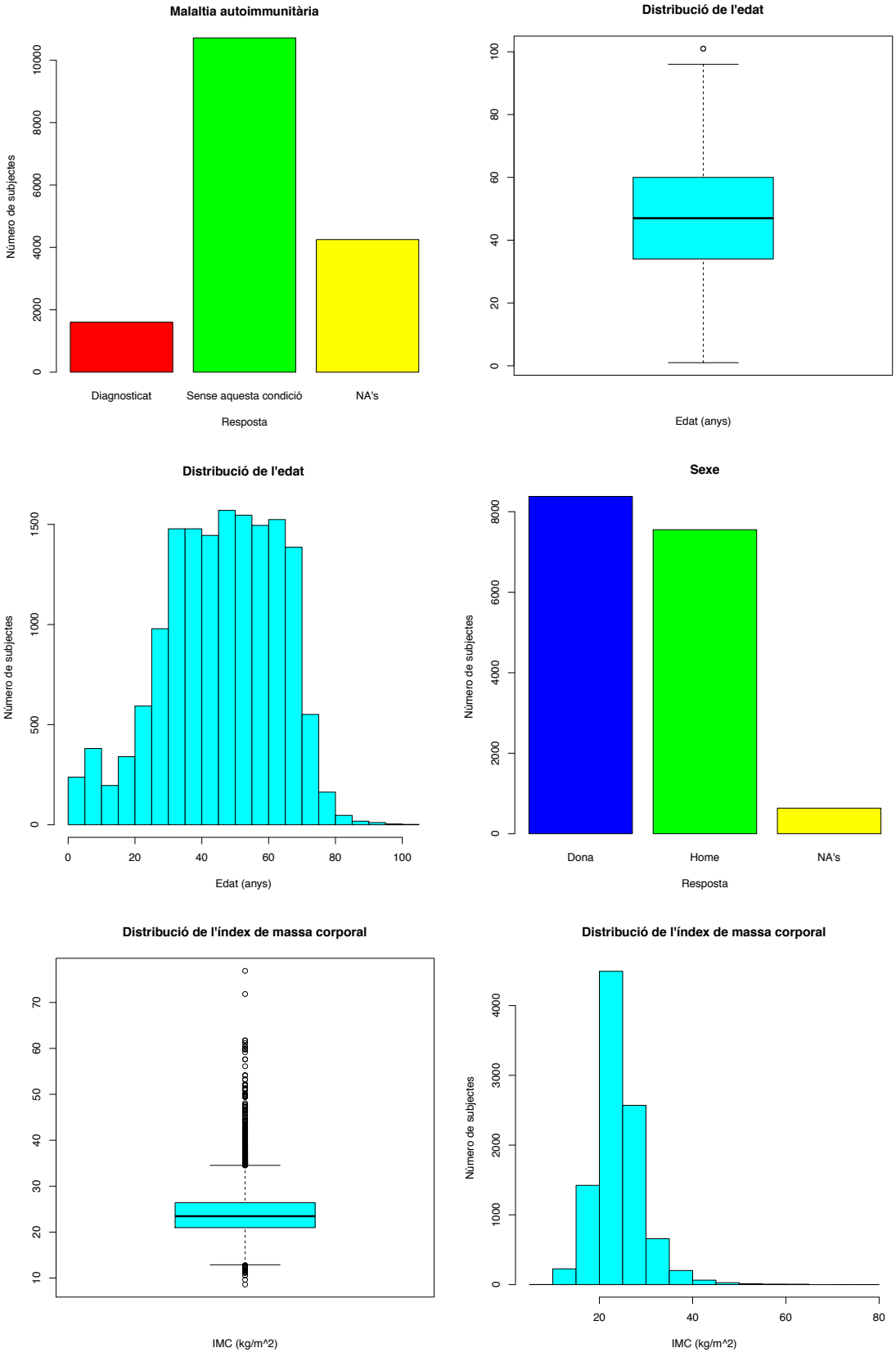


Figura 12. Distribució dels valors de les primeres metadades seleccionades de la base de dades.

3.3 Característiques dels *data sets* d'entrenament i test

Els resultats d'aplicar els diferents mètodes de mostratge per generar els *data sets* d'entrenament i de test són els mostrats a la *Taula 1*.

Taula 1. Número de mostres d'entrenament i test segons el mètode de mostratge.

Mètode de mostratge	Mostres Entrenament	Mostres Test
Cap mètode	1213	596
SMOTE	1421	693
Up-Sampling	2020	994
Down-Sampling	406	198

S'observa que amb el mètode de SMOTE augmenta el nombre de mostres lleugerament. D'aquesta manera, encara que es generen noves mostres, no s'està duplicant molt la informació i s'està equilibrant el nombre de les diferents categories de la variable resposta. Amb l'*up-sampling* augmenta fins quasi el doble la quantitat de mostres. Potser, amb aquest mètode s'estiga duplicant massa informació. Amb el *down-sampling* es redueix a pràcticament un terç el número de mostres per a l'anàlisi. Un gran inconvenient d'aquest mètode és que s'està perdent molta informació amb la reducció del nombre de mostres.

3.4 Resultats dels models de *Random Forest*

A la *Taula 2* es mostren els resultats d'aplicar els diferents mètodes de mostratge i paràmetres d'entrenament per generar models de classificació *Random Forest*.

Taula 2. Mètodes de mostratge, paràmetres d'entrada i estadístics de l'avaluació del rendiment dels diferents models de classificació Random Forest generats.

Mètode de mostratge	Paràmetre	Exactitud	Kappa	Sensibilitat	Especificitat
Cap mètode	**ntree = 500**	**0.8960**	**0.4988**	**0.3737**	**1.0000**
Cap mètode	ntree = 1000	0.8960	0.4988	0.3737	1.0000
Cap mètode	ntree = 10000	0.8943	0.4880	0.3636	1.0000
SMOTE	ntree = 500	0.7590	0.4751	0.4680	0.9773
SMOTE	ntree = 1000	0.7662	0.4906	0.4747	0.9848
SMOTE	ntree = 10000	0.7633	0.4843	0.4714	0.9823
Up-Sampling	ntree = 500	0.7002	0.4004	0.4004	1.0000
Up-Sampling	ntree = 1000	0.7002	0.4004	0.4004	1.0000
Up-Sampling	ntree = 10000	0.7002	0.4004	0.4004	1.0000
Down-Sampling	ntree = 500	0.7172	0.4343	0.5758	0.8586
Down-Sampling	ntree = 1000	0.7172	0.4343	0.5657	0.8687
Down-Sampling	ntree = 10000	0.6970	0.3939	0.5556	0.8384

El millor model generat en aquest cas seria el primer, sense utilitzar cap mètode de mostratge, ja que és el que té el valor Kappa més alt i, a més, el paràmetre "ntree" amb un valor més baix (500 arbres d'entrenament) i, per tant, necessitarà un temps de computació menor que el segon per l'entrenament (que té el mateix valor Kappa). En aquesta taula també s'observa que canviant els mètodes de mostratge no es millora el primer model, pel que fa als valors Kappa. Del mètode de mostratge SMOTE es pot destacar que manté el valor Kappa i equilibra un poc els valors de sensibilitat i especificitat (encara que no el suficient). Amb l'*up-sampling* s'ha baixat molt el valor de Kappa i ha pujat molt poc la sensibilitat respecte a no utilitzar cap mètode de mostratge. Amb el *down-sampling* s'han equilibrat encara més els valors de sensibilitat i d'especificitat però el valor de Kappa ha baixat molt, respecte als dos primers mètodes de mostratge. Aquest model té una exactitud de quasi un 90%. No obstant això, el valor kappa és relativament baix, encara que segons McHugh (2012), es podria considerar que té una adequació moderada. A continuació, s'observa un requadre que conté la matriu de confusió i molts dels estadístics calculats per poder avaluar el millor model utilitzant l'algoritme de *Random Forest*.

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction               Diagnosticat Sense aquesta condició
##   Diagnosticat                     37                       0
##   Sense aquesta condició           62                     497
##
##                Accuracy : 0.896
##                  95% CI : (0.8686, 0.9193)
##     No Information Rate : 0.8339
##     P-Value [Acc > NIR] : 1.100e-05
##
##                   Kappa : 0.4988
##
##  Mcnemar's Test P-Value : 9.408e-15
##
##             Sensitivity : 0.37374
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 0.88909
##              Prevalence : 0.16611
##          Detection Rate : 0.06208
##    Detection Prevalence : 0.06208
##       Balanced Accuracy : 0.68687
##
##        'Positive' Class : Diagnosticat
```

Si s'observa la matriu de confusió, es pot deduir que aquest model de classificació dona com a resposta, quasi sempre, un "Sense aquesta condició", ja que quasi totes les mostres tenien aquesta classe en la variable de resposta. En conclusió, podem dir que aquest classificador no seria el millor model per predir la probabilitat de que una persona té desenvolupada una malaltia autoimmunitària.

3.5 Resultats dels models de *Support Vector Machine*

A la *Taula 3* es mostren els resultats d'aplicar els diferents mètodes de mostratge i *kernels* utilitzats per generar models de classificació SVM.

Taula 3. Mètodes de mostratge, kernels utilitzats i estadístics de l'avaluació del rendiment dels diferents models de classificació SVM generats.

Mètode de mostratge	Paràmetres	Exactitud	Kappa	Sensibilitat	Especificitat
Cap mètode	kernel = vanilladot; lineal	0.8138	0.3195	0.4242	0.8913
Cap mètode	**kernel = rbfdot; gaussià**	**0.8926**	**0.4770**	**0.3535**	**1.0000**
SMOTE	kernel = vanilladot; lineal	0.6320	0.2256	0.4512	0.7677
SMOTE	kernel = rbfdot; gaussià	0.6999	0.3546	0.4478	0.8889

El millor model generat en aquest cas seria el segon, sense haver utilitzat cap mètode de mostratge. Aquest utilitza un *kernel* gaussià, i té un valor Kappa molt més alt que els altres models generats mitjançant l'algoritme SVM. No obstant això, aquestos models no milloren el model que s'havia obtingut amb l'algoritme *Random Forest*. Les matrius de confusió i molts dels estadístics calculats per poder avaluar els 4 models generats utilitzant l'algoritme de SVM es troben en l'annex "6.3 Matrius de confusió i estadístics d'avaluació dels classificadors SVMs generats".

3.6 Resultats dels models de XGBoost

A la *Taula 4* es mostren els 10 millors resultats d'aplicar els diferents mètodes de mostratge i paràmetres d'entrenament per generar models de classificació XGBoost. Per veure la taula completa amb tots els models generats mitjançant aquest algoritme de *Machine Learning*, es pot consultar l'annex "6.4 Resultats de l'avaluació del rendiment dels models de classificació XGBoost utilitzant diferents mètodes de mostratge i paràmetres d'entrenament".

Taula 4. Mètodes de mostratge, paràmetres d'entrada i estadístics de l'avaluació del rendiment dels 10 millors models de classificació XGBoost generats.

Mètode de mostratge	Paràmetres	Exactitud	Kappa	Sensibilitat	Especificitat
SMOTE	max.depth = 1, nrounds = 150	0.8355	0.6572	0.7273	0.9167
SMOTE	max.depth = 1, nrounds = 300	0.8297	0.6464	0.7340	0.9015
SMOTE	**max.depth = 4, nrounds = 50**	**0.8384**	**0.6612**	**0.7071**	**0.9369**
SMOTE	max.depth = 4, nrounds = 150	0.8384	0.6603	0.6970	0.9444
SMOTE	max.depth = 4, nrounds = 300	0.8355	0.6545	0.6970	0.9394
SMOTE	max.depth = 4, nrounds = 400	0.8355	0.6545	0.6970	0.9394
SMOTE	max.depth = 4, nrounds = 500	0.8341	0.6511	0.6902	0.9419
SMOTE	max.depth = 4, nrounds = 600	0.8341	0.6514	0.6936	0.9394
SMOTE	max.depth = 4, nrounds = 1000	0.8341	0.6511	0.6902	0.9419
SMOTE	max.depth = 5, nrounds = 300	0.8312	0.6447	0.6835	0.9419

Amb tots aquestos models entrenats i avaluats, podem dir que el millor és aquell que ha utilitzat l'algoritme de mostratge de SMOTE per generar les dades d'entrenament i test, amb uns paràmetres opcionals de "max.depth = 4" i "nrounds = 50", obtenint els següents resultats a l'avaluació del model:

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction               Diagnosticat Sense aquesta condició
##   Diagnosticat                    210                      25
##   Sense aquesta condició           87                     371
##
##                Accuracy : 0.8384
##                  95% CI : (0.8088, 0.865)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6612
##
##  Mcnemar's Test P-Value : 8.216e-09
##
##             Sensitivity : 0.7071
##             Specificity : 0.9369
##          Pos Pred Value : 0.8936
##          Neg Pred Value : 0.8100
##              Prevalence : 0.4286
##          Detection Rate : 0.3030
##    Detection Prevalence : 0.3391
##       Balanced Accuracy : 0.8220
##
##        'Positive' Class : Diagnosticat
```

Aquestos resultats mostren un valor Kappa acceptable: 0.6612. Segons McHugh (2012), aquest valor de Kappa tindria una adequació a les dades reals substancial. A més, la sensibilitat i l'especificitat del model estan molt més equilibrades i són relativament altes. També es disposa d'una exactitud ajustada d'un 82%. Es podria considerar aquest model de classificació com acceptable, havent millorat tots els anteriors.

3.7 Anàlisi de l'impacte de les diferents variables predictores del model de classificació

En la *Figura 13* s'observen les 25 variables que més importància relativa han tingut a l'hora d'executar els càlculs en el model de classificació. Aquestes variables són:

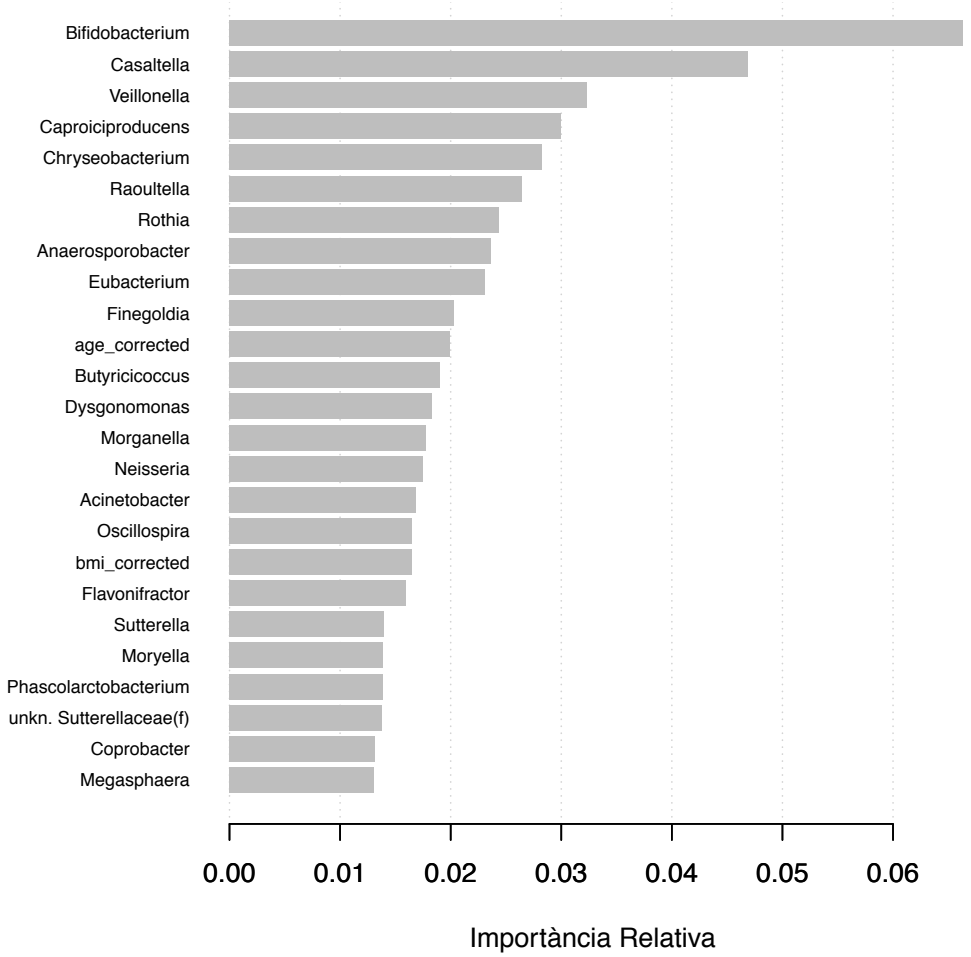


Figura 13. Les 25 variables amb una importància relativa més alta a l'hora d'executar els càlculs en el millor model de classificació XGBoost.

En aquestos resultats es pot observar que el gènere *Bifidobacterium* és, sens dubte, el tipus d'organisme que més ha contribuït en el model de classificació a l'hora de predir si una mostra pertany a un pacient amb o sense una malaltia autoimmunitària. A aquest el segueixen els gèneres *Casatella* i *Veillonella*.

Les els limfòcits T Reguladors (cèl·lules Treg) són moduladors del sistema immune, mantenen la tolerància als antígens propis i prevenen el desenvolupament de malalties autoimmunitàries (Bettelli *et al.*, 2006). Els limfòcits T-*helper* tipus 17 (Th17) són essencials en la defensa davant dels bacteris i fongs, però també són molt importants en el desenvolupament de malalties autoimmunitàries, ja que produeixen citoquines pro-inflamatòries, com la interleuquina-17 (IL-17) i la IL-22. Els limfòcits Th17 s'acumulen a l'intestí, suggerint que la modulació del microbioma intestinal

puga regular el desenvolupament d'aquestes cèl·lules. A més, alguns estudis han demostrat que la microbiota intestinal poden afectar al desenvolupament de l'esclerosi múltiple, una malaltia autoimmunitària que afecta al sistema nerviós central, caracteritzada per reaccions contra les beines de mielina. Els bacteris del gènere *Bifidobacterium* promouen la fermentació dels carbohidrats, generen acetat i lactat, alliberen polifenols i àcid linoleic, i tenen activitats antioxidants. Molts estudis han demostrat que un tractament d'administració de bacteris del gènere *Bifidobacterium* –entre d'altres– han millorat els processos inflamatoris del sistema nerviós central, induint cèl·lules Treg i disminuint la quantitat de cèl·lules Th17 en la mucosa intestinal. Això s'adequa al fet de que els subjectes sans presenten una major quantitat relativa de microorganismes del gènere *Bifidobacterium* que els individus amb alguna malaltia autoimmunitària (Vilela de Oliveira *et al.*, 2017).

En un estudi es va observar que la falta de les dues espècies més abundants del gènere *Bifidobacterium*, és a dir, *Bifidobacterium adolescentis* i *Bifidobacterium pseudocatenulatum*, i una abundància del gènere *Bacteroides*, s'ha associat amb una autoimmunitat de les cèl·lules β en xiquets que, per tant, han desenvolupat un tipus de diabetis autoimmunitària. Alguns estudis funcionals han proposat que una baixa abundància dels bifidobacteris i les espècies productores de butirat pot tindre una relació amb esdeveniments adversos que afecten la funció de barrera epitelial intestinal i la inflamació (De Goffau *et al.*, 2013).

La estimulació de les cèl·lules del sistema immunitari utilitzant lipopolisacàrids produïts per microorganismes repetidament genera un estat refractari temporal en què s'origina una resistència a aquest tipus de biomolècules. Com que els lipopolisacàrids també s'anomenen endotoxines, aquest procés rep el nom de "tolerància a les endotoxines" (Watson & Kim, 1963). En un altre estudi, d'una banda, s'ha observat que les poblacions de Finlàndia i Estònia, que tenen un microbioma amb una predominància del gènere *Bacteroides*, són més propers a no tindre una inducció de la tolerància a les endotoxines durant el primer any de vida d'una persona. D'altra banda, en aquest mateix període de la vida, la població de Rússia, amb una quantitat més alta del gènere *Bifidobacterium*, és més propera a tindre una inducció de la tolerància a les endotoxines. Aquesta intolerància a les endotoxines s'ha associat a un desenvolupament prematur de malalties autoimmunitàries (Vatanen *et al*., 2016).

Finalment, també cal destacar que, de les metadades afegides a l'entrenament del model de classificació, tant l'edat (variable "age_corrected") com l'índex de massa corporal (variable "bmi_corrected") estan entre les 20 variables amb més importància relativa. Les altres metadades seleccionades que s'havien utilitzat en un principi per entrenar el model no s'han inclòs al model o han tingut una importància relativa molt baixa.

4. Conclusions

En el primer objectiu d'aquest treball es pretenia descriure i analitzar la composició taxonòmica dels microorganismes presents al microbioma intestinal humà i de les metadades. Amb els resultats obtinguts podem concloure que, degut a l'important solapament dels clústers analitzats (tant utilitzant 3 com 5 clústers), no podem afirmar que hi haja un número concret d'enterotips que separe la composició del microbioma intestinal dels humans. Això pot ser degut al gran número de mostres i la complexitat de la composició taxonòmica dels microorganismes presents en aquesta part del cos.

El segon objectiu del TFM tenia com a meta la generació d'un model de classificació utilitzant algun algoritme de *Machine Learning* per predir la probabilitat d'haver desenvolupat alguna malaltia autoimmunitària a partir de les dades taxonòmiques i les metadades. Amb les aproximacions preses, és a dir, utilitzant diferents mètodes de mostratge i *Random Forest*, *Support Vector Machine* i XGBoost com a algoritmes de *Machine Learning* s'han conclòs els següents punts:

- Utilitzar el mètode de mostratge SMOTE, a més d'equilibrar les freqüències de les dues classes de la variable resposta "autoimmune", també ha mostrat uns bons resultats si es compara amb els mètodes d'*up-sampling* i *down-sampling*. En el cas de l'algoritme XGBoost, ha demostrat, a més, obtindre un millor rendiment que sense haver utilitzat cap mètode de mostratge.
- Els millors models obtinguts amb els algoritmes *Random Forest* i *Support Vector Machine* han obtingut un valor de Kappa molt semblant que, segons McHugh (2012), es correspon a una adequació moderada dels resultats obtinguts amb els reals. No obstant això, amb l'algoritme XGBoost s'ha obtingut un model de classificació amb un valor de Kappa que es correspon a una adequació substancial dels resultats, segons McHugh (2012). A més, aquest model ha obtingut una sensibilitat i una especificitat molt equilibrades i més altes, en comparació a les obtingudes als altres models, sent del 70% i 93%, respectivament. Encara que l'especificitat és molt alta, en aquest àmbit clínic també convindria tindre una sensibilitat amb uns valors molt alts, ja que és molt important diagnosticar una malaltia en una persona que la pateix. De la mateixa manera, no és tan important la especificitat, ja que diagnosticar que no estan malalts els pacients que no ho estan es podria saber en les properes proves de diagnòstic.

Els resultats del rendiment obtinguts en el millor model generat mitjançant l'algoritme de XGBoost es podria explicar per molts motius. La variable resposta és si s'ha desenvolupat, o no, una malaltia autoimmunitària. Amb això, s'ha de tindre en compte que hi ha molts tipus de malalties autoimmunitàries, molt complexes i diverses entre elles. Per tant, augmenta molt la dificultat de trobar un marcador o una característica comuna en totes elles.

Amb aquestes conclusions, es pot afirmar que els objectius d'aquest treball han estat assolits. No obstant això, es podrien seguir altres metodologies per intentar obtindre un model millor al generat durant aquest projecte. D'una banda, pel que fa a la composició taxonòmica, es podrien haver realitzat altres tipus de filtres, més restrictius o més laxos, als aplicats per reduir el soroll i eliminar les mostres de poca qualitat. D'altra banda, s'hagueren pogut seleccionat altres metadades per entrenar els models. Pel que fa a la generació del model, s'hagueren pogut utilitzar altres mètodes de mostratge o diferents paràmetres d'entrada per a l'entrenament dels models.

Els resultats obtinguts en aquest projecte ens obrin altres vies per investigar en un futur per intentar millorar el model o utilitzar-lo i aplicar-lo a l'àmbit clínic. Per millorar el model es podrien aplicar les modificacions que s'han anomenat a l'anterior paràgraf. A més, també es podrien fer servir altres algoritmes de *Machine Learning*. Amb el model generat també es podria generar un script que agafara unes dades d'entrada i et donara la probabilitat d'haver desenvolupat una malaltia autoimmunitària. D'aquesta manera, en un àmbit clínic, es podria aplicar, per seguir fent un tipus de proves o unes altres, als pacients que encara no se'ls ha pogut diagnosticar cap malaltia. El resultat obtingut a l'últim punt del projecte ens dona una via de recerca sobre el gènere de *Bifidobacterium* per intentar trobar relacions més concretes amb el desenvolupament de malalties autoimmunitàries. Amb això, es podrien generar nous tractaments per a algunes de les malalties d'aquest tipus.

5. Glossari

- API: *Application Programming Interface*.
- ARN: Àcid ribonucleic.
- CSV: *Comma-Separated Values*.
- EMBL-EBI: *European Molecular Biology Laboratory-European Bioinformatics Institue*.
- Enterotip: estrat que representa el tipus de microbioma intestinal humà diferenciat per la seua composició taxonòmica.
- IL: Interleuquina.
- Índex CH: índex Caliński-Harabasz.
- JSD: Jensen-Shannon *divergence*.
- JSON: *JavaScript Object Notation*.
- *k*: número de clústers.
- Kappa: és un estadístic que mesura l'ajustament d'uns valors qualitatius predits amb uns observats considerant l'efecte de l'atzar.
- *Machine Learning*: Tipus d'algoritmes utilitzats per predir dependències entre unes dades introduïdes utilitzant una mostra d'entrenament.
- Malalties autoimmunitàries: malaltia del sistema immunitari en què el es genera una resposta immunitària contra les cèl·lules del propi cos.
- Matriu esparsa: matriu numèrica en què la majoria dels seus valors són 0 o contenen un valor molt proper a 0.
- Metadades: conjunt de dades clíniques i no clíniques que caracteritzen un pacient concret.
- Metagenòmica: ciència que estudia els gens dels microorganismes obtinguts a partir de mostres sense cultivar.
- Microbioma: Col·lecció de tots els genomes dels microorganismes presents en un ambient concret.
- OTU: *Operational Taxonomic Units*.
- PAM: *Partitioning Around Medoids*.
- PCoA: *Principal Coordinates Analysis*.
- *Random Forest*: Algoritme de *Machine Learning* de predicció que utilitza arbres de decisió.
- SMOTE: *Synthetic Minority Over-sampling Technique*.
- *Support Vector Machine*: Algoritme de *Machine Learning* de predicció que utilitza *support vectors*.
- SVM: *Support Vector Machine*.
- Taxonomia: Classificació dels éssers vius en grups.
- TFM: Treball de Fi de Màster
- Th17: limfòcits T-*helper*.
- Treg: limfòcits T Reguladors.
- TSV: *Tab-Separated Values*.
- XGBoost: *Extreme Gradient Boosting*.

6. Bibliografia

1) Angiari, S., & Constantin, G. (2014). Regulation of T cell trafficking by the T cell immunoglobulin and mucin domain 1 glycoprotein. *Trends in molecular medicine, 20*(12): 675-684.

2) Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., De Vos, W. M., Brunak, S., Dore, J., MetaHIT Consortium, Weissenbach, J., Ehrlich, S. D., & Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature, 473*(7346): 174.

3) Bates, D., & Maechler, M. (2019). Matrix: Sparse and Dense Matrix Classes and Methods. R Package version 1.2-17. Accés: 22/12/2019, https://CRAN.R-project.org/package=Matrix.

4) Bettelli, E., Carrier, Y., Gao, W., Korn, T., Strom, T. B., Oukka, M., Weiner, H. L., & Kuchroo, V. K. (2006). Reciprocal developmental pathways for the generation of pathogenic effector T H 17 and regulatory T cells. *Nature, 441*(7090): 235-238.

5) Bork, P. (2011). Enterotyping: the original publication. Accés: 13/10/2019, https://enterotype.embl.de/enterotypes.html.

6) Bougeard, S., & Dray, S. (2018). Supervised multiblock analysis in R with the ade4 package. *Journal of Statistical Software, 86*(1): 1-17.

7) Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1): 1-27.

8) Callahan, B., Proctor, D., Relman, D., Fukuyama, J., & Holmes, S. (2016). Reproducible research workflow in R for the analysis of personalized human microbiome data. *In Biocomputing 2016: Proceedings of the Pacific Symposium*: 183-194.

9) Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2019). xgboost: Extreme Gradient Boosting. R package version 0.90.0.2. Accés: 22/12/2019, https://CRAN.R-project.org/package=xgboost.

10) Cheng, M., & Ning, K. (2019). Stereotypes about enterotype: the old and new ideas. *Genomics Proteomics Bioinformatics 17*: 4-12.

11) Chessel, D., Dufour, A. B., & Thioulouse, J. (2004). The ade4 package-I-One-table methods. *R news, 4*(1): 5-10.

12) Couture-Bile, A. (2018). rjson: JSON for R. R packge version 0.2.20.

13) D'hooghe, M. B., Haentjens, P., Nagels, G., & De Keyser, J. (2012). Alcohol, coffee, fish, smoking and disease progression in multiple sclerosis. *European Journal of Neurology, 19*(4): 616-624.

14) De Goffau, M. C., Luopajärvi, K., Knip, M., Ilonen, J., Ruohtula, T., Härkönen, T., Orivuori, L., Hakala, S., Welling, G. W., Harmsen, H. J., & Vaarala, O. (2013). Fecal microbiota composition differs between children with β-cell autoimmunity and those without. *Diabetes, 62*(4): 1238-1244.

15) Vilela de Oliveira, G. L., Zazeri Leite, A., Stevanato Higuchi, B., Ignácio Gonzaga, M., & Sammartino Mariano, V. (2017). Intestinal dysbiosis and probiotic applications in autoimmune diseases. *Immunology, 152*(1): 1-12.

16) Dray, S., & Dufour, A. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software, 22*(4): 1-20.

17) Dray, S., Dufour, A. B., & Chessel, D. (2007). The ade4 package-II: Two-table and K-table methods. *R news, 7*(2): 47-52.

18) Eichenfield, A. H. (1999). Minocycline and autoimmunity. *Current opinion in pediatrics, 11*(5): 447-456.

19) Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory, 4*(7): 1858-1860.

20) GitLab. Accés: 10/12/2019, https://gitlab.com/.

21) Gleicher, N., Weghofer, A., Lee, I. H., & Barad, D. H. (2011). Association of FMR1 genotypes with in vitro fertilization (IVF) outcomes based on ethnicity/race. *PloS ONE, 6*(4): e18781.

22) Goronzy, J. J. & Weyand, C. M. (2012). Immune aging and autoimmunity. *Cellular and Molecular Life Sciences, 69*(10): 1615-1623.

23) Harpsøe, M. C., Basit, S., Andersson, M., Nielsen, N. M., Frisch, M., Wohlfahrt, J., Nohr, E. A., Linneberg, A., & Jess, T. (2014). Body mass index and risk of autoimmune diseases: a study within the Danish National Birth Cohort. *International Journal of Epidemiology, 43*(3): 843-855.

24) Ishihara, K. (2019). GitHub Gist: data.frame.2.sparseMatrix.r. convert data.frame with both factor and numèric columns into sparse matrix.

Accés 27/11/2019, https://gist.github.com/kotaishr/75a7912ab3e0d0af3847ee0fdec29205.

25) Ji, J., Sundquist, J., & Sundquist, K. (2016). Tonsillectomy associated with an increased risk of autoimmune diseases: A national cohort study. *Journal of Autoimmunity, 72:* 1-7.

26) Kaplan, J. (2019). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. R Package version 1.6.0. Accés: 22/12/2019, https://CRAN.R-project.org/package=fastDummies.

27) Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of statistical software, 11*(9): 1-20.

28) Kho, Z. Y., & Lal, S. K. (2018). The Human Gut Microbiome – A Potential Controller of Wellness and Disease. *Frontiers in Microbiology, 9*: 1835.

29) Kuhn, M. (2019). The caret Package. Accés: 18/11/2019, https://topepo.github.io/caret/subsampling-for-class-imbalances.html#subsampling-techniques.

30) Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Englhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T. (2019). caret: Classification and Regression Training. *Astrophysics Source Code Library*: Houghton, MI, USA.

31) Lantz, B. (2015). *Machine Learning with R: Discover how to build Machine learning algorithms, prepare data, and dig deep into data prediction techniques with R, 2nd Edition.* Packt Publishing Ltd. Birmingham, United Kingdom.

32) Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News, 6*(4): 8-12.

33) Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell, 124*(4): 837-848.

34) Li, Z., Ju, Z., & Frieri, M. (2013). The T-cell immunoglobulin and mucin domain (Tim) gene family in asthma, allergy, and autoimmunity. *Allergy and Asthma Proceedings 34*(1): e21-e26.

35) Liang, C., Tseng, H. C., Chen, H. M., Wang, W. C., Chiu, C. M., Chang, J. Y., Lu, K. Y., Weng, S. L., Chang, T. H., Chang, C. H., Weng, C. T., Wang, H. M., & Huang, H. D. (2017). Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics, 18*(Suppl 1): 932.

36) Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3): 18-22.

37) Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine, 8*(1): 1-11.

38) Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D. & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research 47*(1): 636-641.

39) Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2012). Cluster: cluster analysis basics and extensions. *R package version, 1*(2): 56.

40) McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P. C., Dunn, R. R., Fahimipour, A., K., Gaffney, J., Gilbert, J. A., Gogul, G., Green, J. L., Hugenholtz, P., Humphrey, G., Huttenhower, C., Jackson, M. A., Janssen, S., Jeste, D. V., Jiang, L., Kelley, S. T., Knights, D., Kosciolek, T., Ladau, J., Leach, J., Marotz, C., Meleshko, D., Melnik, A. V., Metcalf, J. L., Mohimani, H., Montassier, E., Navas-Molina, J., Nguyen, T. T., Peddada, S., Pevzner, P., Pollard, K. S., Rahnavard, G., Robbins-Pianka, A., Sangwan, N., Shorenstein, J., Smarr, L., Jin Song, S., Spector, T., Swafford, A. D., Thackray, V. G., Thompson, L. R., Tripathi, A., Vázquez-Baeza, Y., Vrbanac, A., Wischmeyer, P., Wolfe, E., Zhu, Q., The American Gut Consortium, & Knight, R. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems 3*: e00031-18.

41) McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3): 276-282.

42) Mezouar, S., Chantran, Y., Michel, J., Fabre, A., Dubus, J. C., Leone, M., Sereme, Y., Mège, J. L., Ranque, S., Desnues, B., Chanez, P., & Vitte, J. (2018). Microbiome and the immune system: from a healthy steady-state to allergy associated disruption. *Human Microbiome Journal, 10*: 11-20.

43) Nana Teukam, Y. G. (2019). Study of human gut microbiome in terms of taxonomy profiling and describing different indicators that correlates with diabetes (treball de fi de grau). Università di Roma, Roma, Itàlia.

44) R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

45) Ramos, M. (2015). RPubs brought to you by RStudio. Gantt Chart in R. Accés: 13/10/2019, https://rpubs.com/mramos/ganttchart.

46) Rahman, S. F., Olm, M. R., Morowitz, M. J., & Banfield, J. F. (2018). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *mSystems, 3*(1): e00123-17.

47) Riehle, K., Coarfa, C., Jackson, A., Ma, J., Tandon, A., Paithankar, S., Raghuraman, S., Mistretta, T. A., Saulnier, D., Raza, S., Diaz, M. A., Shulman, R., Aagaard , K., Verrsalovic, J., & Milosavljevic, A. (2012). The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences. *BMC Bioinformatics, 13*(Suppl. 13): S11.

48) RStudio Team (2016). RStudio: Integrated development for R. Rstudio, Inc., Boston, MA.

49) Rubtsova, K., Marrack, P., & Rubtsov, A. V. (2015). Sexual dimorphism in autoimmunity. *The Journal of Clinical Investigation, 125*(6): 2187-2193.

50) Saraswat, M. (2019). Machine Learning. Beginners Tutorial on XGBoost and Parameter Tuning in R. Accés: 27/11/2019, https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/.

51) Sharma, M., & Bayry, J. (2015). Autoimmunity: basophils in autoimmune and inflammatory diseases. *Nature Reviews Rheumatology, 11*(3): 129.

52) Torgo, L. (2011). *Data mining with R: learning with case studies.* Chapman and Hall/CRC. New York, NY.

53) Vatanen, T., Kostic, A. D., D'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., Hämäläinen, A. M., Peet, A., Tillmann, V., Uibo, R., Mokurov, S., Dorshakova, N., Ilonen, J., Virtanen, S. M., Szabo, S. J., Porter, J. A., Lähdesmäki, H., Huttenhower, C., Gevers, D., Cullen, T. W., Knip, M., DIABIMMUNE Study Group, & Xavier, R. J. (2016). Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell, 165*(4): 842-853.

54) Virtanen, S. M., Takkinen, H. M., Nwaru, B. I., Kaila, M., Ahonen, S., Nevalainen, J., Niinistö, S., Siljander, H., Simell, O., Ilonen, J., Hyöty, H., Veijola, & R., Knip, M. (2014). Microbial Exposure in Infancy and Subsequent Appearance of Type 1 Diabetes Mellitus–Associated Autoantibodies: A Cohort Study. *JAMA Pediatrics, 168*(8): 755-763.

55) Walesiak, M., & Dudek, A. (2019). clusterSim: Searching for optimal clustering procedure for a data set. R Package version, 0.48-1. Accés: 22/12/2019, https://CRAN.R-project.org/package=clusterSim.

56) Watson, D. W., & Kim, Y. B. (1963). Modification of host responses to bacterial endotoxins: I. Specificity of pyrogenic tolerance and the role of

hypersensitivity in pyrogenicity, lethality, and skin reactivity. *Journal of Experimental Medicine, 118*(3): 425-446.

57) Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1): 1-29.

58) Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, B., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., & Lewis, J. D. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science 334*(6052): 105-108.

59) XGboost developers (2019). XGBoost R Package: XGBoost R Tutorial. Accés 27/11/2019, https://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html.

60) Xie, Y. (2013). knitr: A general-purpose Tool for dynamic report generation in R. *R package version, 1*(1).

61) Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magda, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, A. C., Lauber, C., Clemente, J. C., Knights, D., Knight, R., & Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature, 486*(7402): 222-227.

62) Yazdani, M., Taylor, B. C., Debelius, J. W., Li, W., Knight, R., & Smarr, L. (2016). Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. *En Ia 2016 IEEE International Conference on Big Data (Big Data)*: 1272-1280.

7. Annexos

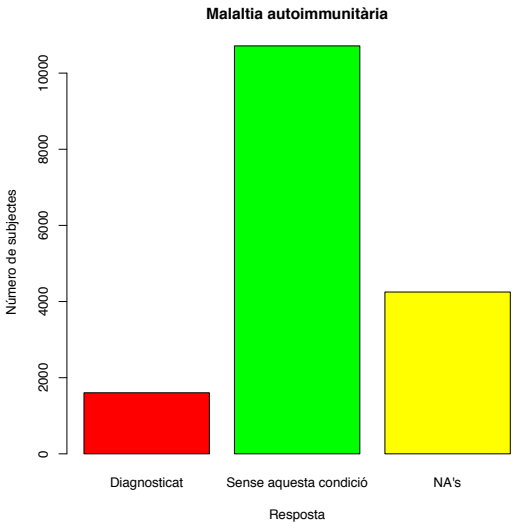
6.1 Llibreries d'**R**

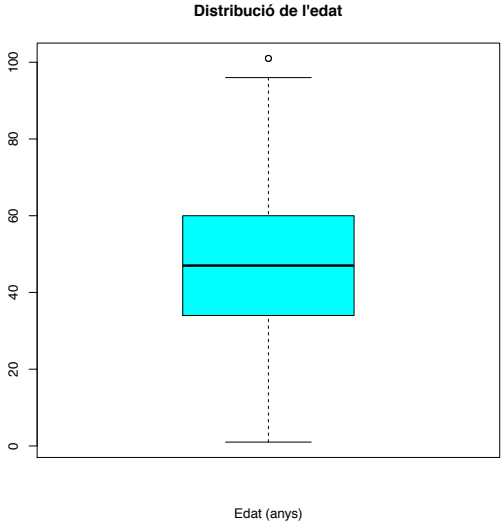
Les llibreries d'**R** utilitzades en aquest projecte són les següents:

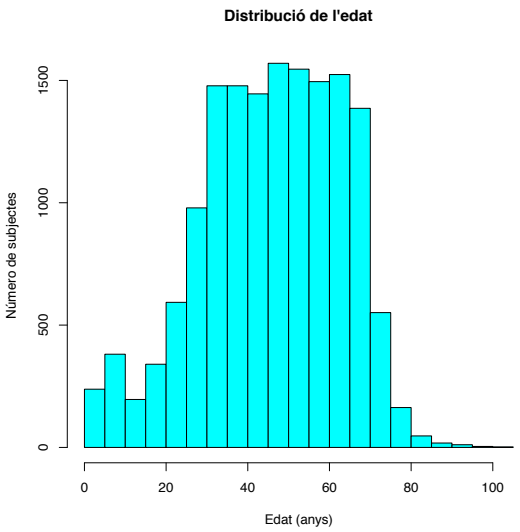
- "knitr" (Xie, 2013).
- "plotrix" (Lemon, 2006).
- "rjson" (Couture-Bile, 2018).
- "plyr" (Wickham, 2011).
- "cluster" (Maechler *et al.*, 2012).
- "clusterSim" (Walesiak & Dudek, 2019).
- "ade4" (Chessel *et al.*, 2004; Dray & Dufour, 2007; Dray *et al.*, 2007; Bougeard & Dray, 2018).
- "caret" (Kuhn *et al.*, 2019).
- "randomForest" (Liaw & Wiener, 2002).
- "DMwR" (Torgo, 2011).
- "fastDummies" (Kaplan, 2019).
- "kernlab" (Karatzoglou *et al.*, 2004).
- "xgboost" (Chen *et al.*, 2019).
- "Matrix" (Bates & Maechler, 2019).

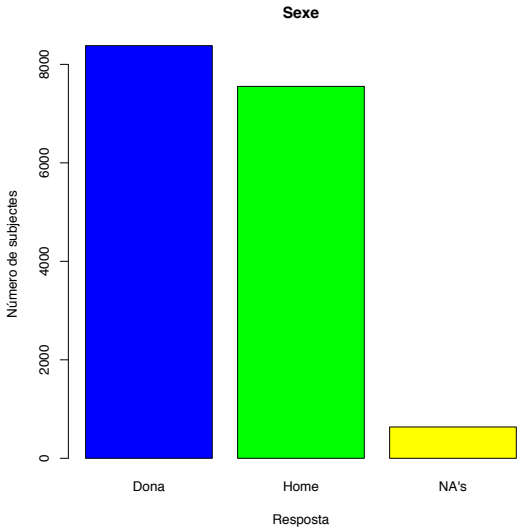
6.2 Figures de la descripció de les metadades

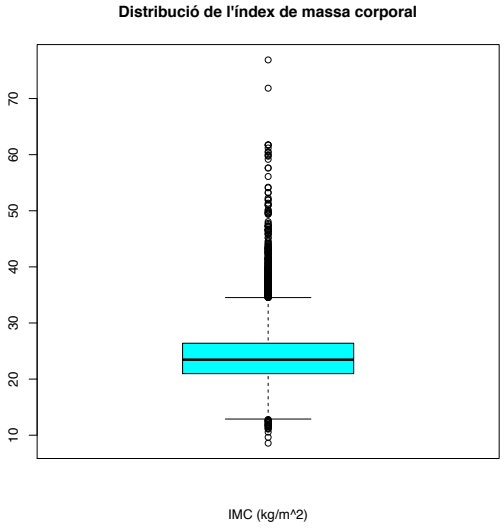
Els resultats obtinguts en la descripció de les metadades són els següents:

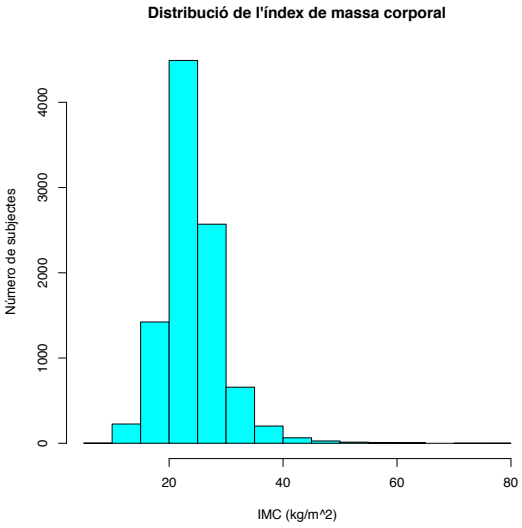


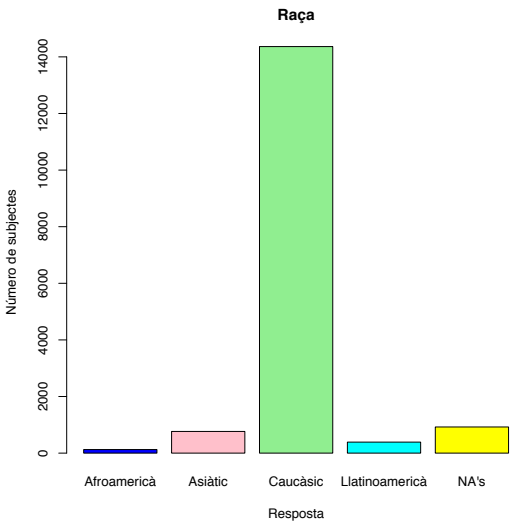


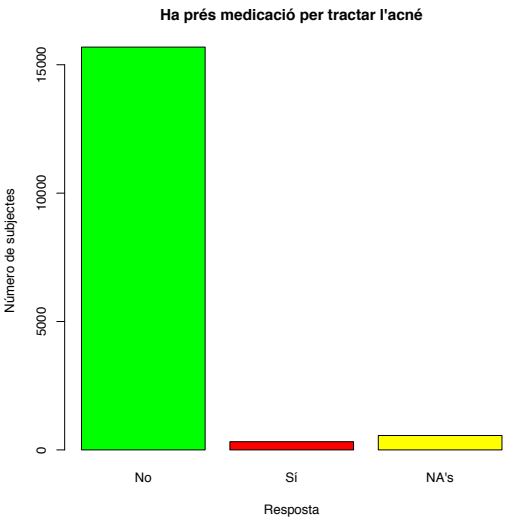

Distribució de l'edat
Número de subjectes
0
500
1000
1500
0
20
40
60
80
100
Edat (anys)

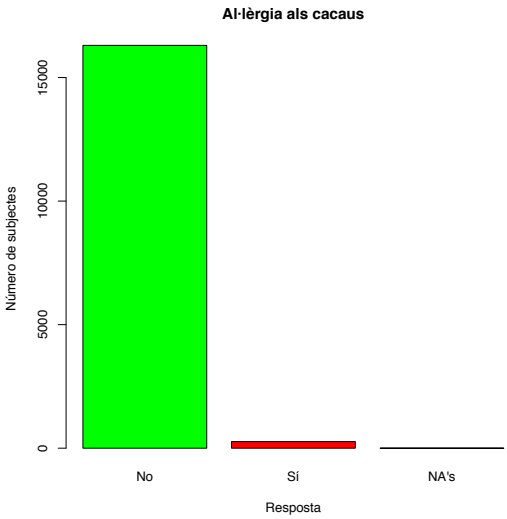

Sexe
Número de subjectes
0
2000
4000
6000
8000
Dona
Home
NA's
Resposta

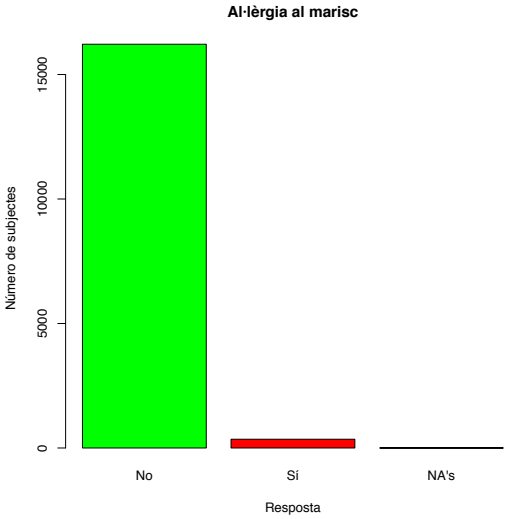

Distribució de l'índex de massa corporal
10
20
30
40
50
60
70
IMC (kg/m^2)

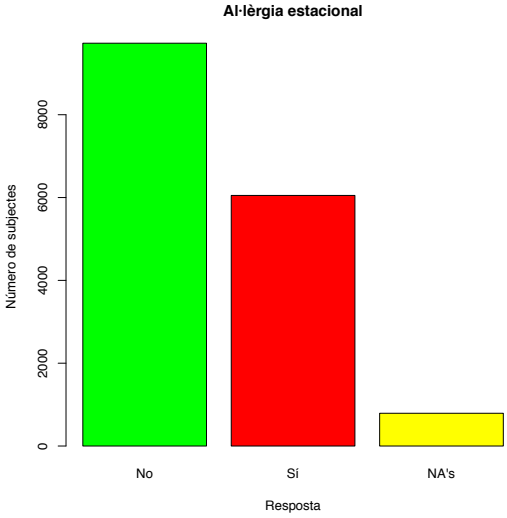

Distribució de l'índex de massa corporal
Número de subjectes
0
1000
2000
3000
4000
20
40
60
80
IMC (kg/m^2)

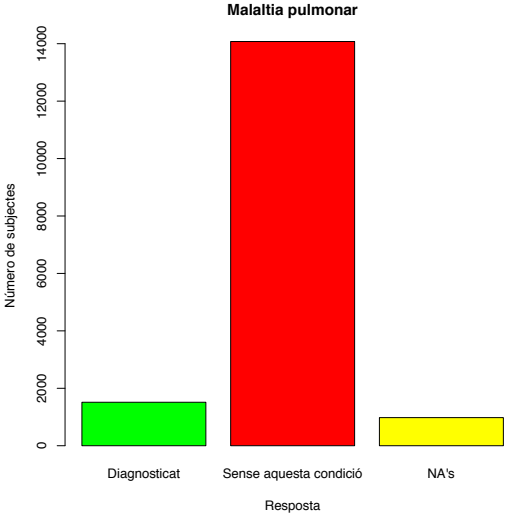

Raça
Número de subjectes
0
2000
4000
6000
8000
10000
12000
14000
Afroamericà
Asiàtic
Caucàsic
Llatinoamericà
NA's
Resposta

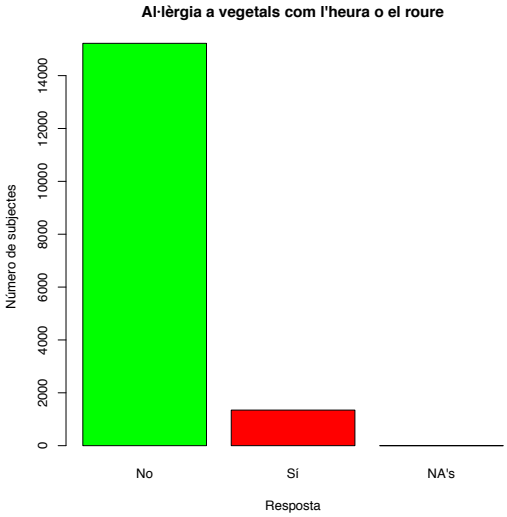

Ha prés medicació per tractar l'acné
Número de subjectes
0
5000
10000
15000
No
Sí
NA's
Resposta

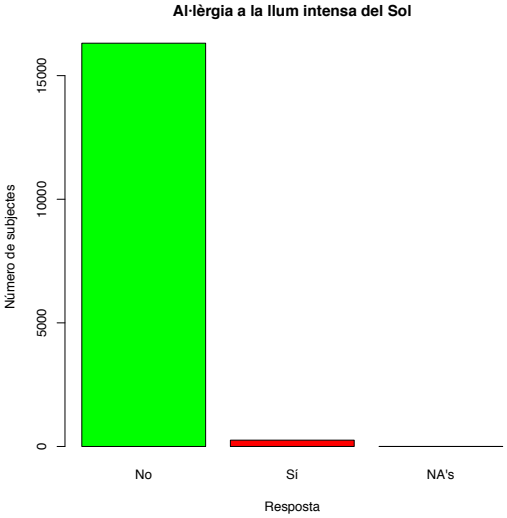

Al·lèrgia als cacaus
Número de subjectes
0
5000
10000
15000
No
Sí
NA's
Resposta

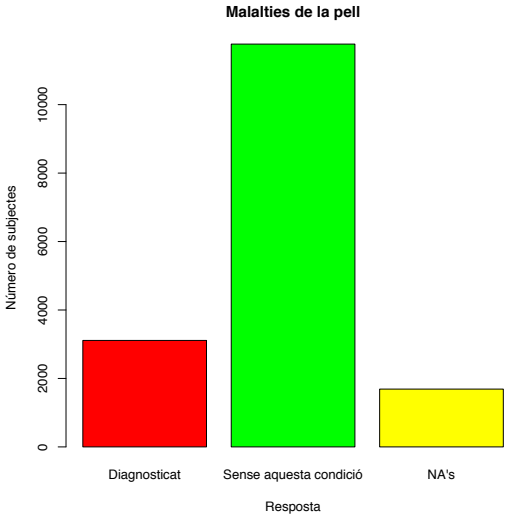

Al·lèrgia al marisc
Número de subjectes
0
5000
10000
15000
No
Sí
NA's
Resposta

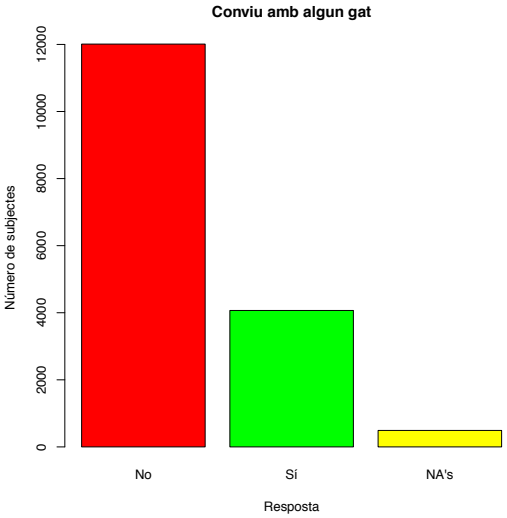

Al·lèrgia estacional
Número de subjectes
0
2000
4000
6000
8000
No
Sí
NA's
Resposta

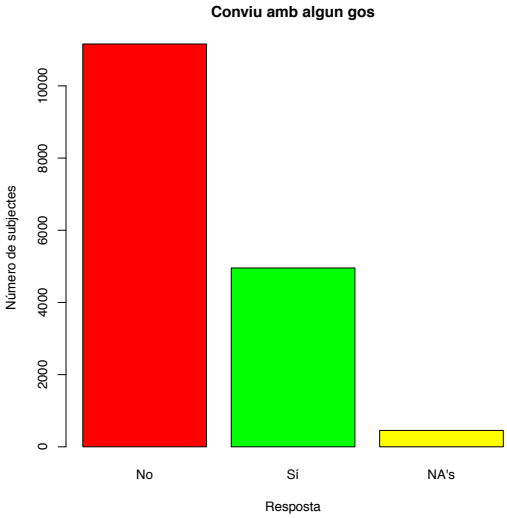

Malaltia pulmonar
Número de subjectes
0
2000
4000
6000
8000
10000
12000
14000
Diagnosticat
Sense aquesta condició
NA's
Resposta

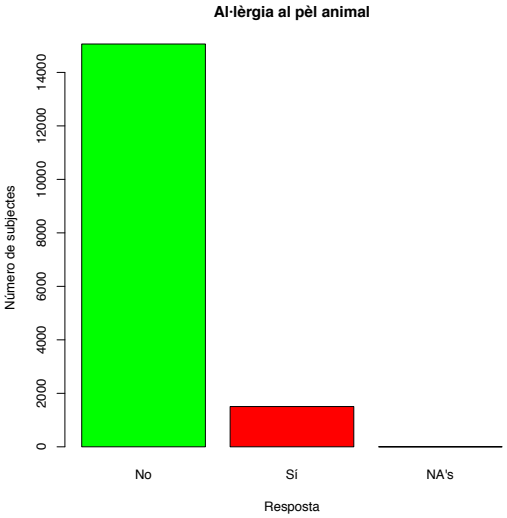

Al·lèrgia a vegetals com l'heura o el roure
Número de subjectes
0
2000
4000
6000
8000
10000
12000
14000
No
Sí
NA's
Resposta

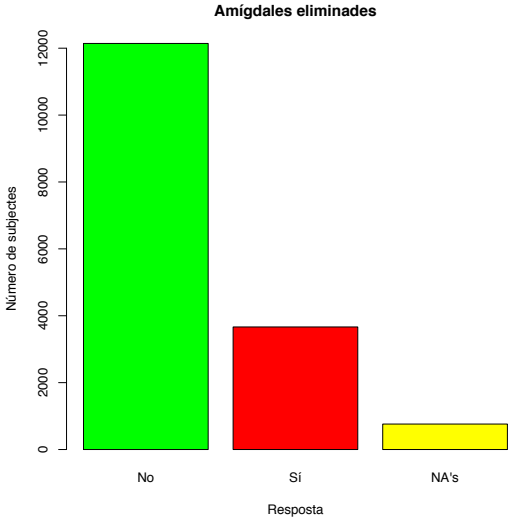

Al·lèrgia a la llum intensa del Sol
Número de subjectes
0
5000
10000
15000
No
Sí
NA's
Resposta

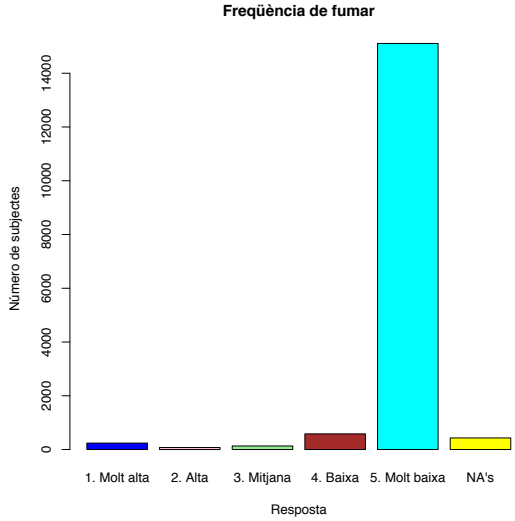

Malalties de la pell
Número de subjectes
0
2000
4000
6000
8000
10000
Diagnosticat
Sense aquesta condició
NA's
Resposta

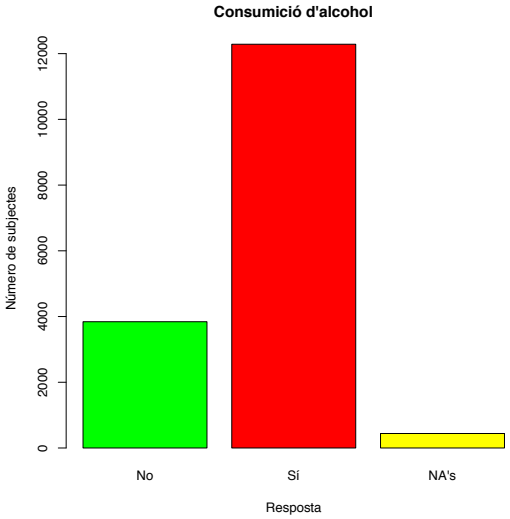

Conviu amb algun gat
Número de subjectes
0
2000
4000
6000
8000
10000
12000
No
Sí
NA's
Resposta


Conviu amb algun gos
Número de subjectes
0
2000
4000
6000
8000
10000
No
Sí
NA's
Resposta


Al·lèrgia al pèl animal
Número de subjectes
0
2000
4000
6000
8000
10000
12000
14000
No
Sí
NA's
Resposta


Amígdales eliminades
Número de subjectes
0
2000
4000
6000
8000
10000
12000
No
Sí
NA's
Resposta


Freqüència de fumar
Número de subjectes
0
2000
4000
6000
8000
10000
12000
14000
1. Molt alta
2. Alta
3. Mitjana
4. Baixa
5. Molt baixa
NA's
Resposta



6.3 Matrius de confusió i estadístics d'avaluació dels classificadors SVMs generats

Resultats del SVM sense utilitzar cap algoritme de mostratge i amb un *kernel* lineal:

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                Diagnosticat Sense aquesta condició
##   Diagnosticat                      42                      54
##   Sense aquesta condició            57                     443
##
##                Accuracy : 0.8138
##                  95% CI : (0.7801, 0.8442)
##     No Information Rate : 0.8339
##     P-Value [Acc > NIR] : 0.9140
##
##                   Kappa : 0.3195
##
##  Mcnemar's Test P-Value : 0.8494
##
##             Sensitivity : 0.42424
##             Specificity : 0.89135
##          Pos Pred Value : 0.43750
##          Neg Pred Value : 0.88600
##              Prevalence : 0.16611
##          Detection Rate : 0.07047
##    Detection Prevalence : 0.16107
##       Balanced Accuracy : 0.65780
##
##        'Positive' Class : Diagnosticat
```

Resultats del SVM sense utilitzar cap algoritme de mostratge i amb un *kernel* gaussià:

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                Diagnosticat Sense aquesta condició
##   Diagnosticat                      35                       0
##   Sense aquesta condició            64                     497
##
##                Accuracy : 0.8926
```

```
##                 95% CI : (0.8649, 0.9163)
##     No Information Rate : 0.8339
##     P-Value [Acc > NIR] : 3.231e-05
##
##                   Kappa : 0.477
##
##  Mcnemar's Test P-Value : 3.407e-15
##
##             Sensitivity : 0.35354
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 0.88592
##              Prevalence : 0.16611
##          Detection Rate : 0.05872
##    Detection Prevalence : 0.05872
##       Balanced Accuracy : 0.67677
##
##        'Positive' Class : Diagnosticat
```

Resultats del SVM utilitzant l'algoritme de mostratge SMOTE i amb un *kernel* lineal:

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction               Diagnosticat Sense aquesta condició
##   Diagnosticat                    134                      92
##   Sense aquesta condició          163                     304
##
##                Accuracy : 0.632
##                  95% CI : (0.5949, 0.668)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.0006724
##
##                   Kappa : 0.2256
##
##  Mcnemar's Test P-Value : 1.168e-05
##
##             Sensitivity : 0.4512
##             Specificity : 0.7677
##          Pos Pred Value : 0.5929
##          Neg Pred Value : 0.6510
##              Prevalence : 0.4286
##          Detection Rate : 0.1934
##    Detection Prevalence : 0.3261
##       Balanced Accuracy : 0.6094
##
##        'Positive' Class : Diagnosticat
```

Resultats del SVM utilitzant l'algoritme de mostratge SMOTE i amb un *kernel* gaussià:

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction               Diagnosticat Sense aquesta condició
##   Diagnosticat                    133                      44
##   Sense aquesta condició          164                     352
##
##                Accuracy : 0.6999
##                  95% CI : (0.6642, 0.7338)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 2.264e-12
##
##                   Kappa : 0.3546
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4478
##             Specificity : 0.8889
```

```
##          Pos Pred Value : 0.7514
##          Neg Pred Value : 0.6822
##              Prevalence : 0.4286
##          Detection Rate : 0.1919
##    Detection Prevalence : 0.2554
##       Balanced Accuracy : 0.6684
##
##        'Positive' Class : Diagnosticat
```

6.4 Resultats de l'avaluació del rendiment dels models de classificació XGBoost utilitzant diferents mètodes de mostratge i paràmetres d'entrenament

Mètode de mostratge	Paràmetres	Exactitud	Kappa	Sensibilitat	Especificitat
Cap mètode	max.depth = 1, nrounds = 10	0.8742	0.3699	0.2727	0.9940
Cap mètode	max.depth = 1, nrounds = 50	0.8809	0.4352	0.3434	0.9879
Cap mètode	max.depth = 1, nrounds = 150	0.8842	0.4735	0.3939	0.9819
Cap mètode	max.depth = 1, nrounds = 300	0.8742	0.4501	0.4040	0.9678
Cap mètode	max.depth = 1, nrounds = 400	0.8641	0.4175	0.3939	0.9577
Cap mètode	max.depth = 1, nrounds = 500	0.8607	0.4144	0.4040	0.9517
Cap mètode	max.depth = 1, nrounds = 600	0.8574	0.4059	0.4040	0.9477
Cap mètode	max.depth = 1, nrounds = 1000	0.8557	0.4179	0.4343	0.9396
Cap mètode	max.depth = 2, nrounds = 10	0.8826	0.4217	0.3131	0.9960
Cap mètode	max.depth = 2, nrounds = 50	0.8943	0.5041	0.3939	0.9940
Cap mètode	max.depth = 2, nrounds = 150	0.8943	0.5193	0.4242	0.9879
Cap mètode	max.depth = 2, nrounds = 300	0.8909	0.5040	0.4141	0.9859
Cap mètode	max.depth = 2, nrounds = 400	0.8859	0.4940	0.4242	0.9779
Cap mètode	max.depth = 2, nrounds = 500	0.8842	0.4891	0.4242	0.9759
Cap mètode	max.depth = 2, nrounds = 600	0.8842	0.4891	0.4242	0.9759
Cap mètode	max.depth = 2, nrounds = 1000	0.8859	0.4990	0.4343	0.9759
Cap mètode	max.depth = 3, nrounds = 10	0.8809	0.4230	0.3232	0.9920
Cap mètode	max.depth = 3, nrounds = 50	0.8960	0.5042	0.3838	0.9980
Cap mètode	max.depth = 3, nrounds = 150	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 3, nrounds = 300	0.8977	0.5199	0.4040	0.9960
Cap mètode	max.depth = 3, nrounds = 400	0.8977	0.5199	0.4040	0.9960
Cap mètode	max.depth = 3, nrounds = 500	0.8977	0.5199	0.4040	0.9960
Cap mètode	max.depth = 3, nrounds = 600	0.8977	0.5199	0.4040	0.9960
Cap mètode	max.depth = 3, nrounds = 1000	0.8993	0.5302	0.4141	0.9960
Cap mètode	max.depth = 4, nrounds = 10	0.8876	0.4670	0.3636	0.9920
Cap mètode	max.depth = 4, nrounds = 50	0.8909	0.4884	0.3838	0.9920
Cap mètode	max.depth = 4, nrounds = 150	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 4, nrounds = 300	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 4, nrounds = 400	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 4, nrounds = 500	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 4, nrounds = 600	0.8977	0.5199	0.4040	0.9960
Cap mètode	max.depth = 4, nrounds = 1000	0.8977	0.5199	0.4040	0.9960
Cap mètode	max.depth = 5, nrounds = 10	0.8926	0.4882	0.3737	0.9960
Cap mètode	max.depth = 5, nrounds = 50	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 5, nrounds = 150	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 5, nrounds = 300	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 5, nrounds = 400	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 5, nrounds = 500	0.8977	0.5148	0.3939	0.9980

Cap mètode	max.depth = 5, nrounds = 600	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 5, nrounds = 1000	0.8943	0.5041	0.3939	0.9940
Cap mètode	max.depth = 6, nrounds = 10	0.8792	0.4362	0.3535	0.9839
Cap mètode	max.depth = 6, nrounds = 50	0.8926	0.4989	0.3939	0.9920
Cap mètode	max.depth = 6, nrounds = 150	0.8909	0.4937	0.3939	0.9899
Cap mètode	max.depth = 6, nrounds = 300	0.8909	0.4937	0.3939	0.9899
Cap mètode	max.depth = 6, nrounds = 400	0.8926	0.4989	0.3939	0.9920
Cap mètode	max.depth = 6, nrounds = 500	0.8943	0.5041	0.3939	0.9940
Cap mètode	max.depth = 6, nrounds = 600	0.8926	0.4989	0.3939	0.9920
Cap mètode	max.depth = 6, nrounds = 1000	0.8926	0.4989	0.3939	0.9920
Cap mètode	max.depth = 7, nrounds = 10	0.8792	0.4303	0.3434	0.9859
Cap mètode	max.depth = 7, nrounds = 50	0.8926	0.4936	0.3838	0.9940
Cap mètode	max.depth = 7, nrounds = 150	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 7, nrounds = 300	0.8943	0.5041	0.3939	0.9940
Cap mètode	max.depth = 7, nrounds = 400	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 7, nrounds = 500	0.8943	0.5041	0.3939	0.9940
Cap mètode	max.depth = 7, nrounds = 600	0.8943	0.5041	0.3939	0.9940
Cap mètode	max.depth = 7, nrounds = 1000	0.8960	0.5094	0.3939	0.9960
Cap mètode	max.depth = 8, nrounds = 10	0.8758	0.4082	0.3232	0.9859
Cap mètode	max.depth = 8, nrounds = 50	0.8943	0.4989	0.3838	0.9960
Cap mètode	max.depth = 8, nrounds = 150	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 8, nrounds = 300	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 8, nrounds = 400	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 8, nrounds = 500	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 8, nrounds = 600	0.8977	0.5148	0.3939	0.9980
Cap mètode	max.depth = 8, nrounds = 1000	0.8960	0.5042	0.3838	0.9980
SMOTE	max.depth = 1, nrounds = 10	0.7258	0.4105	0.4781	0.9116
SMOTE	max.depth = 1, nrounds = 50	0.7807	0.5370	0.6128	0.9066
SMOTE	max.depth = 1, nrounds = 150	0.8355	0.6572	0.7273	0.9167
SMOTE	max.depth = 1, nrounds = 300	0.8297	0.6464	0.7340	0.9015
SMOTE	max.depth = 1, nrounds = 400	0.8240	0.6350	0.7340	0.8914
SMOTE	max.depth = 1, nrounds = 500	0.8268	0.6404	0.7306	0.8990
SMOTE	max.depth = 1, nrounds = 600	0.8225	0.6316	0.7273	0.8939
SMOTE	max.depth = 1, nrounds = 1000	0.8153	0.6184	0.7374	0.8737
SMOTE	max.depth = 2, nrounds = 10	0.7302	0.4200	0.4848	0.9141
SMOTE	max.depth = 2, nrounds = 50	0.8139	0.6096	0.6768	0.9167
SMOTE	max.depth = 2, nrounds = 150	0.8153	0.6131	0.6835	0.9141
SMOTE	max.depth = 2, nrounds = 300	0.8254	0.6329	0.6801	0.9343
SMOTE	max.depth = 2, nrounds = 400	0.8182	0.6175	0.6700	0.9293
SMOTE	max.depth = 2, nrounds = 500	0.8196	0.6204	0.6700	0.9318
SMOTE	max.depth = 2, nrounds = 600	0.8196	0.6204	0.6700	0.9318
SMOTE	max.depth = 2, nrounds = 1000	0.8124	0.6050	0.6599	0.9268
SMOTE	max.depth = 3, nrounds = 10	0.7475	0.4606	0.5286	0.9116
SMOTE	max.depth = 3, nrounds = 50	0.7965	0.5673	0.6027	0.9419
SMOTE	max.depth = 3, nrounds = 150	0.8153	0.6080	0.6330	0.9520
SMOTE	max.depth = 3, nrounds = 300	0.8240	0.6277	0.6566	0.9495
SMOTE	max.depth = 3, nrounds = 400	0.8240	0.6274	0.6532	0.9520
SMOTE	max.depth = 3, nrounds = 500	0.8225	0.6242	0.6498	0.9520
SMOTE	max.depth = 3, nrounds = 600	0.8225	0.6242	0.6498	0.9520
SMOTE	max.depth = 3, nrounds = 1000	0.8139	0.6062	0.6431	0.9419

SMOTE	max.depth = 4, nrounds = 10	0.7648	0.5054	0.6094	0.8813
SMOTE	max.depth = 4, nrounds = 50	0.8384	0.6612	0.7071	0.9369
SMOTE	max.depth = 4, nrounds = 150	0.8384	0.6603	0.6970	0.9444
SMOTE	max.depth = 4, nrounds = 300	0.8355	0.6545	0.6970	0.9394
SMOTE	max.depth = 4, nrounds = 400	0.8355	0.6545	0.6970	0.9394
SMOTE	max.depth = 4, nrounds = 500	0.8341	0.6511	0.6902	0.9419
SMOTE	max.depth = 4, nrounds = 600	0.8341	0.6514	0.6936	0.9394
SMOTE	max.depth = 4, nrounds = 1000	0.8341	0.6511	0.6902	0.9419
SMOTE	max.depth = 5, nrounds = 10	0.7792	0.5317	0.5926	0.9192
SMOTE	max.depth = 5, nrounds = 50	0.8283	0.6380	0.6734	0.9444
SMOTE	max.depth = 5, nrounds = 150	0.8297	0.6415	0.6801	0.9419
SMOTE	max.depth = 5, nrounds = 300	0.8312	0.6447	0.6835	0.9419
SMOTE	max.depth = 5, nrounds = 400	0.8297	0.6418	0.6835	0.9394
SMOTE	max.depth = 5, nrounds = 500	0.8254	0.6322	0.6734	0.9394
SMOTE	max.depth = 5, nrounds = 600	0.8268	0.6354	0.6768	0.9394
SMOTE	max.depth = 5, nrounds = 1000	0.8240	0.6284	0.6633	0.9444
SMOTE	max.depth = 6, nrounds = 10	0.7590	0.4880	0.5623	0.9066
SMOTE	max.depth = 6, nrounds = 50	0.8124	0.6016	0.6263	0.9520
SMOTE	max.depth = 6, nrounds = 150	0.8196	0.6171	0.6364	0.9571
SMOTE	max.depth = 6, nrounds = 300	0.8225	0.6235	0.6431	0.9571
SMOTE	max.depth = 6, nrounds = 400	0.8182	0.6145	0.6397	0.9520
SMOTE	max.depth = 6, nrounds = 500	0.8225	0.6232	0.6397	0.9596
SMOTE	max.depth = 6, nrounds = 600	0.8225	0.6232	0.6397	0.9596
SMOTE	max.depth = 6, nrounds = 1000	0.8182	0.6132	0.6263	0.9621
SMOTE	max.depth = 7, nrounds = 10	0.7662	0.5052	0.5859	0.9015
SMOTE	max.depth = 7, nrounds = 50	0.7994	0.5746	0.6162	0.9369
SMOTE	max.depth = 7, nrounds = 150	0.8023	0.5803	0.6162	0.9419
SMOTE	max.depth = 7, nrounds = 300	0.8095	0.5958	0.6263	0.9470
SMOTE	max.depth = 7, nrounds = 400	0.8110	0.5990	0.6296	0.9470
SMOTE	max.depth = 7, nrounds = 500	0.8081	0.5929	0.6263	0.9444
SMOTE	max.depth = 7, nrounds = 600	0.8095	0.5962	0.6296	0.9444
SMOTE	max.depth = 7, nrounds = 1000	0.8095	0.5947	0.6162	0.9545
SMOTE	max.depth = 8, nrounds = 10	0.7821	0.5390	0.6061	0.9141
SMOTE	max.depth = 8, nrounds = 50	0.8066	0.5911	0.6364	0.9343
SMOTE	max.depth = 8, nrounds = 150	0.8066	0.5918	0.6431	0.9293
SMOTE	max.depth = 8, nrounds = 300	0.8095	0.5976	0.6431	0.9343
SMOTE	max.depth = 8, nrounds = 400	0.8095	0.5976	0.6431	0.9343
SMOTE	max.depth = 8, nrounds = 500	0.8110	0.6008	0.6465	0.9343
SMOTE	max.depth = 8, nrounds = 600	0.8110	0.6008	0.6465	0.9343
SMOTE	max.depth = 8, nrounds = 1000	0.8095	0.5965	0.6330	0.9419

6.5 Codi del diagrama de Gantt i de les tasques corresponents al primer objectiu del TFM

El codi utilitzat per realitzar el diagrama de Gantt utilitzat en la temporalització de la planificació del TFM i per dur a terme les tasques corresponents al primer objectiu d'aquest projecte es troba al següent enllaç de la plataforma *GitLab*. Les diferents seccions estan referenciades en el cos del codi utilitzant format RMarkdown:

https://gitlab.com/jcanetcarbo/tfm_bioinfo_biostats/blob/master/_PROG/Codi_TFM_1.Rmd

6.6 Codi de les tasques corresponents al segon objectiu del TFM

El codi utilitzat per dur a terme les tasques corresponents al segon objectiu d'aquest projecte es troba al següent enllaç de la plataforma *GitLab*. Les diferents seccions estan referenciades en el cos del codi utilitzant format RMarkdown:

https://gitlab.com/jcanetcarbo/tfm_bioinfo_biostats/blob/master/_PROG/Codi_TFM_2.Rmd