Universitat Oberta de Catalunya

uoc.edu

Registres de salut digitals: Tractament de dades i construcció de models de predicció de malaltia.

Sílvia Cros Roura

Máster de bioinformàtica i bioestadística

Àrea d'anàlisi de dades i bioestadística

Consultors: Núria Pérez Álvarez i Esteban Vegas

Professor responsable de l'assignatura: Marc Maceira Duch

Juny 2021

FITXA DEL TREBALL FINAL

Títol del treball:	Registres de salut digitals: Tractament de dades i construcció de models de predicció de malaltia.
Nom de l'autor:	Sílvia Cros Roura
Nom del consultor/a:	Núria Pérez Álvarez i Esteban Vegas
Nom del PRA:	Marc Maceira Duch
Data de lliurament (mm/aaaa):	Juny 2021
Titulació:	Màster bioinformàtica i bioestadística
Àrea del Treball Final:	Àrea d'anàlisi de dades i bioestadística
Idioma del treball:	Català
Nombre de crèdits:	15
Paraules clau	Registres de salut digitals, assajos clínics, artritis reumatoide, factors de risc, models de predicció

Resum del Treball (màxim 250 paraules):

Els objectius principals d'aquest TFM han consistit en conèixer els registres de salut digitals , aprendre a tractar les dades que se'n deriven i simular un estudi de predicció de risc dirigit a l'Artritis Reumatoide.

Per acomplir tals fites, s'ha utilitzat una base de dades amb informació de 100.000 pacients simulats (*EMRbot*s), la qual s'ha adaptat i reduït a 28.572 pacients, la meitat dels quals estaven diagnosticats amb Artritis Reumatoide.

L'anàlisi estadístic s'ha dividit en dos parts: la primera part ha estat enfocada a la identificació de factors de risc associats amb la malaltia en qüestió, recollint que l' hipoalbuminèmia, la proteïnèmia, l'anèmia, la leucositosi, l'hiperplaquetosis, el gènere femení i una edat superior a 45 anys podrien tenir una implicació directe amb aquesta. La segona part ha consistit en utilitzar aquestes variables per entrenar i executar diferents models i algoritmes de classificació de pacients segons diagnòstic. Els models seleccionats han sigut: Regressió logística múltiple, Algoritme de *Naïve Bayes*, *Random Forest*, SVM i ANN. S'ha estimat la seva actuació a partir de diferents paràmetres, tals i com són les corbes ROC, la precisió o l'AUC. En excepció de *Naïve Bayes*, tota la resta de models ha presentat una bona actuació, pel que s'ha conclòs que qualsevol d'ells és vàlid per problemes de classificació de malaltia a partir de predictors numèrics i categòrics.

No obstant això, cal tenir present que tot el treball ha sigut simulat i que les conclusions no són extrapolables a nivell real.

Abstract (in English, 250 words or less):

The main objectives pursued by this TFM were how to deal with data derived from Electronic Health Records and to simulate a prediction study for classifying patients according to the risk of developing rheumatoid arthritis using this data.

For that, a 100.000 virtual patients' dataset was used. This cohort was adapted and reduced to 28.572 patients; half diagnosed with RA.

A statistical analysis was performed, with two differentiated parts: the first one consisted in the identification of risk factors associated with the disease, recognizing hypoalbuminemia, proteinemia, anaemia, leucocytosis, thrombocytosis, the feminine gender, and an age over than 45 years as correlated variables. The second part used these factors to build and execute the different selected models, which were: Logistic multiple regression, *Naïve Bayes* algorithm, *Random Forest*, SVM and ANN. Their performance was evaluated by means of different parameters, such as: ROC curves, accuracy, and AUC. Excluding *Naïve Bayes*, all the other models showed a good performance, hence, all are considered acceptable to be used in classification problems based in numeric and categoric predictors.

Nevertheless, it must be taken into account that the work was done with simulated data, therefore, the conclusions are not comparably to the real patients.

Índex

1.	**Introducció**	**9**
1.1.	Context i justificació del Treball	10
1.2.	Objectius del Treball	12
1.3.	Enfocament	13
1.4.	Planificació del Treball	14
1.5.	Mètodes d'anàlisi estadístic	20
1.6.	Breu sumari de contribucions i productes obtinguts	21
1.7.	Breu descripció dels altres capítols de la memòria	22
2.	**Identificació de factors de risc**	**23**
2.1.	Tractament i adequació inicial de les dades provinents d'EMR	24
2.2.	Anàlisi descriptiu de les dades EMR	34
2.3.	Anàlisi estadístic de les variables	48
3.	**Algoritmes i models de predicció de malaltia**	**60**
3.1	Descripció dels models i algoritmes utilitzats en anàlisi de predicció de risc	61
3.2.	Formulació i execució dels models	66
3.3.	Comparació de l'actuació dels models	83
3.4.	Informe reproduïble i aplicació *Shiny*	85
4.	**Conclusions**	**88**
5.	**Glossari**	**90**
6.	**Bibliografia**	**91**
7.	**Annexos**	**93**

Llista de figures

Figura nº 1. Diagrama de Gantt amb la programació de les tasques. 16
Figura nº 2. Fitxers que conformen l'*EMRbot* de 100.000 pacients. 26
Figura nº 3. Primer esborrany de la base de dades post- selecció de pacients. 29
Figura nº 4. Visualització parcial de la base de dades preparada per l'anàlisi. 32
Figura nº 5. Passos realitzats per la conformació de l'Excel per l'anàlisi amb *Rstudio*. 33
Figura nº 6. Visualització de la frequencia i els patrons dels *missing values.* 36
Figura nº 7. Nº de pacients segons diagnòstic. 38
Figura nº 8. Nº de pacients totals (A) i diagnosticats amb AR (B) segons nº d'admissió. 39
Figura nº 9. Nº de pacients segons génere(A), raça(B), estat civil(C) i idioma parlat (D). 39
Figura nº 10. Descripció gràfica de les variables continues relatives als pacients. 40
Figura nº 11. Distribució dels paramètres de laboratori mesurats amb mg/dl. 40
Figura nº 12. Distribució dels paramètres de laboratori mesurats amb U/L. 41
Figura nº 13. Distribució dels paramètres de laboratori mesurats amb mmol/L. 41
Figura nº 14. Distribució dels paramètres de laboratori mesurats amb gm/dl. 41
Figura nº 15. Distribució dels paramètres de laboratori mesurats unitats de percentatge. 42
Figura nº 16. Distribució del recompte de plaquetes, expressat en k/cumm. 42
Figura nº 17. Distribució del VCM, expressat en fl. 42
Figura nº 18. Distribució del pH de l'orina. 43
Figura nº 19. Relació diagnòstic amb gènere i raça dels pacients segons admissió. 44
Figura nº 20. Gràfic que relaciona el diagnòstic i l'edat dels pacients. 45
Figura nº 21. Relació entre diagnòstic i paràmetres de laboratori (1). 45
Figura nº 22. Relació entre diagnòstic i paràmetres de laboratori (2). 46
Figura nº 23. Relació entre diagnòstic i paràmetres de laboratori (3). 46
Figura nº 24. Relació diagnòstic i paràmetres de laboratori(4). 47
Figura nº 25. QQ-plot de l'edat dels pacients segons grup de diagnòstic. 48
Figura nº 26. QQ-plot de la durada de l'ingrès hospitalari dels pacients. 49
Figura nº 27. QQ-plots dels paràmetres expressats en mg/dl segons grup de diagnòstic. 50
Figura nº 28. QQ-plots dels paràmetres expressats en U/L segons grup de diagnòstic. 51
Figura nº 29. QQ-plots dels paràmetres expressats en mmol/L segons grup de diagnòstic. 52
Figura nº 30. QQ-plots dels paràmetres expressats en gm/dl segons grup de diagnòstic. 53
Figura nº 31. QQ-plots dels paràmetres expressats en % segons grup de diagnòstic 54
Figura nº 32. QQ-plot del recompte de plaquetes segons grup de diagnòstic. 55
Figura nº 33. QQ-plot del VCM segons grup de diagnòstic. 56
Figura nº 34. QQ-plot del pH de l'orina segons grup de diagnòstic. 57
Figura nº 35. Importància de les variables dins el model. 69
Figura nº 36. Representació del model ANN generat. 70
Figura nº 37. Corba ROC del model de regressió logística múltiple. 72
Figura nº 38. Corba ROC del model *Naïve Bayes* amb *Laplace* =0. 73
Figura nº 39. Corba ROC de l'algoritme *Random Forest* amb *ntrees*=500. 73
Figura nº 40. Corba ROC de l'algoritme SVM lineal. 74
Figura nº 41. Corba ROC de l'algoritme ANN amb una neurona a la capa oculta. 74
Figura nº 42. Corba ROC del model de *Naive Bayes* amb *Laplace*=1. 75
Figura nº 43. Corba ROC del model de *Random Forest* amb *ntrees*=1000. 76
Figura nº 44. Corba ROC de l'algoritme SVM lineal. 77
Figura nº 45. Representació dels model ANN amb 3 i 5 nodes a la capa oculta. 78
Figura nº 46. Corbes ROC de l'algoritme ANN amb 3 i 5 nodes a la capa oculta. 79
Figura nº 47. Pestanya 1 de l'aplicació Shiny creada. 86

Figura nº 48. Pestanya 2 de l'aplicació Shiny creada. 87
Figura nº 49. Pestanya 3 de l'aplicació Shiny creada. 87

Llista de taules

Taula nº 1. Paquets d'R disponibles per l'anàlisi de dades EMR. 24
Taula nº 2. Variables que conformen la base de dades. 34
Taula nº 3. Variables amb valors mancants. 35
Taula nº 4. Percentatge que representen els valors mancants dins de cada variable. 35
Taula nº 5. Resum de la comparació entre grups de l'edat dels pacients. 48
Taula nº 6. Resum comparació entre grups de la durada dels ingressos hospitalaris. 49
Taula nº 7. Resum comparació entre grups dels paràmetres expressats en mg/dL. 50
Taula nº 8. Resum comparació entre grups dels paràmetres expressats en U/L. 51
Taula nº 9. Resum comparació entre grups dels paràmetres expressats en mmol/L. 52
Taula nº 10. Resum comparació entre grups dels paràmetres expressats en gm/dL. 53
Taula nº 11. Resum comparació entre grups dels paràmetres expressats en %. 54
Taula nº 12. Resum comparació entre grups del recompte de plaquetes. 55
Taula nº 13. Resum de la comparació entre grups del VCM. 56
Taula nº 14. Resum de la comparació entre grups del pH de l'orina. 57
Taula nº 15. Taula de contingència segons gènere i diagnòstic. 58
Taula nº 16. Taula de contingència segons diagnòstic i raça. 58
Taula nº 17. Chi quadrat i comparacions post-hoc de l'associació raça -diagnostic. 58
Taula nº 18. Recopilació conclusions de la comparació de variables. 59
Taula nº 19. Taula de fortaleses i debilitats de la regressió logística múltiple. 62
Taula nº 20. Taula de fortaleses i debilitats de l'algoritme de *Naive Bayes* . 63
Taula nº 21. Taula de fortaleses i debilitats de l'algoritme de *Random Forest*. 63
Taula nº 22. Taula de fortaleses i debilitats de l'algoritme de l'algoritme SVM. 64
Taula nº 23. Taula de fortaleses i debilitats de l'algoritme de l'algoritme ANN. 65
Taula nº 24. Variables seleccionades per la construcció de models de predicció 66
Taula nº 25. Coeficients i p-valor obtinguts amb la regressió logística múltiple. 67
Taula nº 26. Variables normalitzades i transformades per l'algoritme ANN. 70
Taula nº 27 . Taula de classificació de l'acord segons el valor kappa. 71
Taula nº 28. Taula d'interpretació de l'actuació del model segons el valor AUC. 71
Taula nº 29. Paràmetres mesura actuació del model regressió logística múltiple. 72
Taula nº 30. Paràmetres mesura actuació del model *Naïve Bayes* amb *Laplace* =0. 73
Taula nº 31. Paràmetres mesura actuació de *Random Forest ntrees*=500. 73
Taula nº 32. Paràmetres mesura actuació de l'algoritme SVM lineal. 74
Taula nº 33. Paràmetres mesura actuació ANN amb una neurona a la capa oculta. 74
Taula nº 34. Paràmetres mesura actuació del model *Naive Bayes* amb *Laplace*=1. 75
Taula nº 35. Paràmetres mesura actuació del model *Random Forest ntrees*=1000. 76
Taula nº 36. Paràmetres mesura actuació de l'algoritme SVM radial. 77
Taula nº 37. Paràmetres mesura actuació ANN amb 3 i 5 nodes a la capa oculta. 79
Taula nº 38. Mesures actuació del model de regressió logística múltiple validat. 80
Taula nº 39. Mesures actuació del model *Naïve Bayes* validat. 81
Taula nº 40. Mesures actuació del model *Random Forest* validat. 81
Taula nº 41. Mesures actuació del model SVM validat. 81
Taula nº 42. Precisió i valor kappa segons diferents valors de size i decay. 82
Taula nº 43. Mesures actuació del model ANN validat. 82
Taula nº 44. Mesures actuació de tots els models de predicció executats. 83
Taula nº 45. Avantatges i inconvenients d'una aplicació *Shiny.* 85

1. INTRODUCCIÓ

1.1. CONTEXT I JUSTIFICACIÓ DEL TREBALL

La predicció de malalties a partir de paràmetres clínics i laboratorials, així com la identificació primerenca de futurs pacients, són motius de gran interès en el món de les ciències de la salut. No només com a eines preventives, sinó també com a eines epidemiològiques i econòmiques.

El notable increment de la utilització de registres de salut digitals (EHRs o EMRs, en anglès) ha aportat grans avantatges en aquesta àrea .

Els EMR són versions digitals de les històries clíniques dels pacients. No només faciliten el seguiment i el tractament d'aquests, sinó que el fet d'emmagatzemar moltes variables i informació clínica, comporta que acabin constituint bases de dades molt potents. Aquest fet, permet grans oportunitats al món de la investigació, ja sigui en la realització d'assajos clínics, en recerca genètica o bé en el desenvolupament i refinament algoritmes i models de predicció[4].
Hi ha múltiples avantatges en realitzar prediccions de risc basades en EMR, ja que permeten als investigadors observar més variables, en més individus, en molts més temps i amb menys costos que els estudis tradicionals de cohorts i casos-controls.
No obstant, també hi ha limitacions i molts factors a tenir en compte durant un anàlisi de dades EMR. Per exemple, es necessita molta prudència a l'hora de realitzar inferències, degut a la incompletesa i soroll de les dades, a part, de múltiples fonts de biaix. Segons Kohane et al[9], hi ha sis qüestions a considerar a l'hora de realitzar un anàlisi amb dades EMR:

- Com de completes són les dades? – Per exemple, no tota la informació es troba inclosa en registres digitals, pel que idealment s'hauria d'acabar de completar utilitzant altres fonts. Per altre banda, tot i que teòricament s'incloguin determinades variables, no sempre es registren, pel que la falta d'informació pot comportar malentesos si no s'indica apropiadament.

- Com es recullen i manegen aquestes dades? – Les unitats de mesura o l'escala d'una variable pot diferir entre hospitals, època/any recollida, edat del pacient… Per això caldria intentar estandarditzar i harmonitzar al màxim totes les variables que s'inclouen a l'estudi per intentar disminuir la variabilitat.

- Quin tipus de dades hi ha recollides? – Tant es poden trobar dades codificades com text narratiu. S'haurien d'utilitzar mètodes d'anàlisi que tinguin en compte els diferents tipus de dades per tal de millorar la sensibilitat i l'especificitat dels resultats.

- Es té en compte la variabilitat entre EMR durant l'anàlisi? – Pot variar molt la informació que s'inclou en cada EMR depenent del país, hospital, pràctica mèdica… Això dificulta molt poder obtenir resultats concloents i aplicables a nivell global.

- Són les dades transparents? – Tot i la confidencialitat associada a aquest tipus de dades, els codis i algoritmes utilitzats durant els anàlisi s'haurien de poder compartir en repositoris públics. Això permetria seguir els passos realitzats a altres investigadors i confirmar la transparència del procés.

- L'estudi és multidisciplinari? – Idealment, un estudi amb dades EMR hauria d'incloure col·laboracions amb científics i clínics que coneguin la malaltia, informàtics, experts amb estandardització de dades, epidemiòlegs, estadistes… Cada expert podria contribuir aportant millores en un camp determinat, el que acabaria resultant en un estudi complet i robust.

Per tant, s'ha de tenir molta cura en aquests anàlisi, ja que conclusions massa generalitzades poden conduir a resultats erronis i a hipòtesis falses.

Els objectius principals d'aquest TFM són conèixer els registres de salut digitals , aprendre a tractar les dades que se'n deriven i simular un estudi de predicció de risc dirigit a l'Artritis Reumatoide.

Per això, s'escull una base de dades virtual, anomenada *EMRbot*[8], amb informació de 100.000 pacients, uns 17.000 dels quals, diagnosticats d'Artritis reumatoide.

L'Artritis reumatoide (AR) és una malaltia autoimmune caracteritzada per inflamació crònica, deteriorament progressiu de les articulacions, reducció de la mobilitat… que pot derivar en complicacions cardiovasculars, i fins i tot, mortalitat prematura. Va ser descrita per *Cobb et al*, l'any 1953, i des de llavors, s'han reportat nombroses evidencies que demostren que els pacients d'artritis reumatoide poden arribar a tenir un 50% de risc de mortalitat prematura i que la seva esperança de vida s'ha vist disminuïda de 3 a 10 anys en comparació amb la població general[11].
És àmpliament reconegut que a part dels factors genètics, també hi ha nombrosos factors ambientals, com poden ser la radiació ultraviolada i el fet de ser fumador, que participen en el desenvolupament de malalties autoimmunes com l'AR[1]. Tot i així, encara hi ha una gran desconeixença dels factors que poden incidir i facilitar el seu desenvolupament.

D'aquesta manera, s'intenta identificar els factors de risc per aquesta malaltia i generar models de predicció capaços de classificar els pacients segons el risc de desenvolupar AR.

La temàtica escollida permet treballar molts dels conceptes apresos al llarg del màster, com per exemple: tractament de dades amb R, regressions, models lineals mixtos, models de *Machine Learning*, generació d'informes reproduïbles amb *Rmarkdown*….

1.2. OBJECTIUS DEL TREBALL

A continuació es presenten els objectius generals i específics que persegueix aquest projecte.

1.2.1. Objectius generals

1) Convertir un registre de salut digital (EMR) en una base de dades analitzable estadísticament.

2) Identificar possibles factors de risc que modulin la progressió de l'artritis reumatoide.

3) Comparar l'actuació de diferents algoritmes de predicció.

4) Generar un informe *Rmarkdown* reproduïble. Crear una aplicació *Shiny* que permeti la interacció amb les dades.

1.2.2. Objectius específics

1) Identificació de diferents paquets i codis d'*R* que ajudin al tractament de dades d'EHR.

2) Configuració de la base de dades escollida a Excel.

3) Tractament dels valors mancants, transformació de les variables i realització d'un anàlisi descriptiu de les dades.

4) Associació i relació entre variables/factors amb la el diagnòstic d'Artritis Reumatoide.

5) Definició dels algoritmes i models adequats per l'anàlisi.

6) Construcció models de predicció de malaltia.

7) Comparació dels resultats i l'actuació dels models.

8) Creació d' un informe reproduïble.

9) Realització d'una aplicació Shiny.

1.3. ENFOCAMENT

Hi ha dos estratègies a seguir per desenvolupar aquest projecte: La primera consisteix en una revisió sistemàtica de varis estudis i articles basats en el tema, mentre que la segona es basa en realitzar un treball pràctic a partir de dades reals o simulades.

Des del meu punt de vista, en cas de que hi hagi la possibilitat d'obtenció de dades és millor dur a terme la segona estratègia, ja que això permet tractar directament amb dades crues provinents d'EMR, preparar-les per l'anàlisi, lidiar amb els reptes que se'n deriven… A més a més, la comparació de diferents algoritmes de predicció és millor que es realitzi amb les mateixes dades i variables. La comparació a partir de la revisió sistemàtica podria no ser del tot adequada degut a les diferències en la recollida de les dades, diferències en les variables utilitzades en els models/algoritmes de predicció, etc.

Com que s'ha aconseguit disposar d'una base de dades simulada (*EMRbot*), amb una quantitat de variables i observacions considerable, es realitza un treball pràctic.

1.4. PLANIFICACIÓ DEL TREBALL

En aquest apartat es desglossen els objectius específics prèviament plantejats en una sèrie de tasques, que marquen el desenvolupament del treball. A més, es presenta una planificació sobre calendari (Figura nº1).

1.4.1. Tasques

Objectiu 1: Identificar paquets/codis d'*R* que ajudin al tractament de dades d'EMR.

- Tasca 1.1: Recerca de diferents paquets d'*R* que permetin processar dades EMR.
- Tasca 1.2: Classificar/agrupar els paquets d'*R* segons les funcionalitats que ofereixen.
- Tasca 1.3: Seleccionar els que es considerin més apropiats pel tipus de dades que es treballaran.

Objectiu 2: Configurar la base de dades escollida a l'arxiu *.xlsx* o *.txt* prèvia importació a R.

- Tasca 2.1: Seleccionar les observacions/pacients i les variables que es determinin apropiades per l'anàlisi. Incloure dos grups de pacients: El primer grup que presenti AR com a diagnòstic, i el segon grup que no presenti aquest diagnòstic.
- Tasca 2.2: Filtrar, simplificar i reduir la base de dades en base del seleccionat a la tasca anterior per facilitar-ne el maneig.
- Tasca 2.3: Agrupar els diferents arxius que conformen la base de dades en un sol arxiu. Importar l'arxiu a *R*.

Objectiu 3: Realitzar un anàlisi descriptiu de les dades.

- Tasca 3.1: Tractament dels *missing values.*
- Tasca 3.2: Normalitzar / transformar les variables que es consideri necessari.
- Tasca 3.3: Realitzar un anàlisi descriptiu de les dades.

Objectiu 4: Trobar associacions entre variables/factors i relacionar-los amb la resposta d'interès, en aquest cas, diagnòstic d'artritis reumatoide.

- Tasca 4.1: Comparar variables continues entre grups de pacients.
- Tasca 4.2: Comparar variables categòriques entre grups de pacients.
- Tasca 4.3: Realitzar taules comparatives entre grups de pacients.
- Tasca 4.4: Compilar i identificar totes les variables que tenen associació amb el diagnòstic d'Artritis reumatoide.

Objectiu 5: Definir quins algoritmes/models de predicció són els adequats per l'anàlisi que es vol realitzar.

- Tasca 5.1: Recerca bibliogràfica dels algoritmes més utilitzats en predicció de risc.
- Tasca 5.2: Realitzar una explicació dels diferents algoritmes i descriure els avantatges i inconvenients de cadascun d'ells.
- Tasca 5.3: Seleccionar justificadament els algoritmes adients per utilitzar en el treball.

Objectiu 6: Construir els models de predicció de malaltia per cadascun dels algoritmes/models prèviament seleccionats.

- Tasca 6.1: Formular els diferents models de predicció.
- Tasca 6.2: En el cas dels models de *Machine Learning*, realitzar una primera fase d'entrenament.
- Tasca 6.3: Predicció i avaluació dels models de predicció prèviament formulats.
- Tasca 6.4: Intentar millorar el rendiment dels models anteriors.
- Tasca 6.5: Validació dels models anteriors

Objectiu 7: Comparar els resultats i l'actuació dels models.

- Tasca 7.1: Realitzar una taula comparativa amb els resultats obtinguts dels diferents models.
- Tasca 7.2: Discutir quins models han presentat una millor actuació.

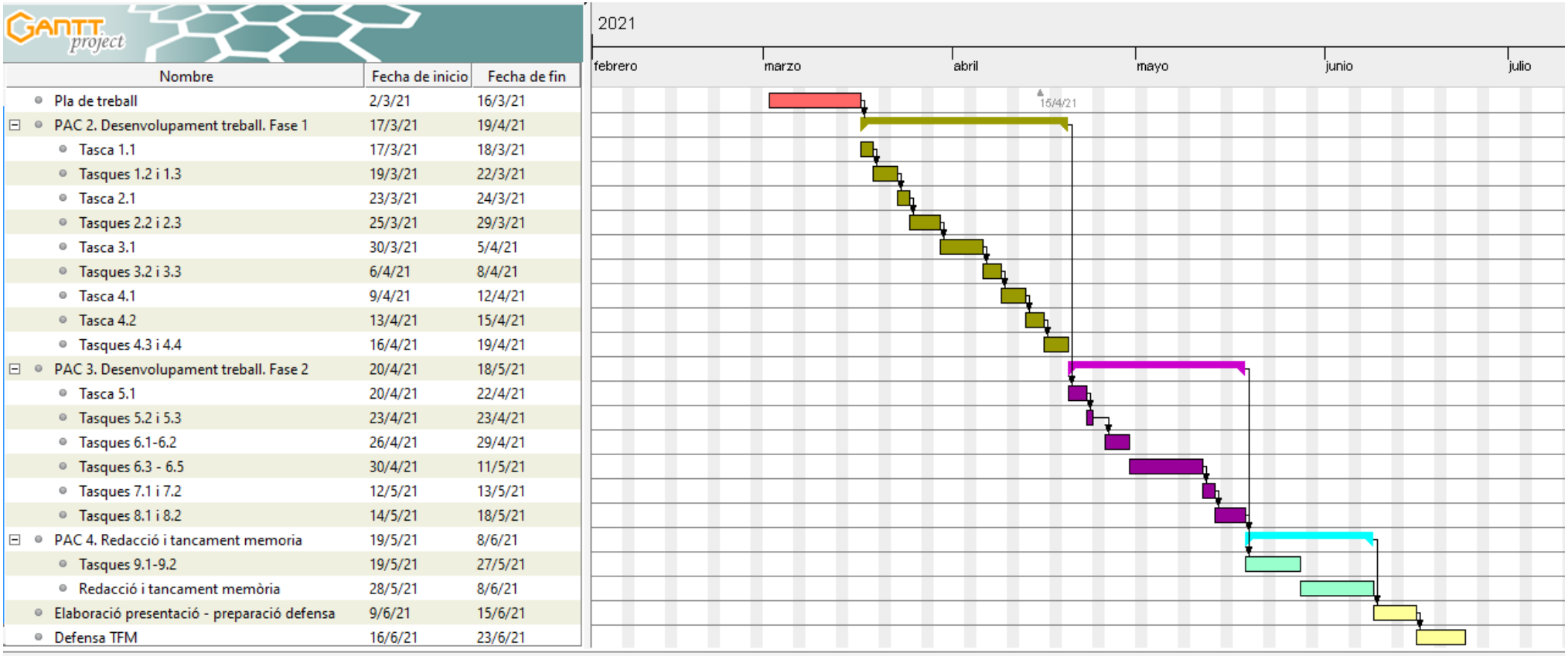
Objectiu 8: Crear un informe reproduïble.

- Tasca 8.1: Adaptar el codi *Rmarkdown* generat per tal de fer-lo dinàmic i adaptable a canvis en les dades.
- Tasca 8.2: Realitzar un informe amb *PDF* resultant de l'execució del codi *Rmarkdown* anterior.

Objectiu 9: Realitzar una aplicació *Shiny*

- Tasca 9.1: Explicar i descriure els avantatges i funcionalitats de les aplicacions *Shiny*. Justificar-ne l'ús en aquest projecte.
- Tasca 9.2: Creació de l'aplicació *Shiny.*

1.4.2. Calendari



Nombre	Fecha de inicio	Fecha de fin
Pla de treball	2/3/21	16/3/21
PAC 2. Desenvolupament treball. Fase 1	17/3/21	19/4/21
Tasca 1.1	17/3/21	18/3/21
Tasques 1.2 i 1.3	19/3/21	22/3/21
Tasca 2.1	23/3/21	24/3/21
Tasques 2.2 i 2.3	25/3/21	29/3/21
Tasca 3.1	30/3/21	5/4/21
Tasques 3.2 i 3.3	6/4/21	8/4/21
Tasca 4.1	9/4/21	12/4/21
Tasca 4.2	13/4/21	15/4/21
Tasques 4.3 i 4.4	16/4/21	19/4/21
PAC 3. Desenvolupament treball. Fase 2	20/4/21	18/5/21
Tasca 5.1	20/4/21	22/4/21
Tasques 5.2 i 5.3	23/4/21	23/4/21
Tasques 6.1-6.2	26/4/21	29/4/21
Tasques 6.3 - 6.5	30/4/21	11/5/21
Tasques 7.1 i 7.2	12/5/21	13/5/21
Tasques 8.1 i 8.2	14/5/21	18/5/21
PAC 4. Redacció i tancament memoria	19/5/21	8/6/21
Tasques 9.1-9.2	19/5/21	27/5/21
Redacció i tancament memòria	28/5/21	8/6/21
Elaboració presentació - preparació defensa	9/6/21	15/6/21
Defensa TFM	16/6/21	23/6/21

Figura nº1. Diagrama de Gantt amb la programació de les tasques.

1.4.3. Fites:

El següent quadre relaciona els objectius generals, els específics i les diferents tasques anteriorment mencionades. A part, també es marquen les fites per avançar al següent pas del projecte.

Objectiu general	Objectiu específic	Tasques	Fita	Data compleció
Convertir un registre mèdic electrònic en una base de dades analitzable estadísticament. **PAC 2**	Identificar paquets/codis d'*R* que ajudin al tractament de dades d'EMR.	Recerca de diferents paquets/codis d'*R*. Agrupament segons funcionalitat. Selecció dels més apropiats.	Quadre resum dels paquets R més coneguts per anàlisi EHR, agrupats per funcionalitat.	22/03/21
	Configurar la base de dades escollida.	Selecció dels pacients i variables apropiades per l'anàlisi. Polir la base de dades. Conformar un sol arxiu i importar-lo a R.	Obtenir un arxiu *.txt* o *.xlsx* polit i preparat pel posterior anàlisi a R.	29/03/21
Identificar possibles factors de risc que modulen la progressió d'AR. **PAC 2**	Inferir i millorar la qualitat de les dades.	Tractar els *missing vàlues*. Transformació de les variables. Realitzar un anàlisi descriptiu.	Resum descriptiu de les variables d'estudi.	8/04/21
	Trobar associacions entre variables/factors i relacionar-los amb la resposta d'interès.	Comparació variables continues. Comparació variables categòriques. Realització taula comparativa. Identificació variables associades resposta interès.	Taula comparativa entre grups de pacients. Remarcar variables que tinguin associació amb AR.	19/04/21

Objectiu general	Objectiu específic	Tasques	Fita	Data compleció
Comparació rendiment algoritmes de predicció. **PAC 3**	Definir algoritmes/models de predicció adequats	Recerca bibliogràfica. Explicació dels diferents algoritmes. Selecció dels algoritmes adients.	Taula resum amb característiques dels models predicció seleccionats.	23/04/21
	Construir els models de predicció de malaltia	Formulació models de predicció. Realitzar fase d'entrenament (si cal). Execució, predicció i avaluació dels models. Millora del rendiment dels models (si cal). Validació dels models	Obtenció resultats per tots els models de predicció empleats.	11/05/21
	Comparar els resultats i l'actuació dels models.	Realització una taula comparativa resultats. Discussió de les actuacions dels models.	Quadre comparatiu entre resultats dels models.	13/05/21
Generació informe *Rmarkdown* reproduïble **PAC 3**	Creació informe reproduïble	Adaptar el codi *Rmarkdown* generat . Realitzar un informe amb *PDF.*	Informe *PDF* automàtic.	18/05/21
Creació aplicació *Shiny* **PAC 4**	Realitzar una aplicació Shiny	Explicació aplicacions *Shiny*. Creació de l'aplicació *Shiny*.	Aplicació *Shiny* funcional.	18/05/21- 08/06/21

1.4.4. Anàlisi de riscos:

En aquest apartat es relaciona els objectius específics amb factors de risc que podrien contribuir negativament en el desenvolupament del projecte.

Objectiu específic	Risc associat
Identificar paquets/codis d'*R* que ajudin al tractament de dades d'EMR.	Paquets trobats no adients o obsolets.
Configurar la base de dades escollida.	Degut a la grandària de la base dades, problemes per obrir-la i/o treballar-hi amb els programes convencionals (Excel, bloc notes…).
Trobar associacions entre variables/factors i relacionar-los amb la resposta d'interès.	No trobar resultats suficientment concloents per definir associacions entre variables i la resposta d'interès, degut a que es tracta d'una base de dades generades de manera totalment aleatòria. No extrapolació en pacients reals.
Definir algoritmes/models de predicció adequats	Solapament de fases i tasques amb el següent objectiu.
Globalment	Incompatibilitat amb horari laboral, imprevistos personals i familiars, imprevistos tècnics… que provoquen un retràs.

1.5. MÈTODES D'ANÀLISI ESTADÍSTIC

A continuació, es descriu de forma resumida els passos que es segueixen pel tractament i l'anàlisi estadístic de les dades. L'explicació detallada es troba *adhoc* en els capítols, a mesura que es va avançant amb el projecte.

1.5.1. Obtenció de dades

S'obté un mínim de 10.000 pacients diagnosticats amb Artritis Reumatoide i el mateix número de pacients no diagnosticats d'AR de la base de dades *EMRbot*s (100.000 pacients inicials).
El criteri d'inclusió utilitzat és el número d'admissions hospitalàries, prioritzant els pacients diagnosticats d'AR durant les primeres admissions.
Per altre, es realitza un filtratge de les variables, per tal d'incloure només les considerades rellevants per l'anàlisi que es vol dur a terme. Finalment, es classifica els pacients en dos grups segons el seu diagnòstic.

1.5.2. Estudi de la població

Es realitza una descriptiva numèrica i gràfica (general i per grup d'interès) de la població d'estudi, com també comparacions entre grups amb els objectiu d'intentar establir una relació entre les diferents variables incloses i el diagnòstic d'Artritis Reumatoide.
En el cas de les variables numèriques, s'utilitza la mitjana i la desviació estàndard com a mesures bàsiques de comparació i així mateix, s'executa un test paramètric (*t-student*) o no paramètric (*Mann-Whitney*) en funció de la normalitat de les dades.
En el cas de les variables categòriques, es creen taules de contingència i computen testos estadístics adients per establir associacions com són la chi-quadrat o el test de Fisher.
Les variables que presenten significació estadística entre grups són les utilitzades en els models de predicció.

1.5.3. Models de predicció

Es construeix diferents models de predicció enfocats a la classificació dels pacients segons si tenen risc o no de desenvolupar AR. Per a tal fita, es seleccionen els models més adients segons bibliografia i es divideixen les dades en dos conjunts: entrenament i test. El primer serveix per l'aprenentatge dels models, és a dir, per ensenyar als models com classificar els pacients segons els valors de les seves variables, mentre que el segon conjunt s'utilitza per l'avaluació de l'actuació dels models, és a dir, per comprovar quin % de classificacions són correctes o no.
Es mesura l'actuació dels models a partir de diferents paràmetres, tals i com són les corbes ROC, la precisió o l'AUC, com igualment es testen alternatives per cadascun dels models a fi d'intentar millorar el seu rendiment.
Finalment, s'estudia la robustesa de tots els models amb mètodes de *k-fold cross validation i bootstrapping.*

Per l'anàlisi estadístic s'utilitza *Rstudio* versió 4.0.0 i l'eina *Rmarkdown*.

1.6. BREU SUMARI DE CONTRIBUCIONS I PRODUCTES OBTINGUTS

Al final del TFM s'obtenen les següents contribucions/productes:

- Coneixement i pràctica de com tractar dades provinents de la base de dades "*EMRbot*s".

- Comparació estadística de varies variables entre dos grups de poblacions. Taula comparativa entre aquests dos grups de poblacions (diagnosticats o no d'AR) per discernir quines variables estan associades amb el diagnòstic de la malaltia en qüestió.

- Models/Algoritmes de predicció optimitzats que permetin classificar la població en dos categories. Taula comparativa de l'actuació dels diferents models i algoritmes per tal d'identificar el/s més adients.

- Aplicació interactiva *Shiny* de visualització gràfica de la base de dades.

- Recopilatori d'articles i material publicat en el tema dels EMR.

El llistat de productes anteriors s'inclou la memòria, estructurada en forma de capítols.

Per altra banda, la resta de productes s'inclouen com a Annexos a la memòria:

- La base de dades *EMRBots* filtrada i preparada, en format Excel.
- El codi *Rmarkdown* utilitzat per l'anàlisi.
- Un informe reproduïble en PDF.
- El codi R que ha permès el desenvolupament de l'aplicació *Shiny*.
- Transparències amb els resultats.

1.7. BREU DESCRIPCIÓ DELS ALTRES CAPÍTOLS DE LA MEMÒRIA

La memòria es divideix en els següents capítols:

Capítol 1. Introducció

Descripció i explicació de la utilitat dels registres electrònics de salut, avantatges, limitacions, factors a tenir en compte a l'hora de realitzar un estudi d'aquest calibre. Inclou una justificació del tema escollit. Finalment, s'explica quins mètodes estadístics s'han aplicat, el seu propòsit i l'ordre en que s'han utilitzat.

Capítol 2. Identificació factors de risc.

Primera part del cas pràctic, on s'adequa la base de dades escollida i es realitza estadística descriptiva. Per altre, es realitza un tractament estadístic de les dades per trobar associacions entre variables a partir del software *Rstudio* versió 4.0.0.

Capítol 3. Comparació d'algoritmes i models de predicció.

Segona part del cas pràctic, on s'aprofundeix sobre els models i algoritmes de predicció de malaltia, s'executen i es realitzen comparacions entre ells.

Capítol 4. Discussió i conclusions

Discussió i reflexió del treball. Es tracten diferents aspectes: des de la planificació de tasques i compliment d'objectius, fins els resultats obtinguts.

Capítol 5. Glossari

Definició dels termes i acrònims més utilitzats al llarg de la memòria.

Capítol 6. Bibliografia

Llistat de referencies, articles i fonts que s'han consultat i fet referència al llarg del desenvolupament del treball. Es presenten alfabèticament ordenats.

Capítol 7. Annexos:

- Annex 1: Excel amb la base de dades *EMRbot* preparada per l'anàlisi.
- Annex 2: Codi R utilitzat per generar l'informe reproduïble en format *Rmarkdown*.
- Annex 3: Informe estadístic reproduïble en *PDF.*
- Annex 4: Codi R utilitzat per la generació de l'aplicació *Shiny*.

2. IDENTIFICACIÓ DE FACTORS DE RISC

2.1. TRACTAMENT I ADEQUACIÓ INICIAL DE LES DADES PROVINENTS D'EMR

2.1.1. Eines del software R per processar i extreure informació dels EMR.

La estructura i els components dels sistemes EMR solen ser complicats i necessitar de prèvia familiarització. Això comporta que, en la major part dels estudis, els investigadors, clínics i científics, necessitin realitzar col·laboracions amb diferents grups d'experts per tal d'obtenir assessorament [5].

El fet de disposar d'aplicacions, eines web i paquets de software adients poden ajudar a la visualització i anàlisi d'aquestes dades, igualment que faciliten la iniciació dels investigadors.

A continuació, es presenta un resum (Taula nº1) de diferents paquets d'*R* disponibles dissenyats per facilitar l'anàlisi de dades provinents d'EMR.

Paquet R	Funcionalitats	Funcions útils per cas pràctic
EHR	-Processar i analitzar dades EMR per realitzar estudis relacionats amb tractaments. -Realitzar anàlisi farmacocinètica i farmacodinàmica utilitzant EMR. -Realitzar estudis PheWAS (estudis d'associacions entre exposició a tractaments i fenotips).	>dataTransformation(): realitzar petites modificacions a la base de dades. >zeroOneTable(): Realitzar taules de contingència amb variables binaries.
rEHR [13]	-Accelerar i automatitzar l'extracció i l'anàlisi de dades EMR.	No funciona per versions actuals d'R. Només funciona en la versió 3.3.2.
cleanEHR	-Plataforma per netejar i processar dades d'EMR.	Es troba eliminat del repositori CRAN, pel que no funciona amb les versions d'R actuals.
ROMOP [5]	-Paquet d'R per interactuar més fàcilment amb dades EMR de format OMOP.	No aplica ja que les dades d'*EMRbot*s no es troben en format OMOP.
Rdrugtrajectory [12]	-Dissenyat per l'anàlisi de dades provinents de la CPRD (Clinical Practice Research Datalink dataset), una EMR de Regne Unit.	No aplica ja que les dades d'*EMRbot*s no provenen de la CPRD.
EHRtemporalVariability	-Traçar canvis temporals en EMR, sobretot en bases de dades de llarga duració.	No aplica pel tipus d'anàlisi que es vol realitzar.

ComoRbidity [6]	-Paquet d'*R* per ajudar en l'anàlisi de comorbiditats utilitzant dades obtingudes de EMR.	>summaryDB(): analitzar i caracteritzar la població d'estudi. >populationAge(): analitzar la distribució d'edat dels pacients diagnosticats amb la malaltia d'estudi en comparació a la resta. >diseasePrevalence(): calcular la prevalença de la malaltia. >diagnosticUse(): estudiar un desordre que engloba varis codis de diagnòstic >comorbidityAnalysis(): realitzar un anàlisi de comorbiditats i guardar-lo en un objecte "class". >network() + heatmapPlot(): visualitzar gràficament els resultats de l'anàlisi de comorbiditat. >sex*Ratio*(): visualitzar la distribució de la comorbiditat per sexe.

Taula nº1. Paquets d'R disponibles per l'anàlisi de dades EMR.

2.1.2. EMRbots

Les EMR reals contenen informació molt confidencial i sensible, ja que en elles es descriu informació personal de pacients. Per això mateix, el seu accés és limitat a grups de recerca molt específics, que normalment es troben associats als hospitals o sistemes de salut.

Per realitzar aquest projecte, s'ha utilitzat una EMR generada artificialment que conté dades virtuals. Es tracta de la base de dades *EMRbot*s. Aquesta base de dades engloba tres cohorts diferents que contenen informació de 100, 10.000 i 100.000 pacients respectivament, que es poden descarregar directament des de la pàgina web: www.*EMRbots*.org.

Aquestes dades són molt útils per familiaritzar-se amb dades EMR o per desenvolupar nous algoritmes de *Machine Learning,* però realment no serveixen per extreure conclusions extrapolables a pacients reals. Durant el seu procés de creació no es van tenir en compte les complexes interaccions factors-pacients, ja que hi havia massa factors, associacions i assumpcions a tenir present, el que suposava un enorme repte.

2.1.3. Configuració inicial base de dades

La cohort escollida per aquest treball és la que conté informació de 100.000 pacients. Per la seva descarrega, es troba comprimida en *.rar* i conformada per 4 arxius *.txt* (Figura nº2).

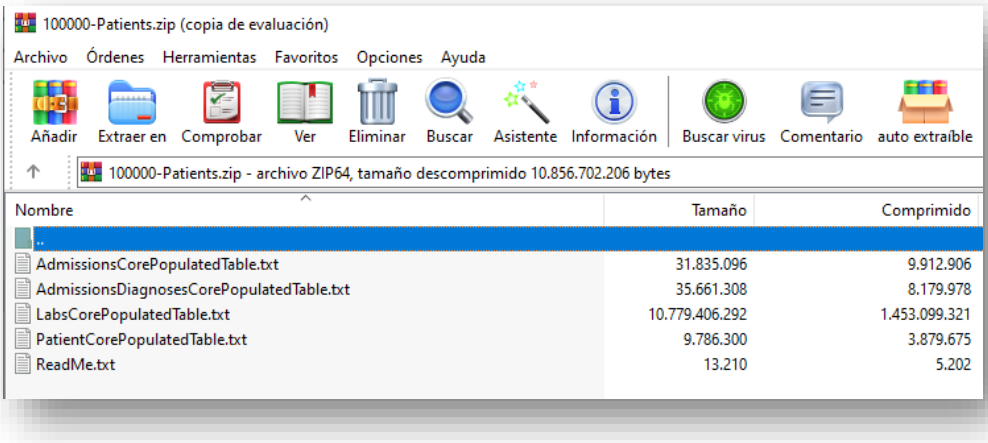


Figura nº2. Fitxers que conformen l'*EMRbot* de 100.000 pacients.

Aquests arxius contenen la següent informació:

- *PatientCorePopulatedTable.txt* : Informació dels 100.000 pacients virtuals. Inclou les següents variables:
  - Gènere.
  - Data naixement (per poder relacionar-ho amb edat).
  - Raça.
  - Estat civil.
  - Llengua.
  - Grau de pobresa.
- *AdmissionsCorePopulatedTable.txt*: Informació de les diferents admissions associades als pacients anterior. Cada pacient virtual està associat amb 1-10 admissions. Per tant, hi ha 361.761 admissions.

- *AdmissionsDiagnosesCorePopulatedTable.txt*: Informació dels diagnòstic associats a cada admissió i pacient. Per tant, 361.761 diagnòstics.

- *LabsCorePopulatedTable.txt*: Informació de diferents paràmetres de laboratori associats a cada pacient i admissió. En total, més de 30 paràmetres, inclosos en: anàlisi d'orina, paràmetres metabòlics i paràmetres sanguinis. En resum, unes 107.535.387 mesures de laboratori.

Un cop descarregats els diferents arxius *.txt*, es procedeix a seleccionar els pacients i les variables adients pel posterior anàlisi. Els passos seguits es descriuen a continuació:

1) Importació de l'arxiu *AdmissionsDiagnosesCorePopulatedTable.txt* a Excel. Identificació del codi de diagnòstic associat a Artritis reumatoide → **M05.** En total hi ha 19.112 diagnòstics d'AR, tant com a única condició com incloent complicacions associades.

2) Filtratge a Excel segons admissió → El 75 % (14.283/19.112)dels diagnòstics d'AR es troba entre les primeres tres admissions. Es decideix incloure només pacients diagnosticats d'AR en les tres primeres admissions:
   - 5.368 pacients (37,6 %) diagnosticats en la primera admissió.
   - 4.943 pacients (34,6 %) diagnosticats en la segona admissió.
   - 3.971 pacients (27,8 %) diagnosticats en la tercera admissió.

3) En el mateix arxiu i en l'arxiu *AdmissionsCorePopulatedTable.txt* , es filtra segons admissió (AdmissionID), trobant **100.000 pacients de primera admissió**, que es corresponen amb els mateixos pacients de l'arxiu *PatientCorePopulatedTable.txt.* Així doncs, s'ordenen alfabèticament els tres arxius segons PacientID (*AdmissionsDiagnosesCorePopulatedTable.txt + AdmissionsCorePopulatedTable.txt + PatientCorePopulatedTable.txt.) i* s'agrupa tota la informació dels tres arxius en una sola fulla d'Excel.

4) A continuació, es seleccionen els 5.368 pacients diagnosticats d'AR. Per altra banda, es seleccionen la resta de pacients no diagnosticats d'AR a la primera admissió (n= 94.631). Utilitzant la funció **=ALEATORIO()** d'Excel, es creen números aleatoris per cadascun d'aquests pacients i s'ordenen de menys a més. S' escullen els primers 5.368 pacients. D'aquesta manera, es seleccionen aleatòriament el mateix nombre de pacients no diagnosticats d'AR que diagnosticats d'AR a la primera admissió.
   Per tant, en total s'escullen **10.736 pacients de 1a admissió.**

5) Es repeteix el mateix procediment que l'explicat en els dos punts anteriors **pels pacients de segona i tercera admissió.** La diferència és que l'Excel resultant només conté informació de 2 arxius: *AdmissionsDiagnosesCorePopulatedTable.txt + AdmissionsCorePopulatedTable.txt,* ja que el número de pacients de 2a i 3a admissió és inferior a 100.000 i no es poden agrupar tant fàcilment amb la informació de *PatientCorePopulatedTable.txt.* Així doncs, s'obté:

   - Segona admissió: 4.943 pacients diagnosticats d'AR + 4.952 pacients no diagnosticats d'AR seleccionats de forma aleatòria **= 9.895 pacients de 2a admissió.**
   - Tercera admissió: 3.971 pacients diagnosticats d'AR + 3.970 pacients no diagnosticats d'AR seleccionats de forma aleatòria = **7.941 pacients de 3a admissió.**

   Tal i com s'ha explicat, la selecció resultant no conté la informació de les variables relatives als pacients.

6) Per incloure aquestes variables, s'utilitza la funció **=BUSCARV().** La variable comuna entre els diferents arxius és *PatientID*. D'aquesta manera, es transformen les diferents fulles d'Excel en taules i se li demana que si coincideix el Pacient ID, s'importi a la mateixa fulla la informació relativa a gènere, data naixement, raça, estat civil, llengua i % per sota la pobresa.

En aquest punt, s'obté una primera base de dades amb informació de 28.572 pacients virtuals, la meitat diagnosticats amb AR i la resta amb altres diagnòstics (Figura nº 3) .

PatientID	AdmissionID	AdmissionStartDate	AdmissionEndDate	Primary DiagnosisCo	Primary Diagnosis Description	Patient Gende	Patient DateOfBirth	Patient Race	Patient MaritalStat	Patient Languag	PatientPopulation PercentageBelow Poverty
0000585C-5C9D-49BD-8F9E-41345464F832	1	2005-08-23 18:45:59.687	2005-09-02 23:18:49.750	O03.33	Metabolic disorder	Female	1984-04-28 23:40:27.2	White	Single	English	17.31
00005E17-D776-4260-8566-3BF18422A777	1	1952-03-10 03:34:16.947	1952-03-17 00:50:44.730	C05.0	Malignant neoplasr	Female	1924-08-02 09:42:39.8	White	Unknown	English	17.6
0002DD3B-F8AA-4B4B-B5B4-33B933B18FEC	1	1958-02-26 20:32:56.993	1958-03-11 11:32:29.733	H11.41	Vascular abnormali	Male	1937-12-18 10:54:52.8	White	Single	Unknown	14.85
00039D6B-6008-46B9-BCB3-1FB108E739A3	2	1995-12-25 17:00:59.673	1995-12-28 12:13:55.447	O99.713	Diseases of the skir	Male	1973-01-25 15:25:13.7	White	Divorced	Spanish	18.98
0006542E-91B9-4A24-9A6E-05B18176085E	1	1967-05-16 11:06:26.923	1967-05-26 00:25:51.593	B40.0	Acute pulmonary bla	Female	1947-03-15 10:43:05.6	White	Separated	English	19.55
00124F16-6D6D-433E-A8CE-7096B6755607	1	2005-01-05 11:35:49.823	2005-01-12 08:42:19.623	D13.9	Benign neoplasm of	Female	1982-10-29 01:09:07.8	Asian	Single	English	19.14
0012A538-0485-4A96-BFE5-F196179468C0	2	1997-09-03 21:36:04.943	1997-09-10 16:58:14.097	Z95.0	Presence of cardiac	Female	1947-10-02 06:05:45.7	African Ameri	Married	English	15.1
0012E2AC-9CB2-4977-B086-2D50F0A2ED07	3	2005-12-10 10:20:05.990	2005-12-22 15:23:20.520	M05.461	Rheumatoid myopat	Male	1941-03-04 07:29:46.3	White	Divorced	Unknown	15.61
001376F1-3CE1-4E1A-A8EF-82306C0F46E1	2	1995-08-25 20:39:02.113	1995-09-08 09:46:50.810	M63.841	Disorders of muscle	Female	1966-07-21 04:40:35.1	African Ameri	Married	Spanish	19.21
0013E76D-E5DC-4878-9490-02F47168A3FD	3	2011-02-26 09:07:47.570	2011-03-09 21:57:16.413	M05.4	Rheumatoid myopat	Female	1986-09-29 23:52:12.1	White	Single	English	11.96
0015DC00-E55A-4D69-9ECF-C971D617C595	1	1991-12-19 12:35:27.330	1991-12-30 11:39:17.083	C94.31	Mast cell leukemia,	Female	1967-06-03 07:58:02.5	Asian	Single	Spanish	16.93
0016859E-FA7D-432A-A5FF-0A081B698E81	3	2011-03-18 18:13:07.540	2011-03-24 23:54:01.290	F11	Opioid related diso	Female	1959-07-09 15:05:20.8	Asian	Married	English	6.4
001B25D0-EA57-4493-8C81-DA85305EE15F	2	1997-09-12 03:56:19.223	1997-09-25 06:12:47.053	E08.359	Diabetes mellitus d	Female	1952-01-06 01:20:51.8	African Ameri	Single	English	15.66
001C8126-5353-4FC5-9C43-632E513F1474	1	1993-08-17 21:01:37.630	1993-08-23 20:09:53.680	M05.45	Rheumatoid myopat	Female	1969-06-15 20:47:42.3	Unknown	Single	English	96.05
001C8126-5353-4FC5-9C43-632E513F1474	2	1998-10-12 05:29:43.173	1998-10-21 01:43:31.553	E75.02	Tay-Sachs disease	Female	1969-06-15 20:47:42.3	Unknown	Single	English	96.05
001D4569-0086-43FE-AED7-9C204A082A81	1	1958-12-16 22:32:08.113	1958-12-21 03:51:46.400	I21.02	ST elevation (STEMI)	Female	1940-10-04 21:15:46.2	White	Single	English	15.22
00216F7C-1B6B-4636-A708-91DDD1BFD7B2	2	1999-06-29 09:05:32.983	1999-07-13 04:27:00.620	M05.321	Rheumatoid heart d	Male	1963-08-24 01:50:34.5	African Ameri	Married	English	15.28
0021E779-4C77-4999-8E02-FC7A6EE1D436	2	2005-03-19 04:05:00.603	2005-03-29 16:09:32.720	M05.441	Rheumatoid myopat	Female	1973-03-07 17:14:22.9	African Ameri	Separated	English	16.15
002373E1-06CE-4DBF-AA5C-CECE44A56A94	1	2009-03-19 00:07:58.713	2009-03-31 09:34:34.893	M05.341	Rheumatoid heart d	Female	1987-12-15 22:15:55.1	White	Divorced	English	4.2
002639CD-92D5-4AB2-A73D-4DE3C13E46A8	2	1996-08-04 17:05:11.210	1996-08-12 21:36:05.790	O99.62	Diseases of the dig	Female	1961-02-10 23:24:12.4	White	Single	Spanish	0.8
002B0E4B-621E-4D4D-9869-200891A4EFDE	1	1961-01-22 12:31:24.890	1961-02-09 10:38:35.110	M05.312	Rheumatoid heart d	Female	1934-07-02 03:21:36.3	African Ameri	Divorced	English	18.07
002B0E4B-621E-4D4D-9869-200891A4EFDE	3	2008-11-26 21:31:28.437	2008-12-14 07:12:34.080	M05.322	Rheumatoid heart d	Female	1934-07-02 03:21:36.3	African Ameri	Divorced	English	18.07
002BC9F1-8C5D-4087-AF7E-0FAF8647A26E	1	1948-05-18 21:30:38.177	1948-05-24 00:29:47.523	E11.42	Type 2 diabetes mel	Male	1923-02-09 15:49:10.2	White	Married	English	13.97
002C5D2D-75CB-4160-9B6D-8E9613F7B4BF	1	1949-12-13 12:49:04.133	1949-12-27 22:10:51.913	C91.52	Adult T-cell lymphor	Female	1923-11-16 16:44:57.4	Asian	Single	English	13.39
002F3761-44FC-4B54-A189-3BD02A7FC703	2	1998-12-17 12:05:58.313	1998-12-20 02:52:19.810	M05.142	Rheumatoid lung di	Male	1972-05-05 21:56:05.9	Unknown	Single	English	18.96

Recuento: 28572

Figura nº 3. Primer esborrany de la base de dades post- selecció de pacients.

El següent pas correspon a escollir quines variables de laboratori es volen incloure. Tal i com s'ha explicat anteriorment, la base de dades original conté 31 variables de laboratori. Aquestes 31 variables de laboratori es troben agrupades de la següent forma:

- Paràmetres sanguinis: 12 variables.
- Paràmetres metabòlics: 15 variables.
- Anàlisi d'orina: 4 variables.

Es considera que amb 20 d'aquests paràmetres n'hi ha suficient. Els paràmetres seleccionats són:

- 6 paràmetres sanguinis: Neutròfils absoluts, Limfòcits absoluts, Hematòcrit, Hemoglobina, Plaquetes i Volum Corpuscular mig.
- 13 paràmetres metabòlics: Albúmina, ALK, ALT, AST, Bilirubina, BUN, Calci, Clor, Creatinina, Glucosa, Potassi, Sodi i Proteïnes totals.
- 1 paràmetre d'Anàlisi d'orina: pH

L'arxiu *LabsCorePopulatedTable.txt* té una mida de més de 10GB, pel que no és possible obrir l'arxiu complet amb Excel. Per resoldre aquest problema s'utilitza una eina incorporada a Excel, anomenada **Power Query.** Aquest és un complement gratuït que permet extreure dades de diferents fonts, transformar-les i carregar-les pel seu ús posterior. Permet treballar amb dades de mida molt gran.

Els passos realitzats amb *Power Query* per obtenir la informació dels paràmetres de laboratori seleccionats es resumeixen a continuació:

1) S'obre un arxiu d'Excel. Pestanya *Datos* → *Obtener datos desde el texto.* Es selecciona l'arxiu *LabsCorePopulatedTable.txt,* el que genera que s'obri una finestra de previsualització de les dades.

2) Si es clica *Transformar datos* s'obre l'editor de *Power Query* que permet realitzar una sèrie de transformacions a les dades. A partir d'aquí, es filtren les dades segons *AdmissionID* (1-2-3) i tot seguit, es fa un segon filtratge segons paràmetre de laboratori. Així doncs, per cada admissió es generen 20 arxius d'Excel diferents, cadascun amb informació relativa a un paràmetre de laboratori.

3) Per tal d'obtenir només els valors corresponents als pacients prèviament seleccionats (n=28.752), es torna a utilitzar la funció **=BUSCARV(),** a partir del *PatientID*. Es crea un nou arxiu per cada paràmetre de laboratori només amb la informació dels pacients seleccionats.

4) Com que hi ha varis valors del mateix paràmetre per cada pacient i admissió, es realitza una mitjana i així s'obté només un valor per paràmetre/pacient/admissió. Per aconseguir-ho, s'utilitza la funció **=PROMEDIO.SI().** Aquesta, funciona de tal manera que només realitza la mitjana dels valors que coincideixen amb la *PatientID*. A continuació, s'eliminen les files duplicades, pel que finalment només es conté un valor per pacient seleccionat.

5) Aquests passos es realitzen per cada paràmetre de laboratori i admissió inclosa dins l'estudi (20 paràmetres de laboratori x 3 admissions = 60 fulles d'Excel).

6) Finalment, s'integren tots els paràmetres de laboratori amb la resta de variables. Per fer-ho, es torna a utilitzar la funció **=BUSCARV()** i es cerca la *PatientID* corresponent a cada valor. Es fa per cada admissió (3 fulles d'Excel) I **finalment** es conforma un sol arxiu que conté tots els pacients i variables seleccionats (Figura nº4 i Annex 1).

Tots els passos descrits es visualitzen de forma resumida en la Figura nº5.

PatientID	ALBUMIN (gm/dl)	ALK (U/L)	ALT (U/L)	AST (U/L)	BILI (mg/dl)	BUN (mg/dL)	CALCIUM (mg/dL)	CHLORIDE (mmol/L)	CREA (mg/dl)	GLUC (mg/dl)	POTASSIUM (mmol/L)	SODIUM (mmol/L)	TOTAL PROTEIN (gm/dL)	Absolute Lymphocytes (%)	Absolute neutrophils (%)	Hematocrit (%)	Hemoglobin (gm/dl)	Platelet count (k/cumm)	VCM (fl)	pH_Urine
0000585C-5C9	3,77	70,80	47,55	23,92	0,44	17,13	9,99	91,34	0,83	91,11	4,67	136,20	7,90	23,85	70,03	36,63	13,08	285,30	70,33	6,19
00005E17-D776	4,87	85,70	37,97	27,41	0,62	19,70	8,70	102,00	0,67	101,00	2,43	77,93	5,92	23,27	61,97	43,31	12,37	319,53	94,62	5,62
0002DD3B-F8A	3,39	80,14	45,89	20,67	0,64	17,48	10,11	103,88	0,72	111,38	3,67	125,61	8,20	22,13	65,55	37,16	12,44	255,23	71,38	4,75
00039D6B-6009	3,20	97,35	29,07	19,50	0,47	21,50	7,30	99,27	0,70	34,45	3,80	135,50	7,53	23,20	74,93	#N/D	14,90	429,15	66,57	4,03
0006542E-91B9	3,76	91,00	35,98	22,93	0,64	15,28	10,12	90,01	0,86	72,00	4,20	122,78	7,61	24,54	65,70	35,87	12,19	272,94	88,23	4,34
00124F16-6D6D	3,18	91,73	48,10	21,50	0,48	3,17	9,06	103,75	0,38	100,34	2,87	115,08	7,08	25,04	60,79	41,23	16,97	289,05	73,55	6,38
0012A538-0485	4,40	59,87	50,83	24,20	0,23	13,84	6,47	106,73	0,80	83,94	4,70	113,64	7,54	26,00	71,11	49,17	13,68	125,15	71,62	5,72
0012E2AC-9CB	3,11	73,04	46,55	29,52	0,59	22,26	9,84	96,35	0,88	81,30	3,50	137,30	6,73	25,40	61,49	36,73	13,14	230,44	79,42	6,01
001376F1-3CE1	3,78	90,35	36,73	20,87	0,27	14,72	9,23	103,42	0,73	100,98	4,25	109,42	6,68	17,51	60,30	29,82	11,90	280,87	70,88	5,80
0013E76D-E5D(	4,26	97,51	42,23	26,35	0,58	16,62	9,27	100,46	0,70	87,34	3,85	118,28	6,33	24,04	69,36	35,03	16,86	247,50	86,28	5,08
0015DC00-E55A	4,14	98,04	35,55	19,38	0,70	12,76	7,72	93,58	0,66	83,57	4,38	137,44	6,90	26,36	69,86	32,70	14,57	296,90	72,11	6,60
0016859E-FA7D	3,69	105,73	38,99	20,43	0,48	11,40	9,10	84,66	0,67	95,51	3,94	141,64	7,07	23,60	69,55	39,85	14,50	365,47	78,80	4,65
001B25D0-EA5	4,47	88,22	54,74	24,16	0,42	19,75	8,35	86,13	0,72	83,66	4,12	121,42	7,16	21,24	71,38	42,76	13,53	337,26	77,51	5,35
001C8126-5353-	4,25	95,72	34,94	32,58	0,50	12,66	9,48	102,72	0,76	87,80	4,33	147,88	7,40	26,52	65,50	41,08	13,25	256,03	83,02	5,02
001C8126-5353-	3,37	82,47	40,38	23,73	0,43	18,60	8,48	91,50	0,87	109,79	2,97	137,50	6,80	22,54	69,99	45,25	10,79	278,53	71,41	6,14
001D4569-0086-	5,18	90,62	53,10	26,18	0,62	21,90	5,94	82,45	0,88	73,40	5,06	145,70	7,66	24,83	67,48	36,74	13,73	202,98	49,38	6,72
00216F7C-1B6E	3,59	98,71	49,69	27,65	0,54	13,71	9,50	87,48	0,81	93,58	3,00	132,33	7,03	23,48	64,65	40,77	11,39	233,13	84,99	5,23
0021E779-4C77	3,12	78,43	35,55	18,68	0,44	16,85	8,76	102,66	0,50	78,91	4,61	100,96	5,59	15,88	63,74	35,00	9,54	263,95	76,61	5,84
002373E1-06CE	4,62	118,56	38,57	25,92	0,43	13,71	8,76	98,02	0,86	103,58	4,72	130,58	7,23	20,63	52,81	29,69	13,92	276,10	67,88	5,49
002639CD-92D	3,40	100,07	39,24	25,29	0,73	20,97	9,25	100,03	0,61	92,08	4,51	137,60	7,31	19,30	73,33	33,14	12,00	343,24	87,29	5,12
002B0E4B-621E	3,88	90,90	38,13	25,99	0,57	21,08	8,75	96,16	0,61	84,26	3,95	120,60	7,50	19,59	68,58	42,82	12,70	257,79	78,42	4,95
002B0E4B-621E	4,11	76,28	50,44	18,00	0,55	14,76	9,51	83,15	0,83	73,93	4,01	139,70	7,95	26,62	51,91	36,33	14,49	316,48	83,72	4,72
002BC9F1-8C5I	3,78	71,75	49,15	32,05	0,10	16,78	9,70	54,83	0,70	55,48	5,35	67,92	5,88	19,83	57,32	43,53	9,42	279,10	72,35	5,70
002C5D2D-75C	3,58	85,20	29,34	26,26	0,28	13,75	8,08	102,76	0,65	82,41	3,23	141,42	7,11	19,85	50,91	33,34	12,69	359,35	67,15	4,77
002F3761-44FC	3,23	116,20	41,00	34,13	0,73	22,27	10,87	95,45	0,87	61,93	#N/D	147,27	4,70	17,60	71,90	39,33	12,30	436,70	55,80	6,65
00317A7D-F7A	3,73	86,83	34,68	21,55	0,53	19,58	9,49	89,02	0,67	74,68	4,10	132,73	7,20	20,64	58,87	35,74	12,91	272,41	84,86	6,18
0032756C-FF82	3,39	93,45	47,15	24,58	0,41	17,27	9,91	102,30	0,65	84,19	3,84	128,60	6,05	21,95	68,73	36,75	13,71	269,04	76,56	5,29
00353259-2DD1	3,85	89,69	38,03	28,16	0,49	21,08	9,77	94,56	0,85	83,13	4,10	129,94	7,23	21,04	63,18	43,24	12,19	229,70	90,68	4,85
003ADAF9-BA	4,19	84,24	27,61	20,47	0,69	10,71	7,99	96,08	0,74	102,33	4,18	144,63	6,39	21,20	61,09	31,52	13,93	284,91	64,70	6,78
003BE60F-5078	4,84	91,25	40,60	20,66	0,37	15,83	8,26	73,65	0,78	75,64	4,42	143,15	8,87	27,15	75,20	47,68	7,98	247,68	52,62	5,10
003E1B33-3F6C	4,82	85,78	28,56	29,93	0,50	18,48	9,43	101,17	0,78	96,18	3,92	137,80	6,64	22,58	69,80	42,13	14,09	217,11	80,58	6,50
0043A713-2A6D	3,62	77,09	41,28	23,85	0,50	17,39	7,66	96,83	0,82	79,72	4,46	135,18	7,18	22,45	60,76	35,50	10,67	332,91	84,14	5,45
004530F2-9FC2	2,93	122,84	42,20	25,37	0,61	21,26	8,60	101,38	0,72	99,30	4,33	135,34	7,28	25,81	60,97	44,22	13,92	186,40	87,87	6,64
0046C441-9530	3,99	93,83	43,68	20,46	0,55	21,15	9,06	93,52	0,77	91,03	3,56	142,34	6,74	21,16	69,49	44,60	11,84	288,55	67,70	4,18
0046DF03-4E69	2,89	88,69	43,41	20,95	0,74	22,90	9,06	83,93	0,90	86,30	3,32	120,81	7,05	23,63	60,76	37,40	15,43	259,30	53,22	5,55

Figura nº 4. Visualització parcial de la base de dades preparada per l'anàlisi

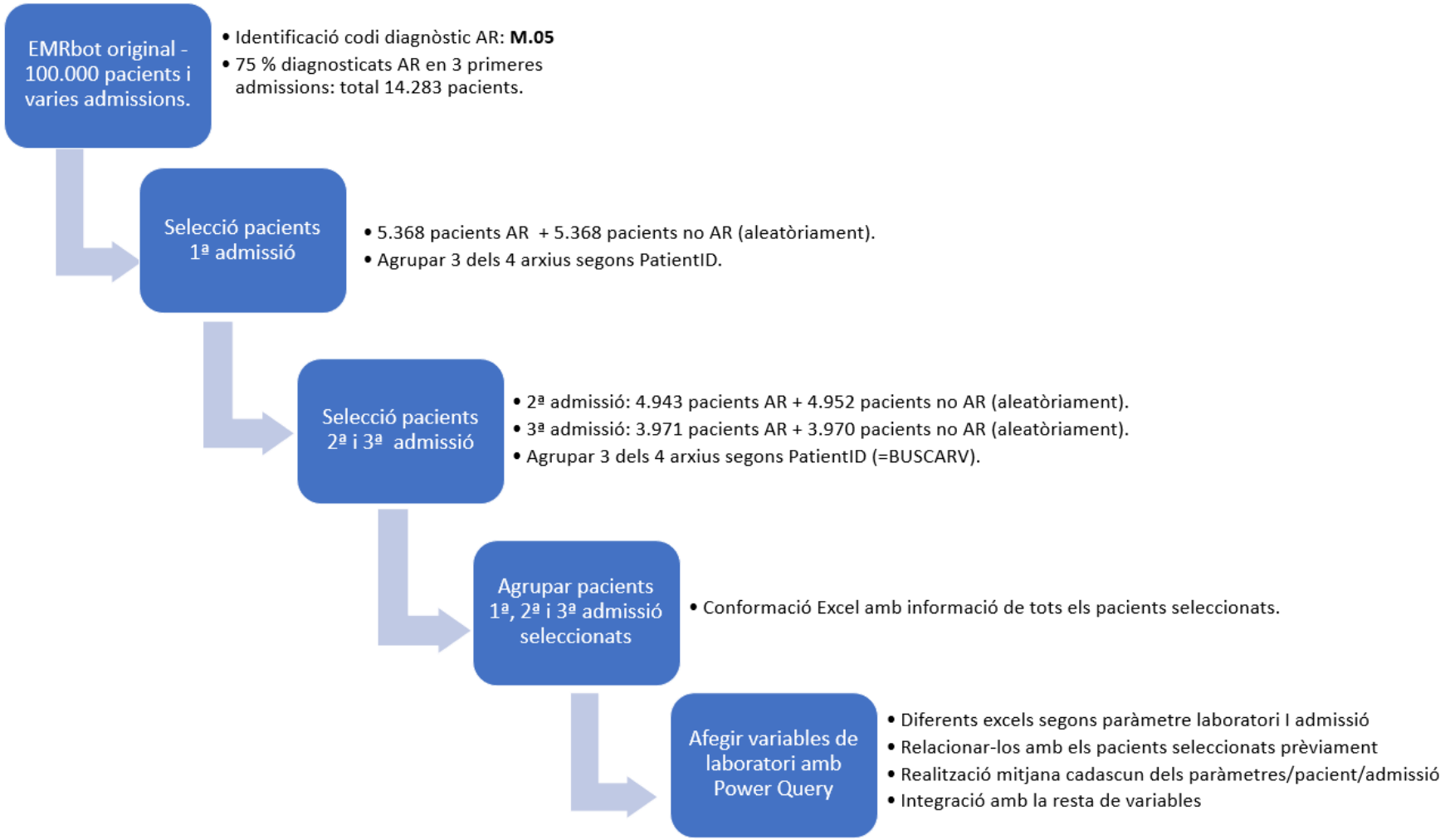


Figura nº5. Passos realitzats per la conformació de l'Excel preparat per l'anàlisi amb *Rstudio*.

2.2. ANÀLISI DESCRIPTIU DE LES DADES EMR

Un cop importat l'Excel anterior a *Rstudio*, es canvia el format d'algunes variables i se'n creen dos de noves:

- *PatientAge:* Informació sobre l'edat dels pacients. Es calcula a partir de la resta entre la data d'admissió i la data de naixement dels pacients. Expressada en anys.

- *LengthOfStay* (LOS): Informació sobre la durada de l'ingrés a l'hospital. Es calcula a partir de la resta entre la data de sortida i la data d'admissió dels pacients. Expressada en dies.

Finalment, es crea una nova variable binaria (RA) que indica si els pacients són diagnosticats d'Artritis reumatoide (AR) o d'altres diagnòstics (NO_AR) a partir del codi de diagnòstic.

Per tant, la base de dades queda finalment conformada per 28.572 observacions i 35 variables:

Variable	Descripció	Tipus
PatientID	Codi pacient	Caràcter
AdmissionID	Nº admissió (1-3)	Factor
Primary.DiagnosisCode	Codi diagnòstic	Factor
DiagnosisDescription	Descripció diagnòstic	Caràcter
Admission StartDate	Data entrada	Data
Admission EndDate	Data Sortida	Data
Patient.Gender	Gènere	Factor
Patient.Date.Of.Birth	Data naixement	Data
PatientRace	Raça pacient	Factor
PatientMarital.Status	Estat Civil	Factor
Patient.Language	Llengua	Factor
Percentage.BelowPoverty	% Pobresa	Contínua
Albumin	Valors d'albumina	Contínua
ALT	Valors ALT	Contínua
ALK	Valors ALK	Contínua
AST	Valors AST	Contínua
Bilirubin	Valors d'albumina	Contínua
BUN	Valors BUN	Contínua
Calcium	Valors Calci	Contínua
Chloride	Valors clor	Contínua
Creatinine	Valors creatinina	Contínua
Glucose	Valors glucosa	Contínua
Potassium	Valors potassi	Contínua
Sodium	Valors sodi	Contínua
Total protein	Valors proteïnes totals	Contínua
Absolute neutrophils	Neutròfils absoluts	Contínua
Absolute limphocytes	Limfòcits absoluts	Contínua
Hematocrit	Valors hematòcrit	Contínua
Hemoglobin	Valors hemoglobina	Contínua
Platelet count	Valors plaquetes	Contínua
VCM	Valors VCM	Contínua
Ph_urine	Valors pH orina	Contínua
Patient Age	Edat	Contínua
LOS	Durada ingrès	Contínua
RA	Diagnòstic Artritis	Factor

Taula nº 2. Variables que conformen la base de dades.

2.2.1. Tractament dels valors mancants o *missing values*

El primer pas consisteix en comptabilitzar i tractar els valors mancants de la base de dades. Per això, es contabilitza el nombre de NAs dins de cada variable. Les variables que presenten valors faltants es presenten a la Taula nº 3:

Variable	Valors mancants	Variable	Valors mancants
Albumin	40	Hemoglobin	45
ALT	38	VCM	34
Bilirubin	37	ALK	27
Calcium	44	AST	48
Creatinine	39	BUN	41
Potassium	40	Chloride	35
Total protein	34	Glucose	43
Absolute neutrophils	37	Sodium	44
Hematocrit	40	Absolute limphocytes	40
Ph_urine	33	Platelet count	36

Taula nº 3. Variables amb valors mancants.

S'observa que els valors mancants es presenten en les variables de laboratori. La variable AST és la que en presenta un nombre més alt (n=48).

Segons la classificació descrita a la pàgina https://datascienceplus.com/imputing-missing-data-with-r-mice-package/ [15], els valors mancants en aquest cas es considerarien MCAR (valors perduts de manera totalment aleatoria), ja que és una base de dades simulada. S'estipula que un límit d'un 5% de valors MCARs és acceptable, mentres que si es supera aquest llindar, es recomana descartar les variables afectades. Per tant, es comptabilitza quin % suposen aquests dins de cada variable (Taula nº4) i d'aquesta manera es valora si cal eliminar alguna variable.

Variable	% valors mancants	Variable	% valors mancants
Albumin	0,140	Hemoglobin	0,157
ALT	0,133	VCM	0,119
Bilirubin	0,129	ALK	0,094
Calcium	0,154	AST	0,168
Creatinine	0,136	BUN	0,143
Potassium	0,140	Chloride	0,122
Total protein	0,119	Glucose	0,150
Absolute neutrophils	0,129	Sodium	0,154
Hematocrit	0,140	Absolute limphocytes	0,140
Ph_urine	0,115	Platelet count	0,126

Taula nº 4. Percentatge que representen els valors mancants dins de cada variable.

Tal i com s'observa, el percentatge més alt correspon a un 0,168% (AST), pel que no és necessari ometre cap variable de l 'anàlisi.

A continuació es realitza un gràfic a través del paquet *VIM*(Figura nº 6) per visualitzar la frequencia i el patró dels valors mancants, és a dir, si els *missing values* de les diferents variables coincideixen en els mateixos pacients.

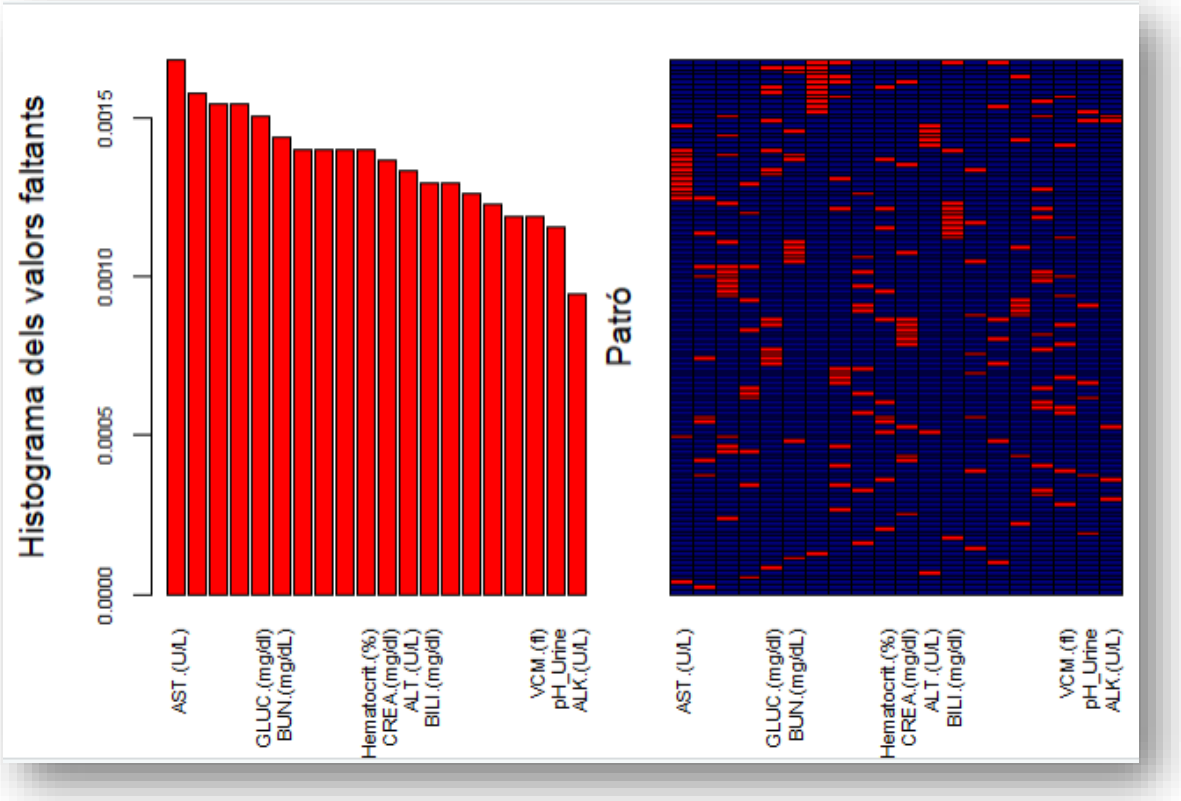


Figura nº 6. Visualització de la frequencia i els patrons dels *missing values.*

Tot i que hi ha múltiples combinacions, sembla que hi ha algunes coincidències de variables amb valors mancants pels mateixos pacients. Així doncs, se li demana a *Rstudio* que retorni el número total de pacients sense *missing values* en cap de les seves variables (casos complets). El número retornat és de 27.927 pacients, el que indica que 645 pacients presenten algun NA entre les seves variables.

Per tal de no eliminar aquests 645 pacients de la base de dades, es realitza una imputació per k-Nearest Neighbours (*kNN*) a través del paquet *DMwR*, ja que totes les variables a imputar són continues. S'utilitza un valor de k=10, ja que és el que utilitza per defecte aquest paquet. Es confirma que després de la imputació no queda cap valor mancant.

2.2.2. Transformació de variables

En aquest cas no es necessita normalitzar ni estandaritzar cap variable, ja que totes es troben expresades en unitats de mesura internacionals.

Tot i així, es pretèn modificar valors de certes variables, ja que per la creació de l'*EMRbot* s'han utilitzat algoritmes que simulen dades aleatoriament a partir dels rangs de referència de cada paràmetre, sense tenir en compte els diagnòstics associats[9]. Per això mateix, no es preveu trobar diferències amb la distribució de variables entre els dos grups de pacients (diagnosticats amb AR / no diagnosticats AR), pel que es creu difícil poder assolir els objectius del projecte si no hi ha una transformació prèvia.

Consequentment, es realitza una recerca bibliografica per intentar detectar quines de les variables incloses a la base de dades es poden trobar associades amb la malaltia d'interès (AR). Segons Fernandez et Llorente (2012)[3], els següents paràmetres es podrien donar en un curs clínic agut d'AR:

- Anèmia, tot i que amb uns valors d'hemoglobina >9.0.
- Tromobocitosis.
- Leucocitosis.
- Hipoalbuminemia.
- Elevació de fraccions de gammaglobulines.

També sembla que la malaltia sol presentar-se en una població d'entre 40 i 65 anys d'edat i té una prevalença més elevada en dones.

Per això mateix, es realitzen les següent modificacions en les variables del grup de pacients diagnòsticat amb la malaltia d'interès:

- Es divideixen els valors d'hemoglobina per 1,05 i els d'hematòcrit per 1,08.
- Es multipliquen els valors de plaquetes per 1,11.
- Es multipliquen els valors de limfòcits per 1,09 i els dels neutròfils per 1,05.
- Es divideixen els valors d'albúmina per 1,04 i es multipliquen els valors de proteïnes totals per 1,06.
- Es multiplica l'edat dels pacients diagnosticats amb AR per 1,30. També es multiplica l'edat dels no diagnosticats amb AR per 1,15.

2.2.3. Exploració gràfica de les dades

Ara sí que ja es pot procedir l' exploració gràfica de les dades. Per fer-ho, es plantegen una sèrie de preguntes :

1. Quants pacients inclosos a l'estudi són diagnosticats d'AR?
2. Quants pacients havien estat ingressats prèviament?
3. Quina és la distribució per génere dels pacients?
4. Quina diversitat de pacients hi ha?
5. Quin és l'estat civil dels pacients inclosos?
6. En quin idioma parlen?
7. En quin estat econòmic es troben els pacients inclosos?
8. Quina edat tenen els pacients al moment de l'ingrès?
9. Quina durada tenen els ingressos hospitalaris?
10. Com són les distribucions dels diferents paràmetres de laboratori analitzats?

En primer lloc, interessa respondre totes les preguntes relacionades amb els pacients inclosos, pel que es realitzen una sèrie de gràfics inicials (Figures nº7- 10).

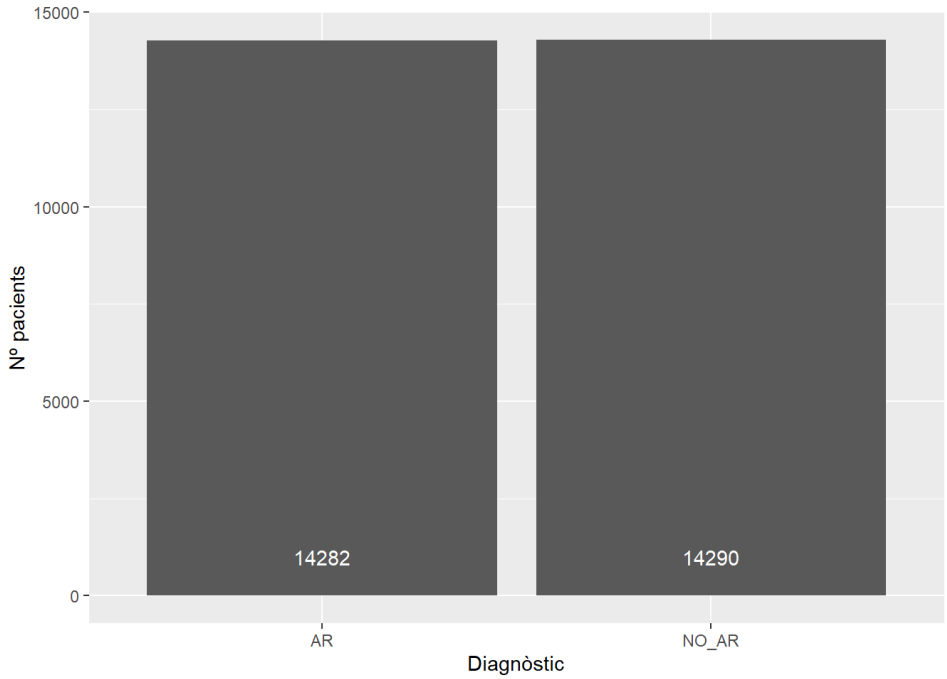


Figura nº 7. Nº de pacients segons diagnòstic.

Hi ha 14.282 pacients inclosos a l'estudi diagnosticats amb artritis reumatoide.

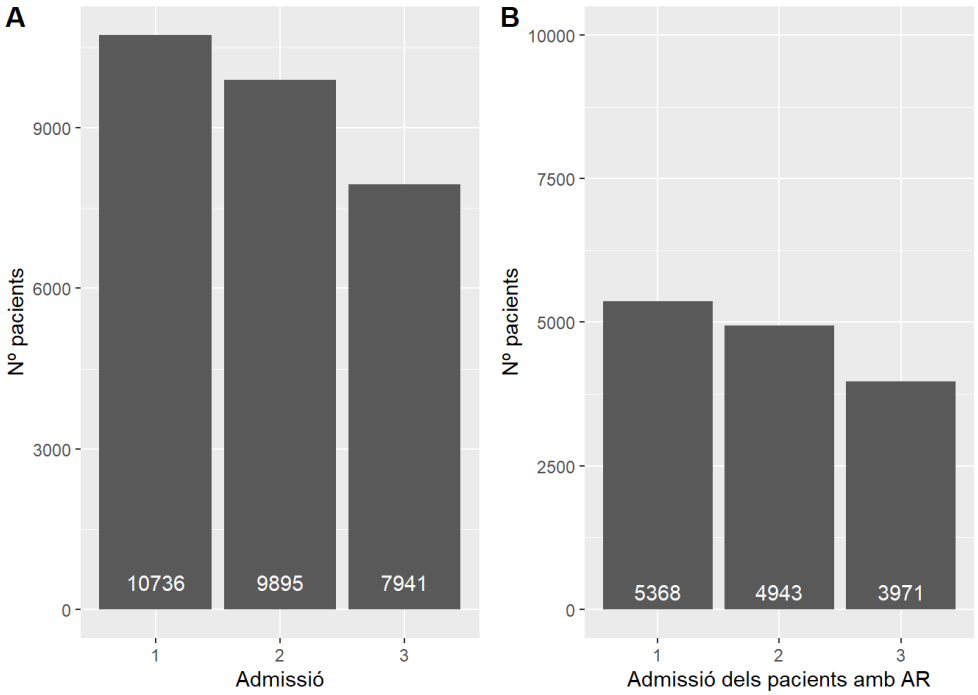


Figura nº 8. Nº de pacients totals (A) i diagnosticats amb AR (B) segons nº d'admissió.

Hi ha 62,42% dels pacients totals i diagnosticats amb AR que ja havien estat ingressats prèviament per altres patologies.

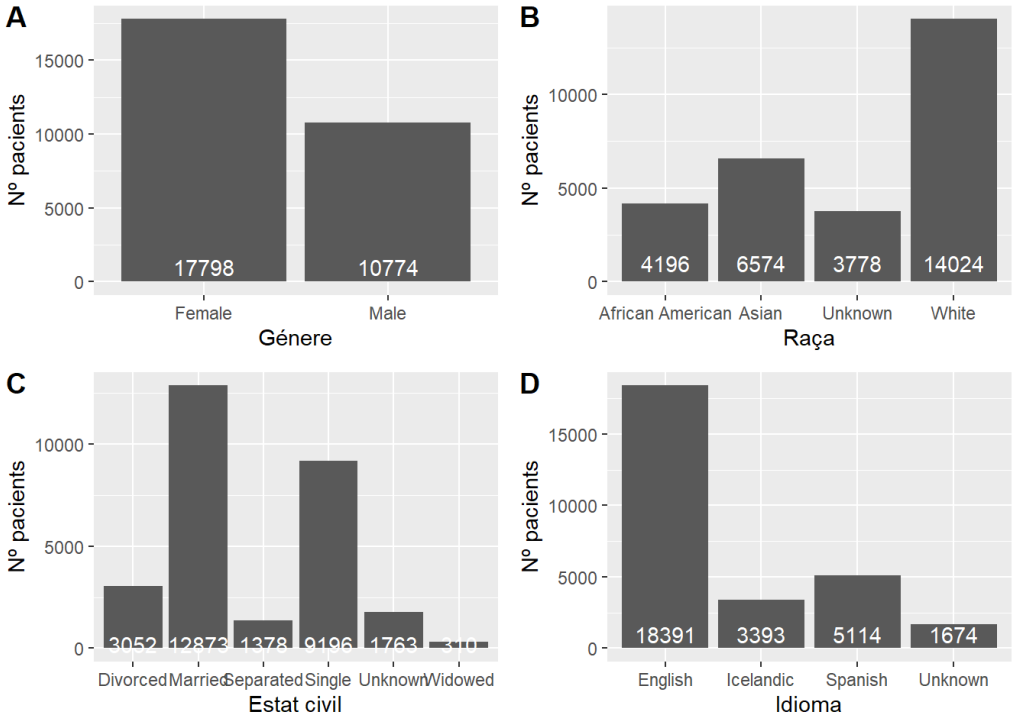


Figura nº 9. Nº de pacients segons género(A), raça(B), estat civil(C) i idioma parlat (D).

S'observa que hi ha predominància de pacients que són dones (17.798 vs 10.774), que casi la meitat dels pacients són de raça blanca i que la majoria són de parla anglesa. Respecte l'estat civil, majoritariament es troben casats o solters.

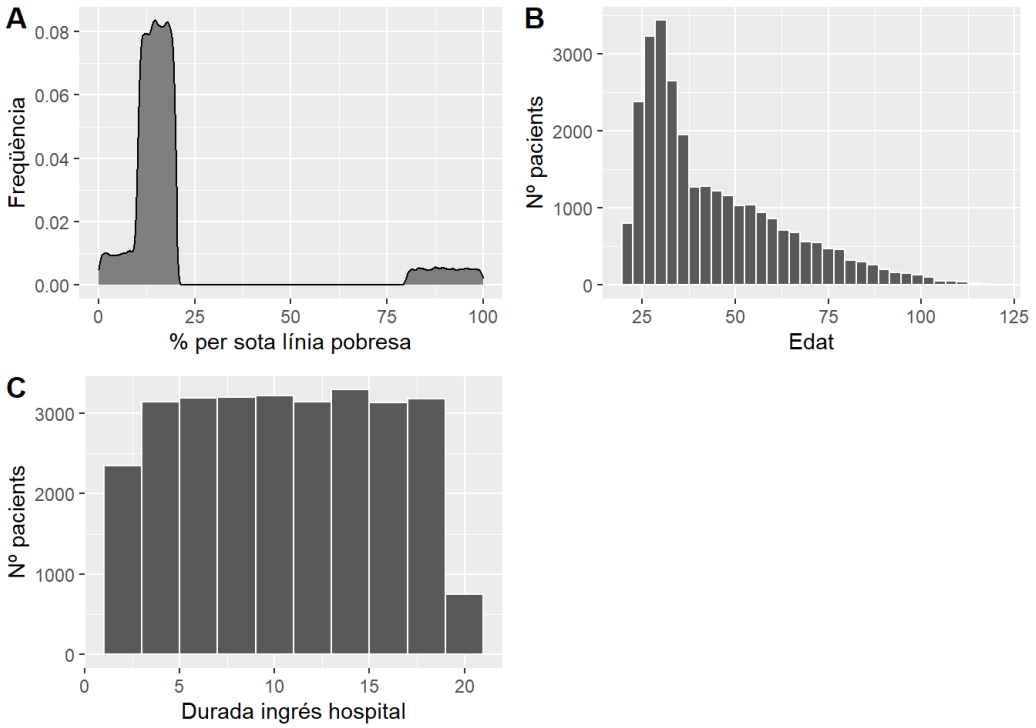


Figura nº 10. Descripció gràfica de les variables continues relatives als pacients: Estat socioeconòmic (A), edat (B) i durada d'estada a l'hospital (C).

La majoria dels pacients es troben poc per sota de la línia de pobresa, a l'admissió la majoria de pacients presenten una edat entre 25 i 40 anys i les durades dels ingressos solen ser d'entre 3 i 18 dies.

En segon lloc, es vol explorar les distribucions dels diferents paràmetres de laboratori. Per això es gràfiquen diagrames de caixes dels diferents paràmetres segons les unitats de mesura (Figures 11-18):

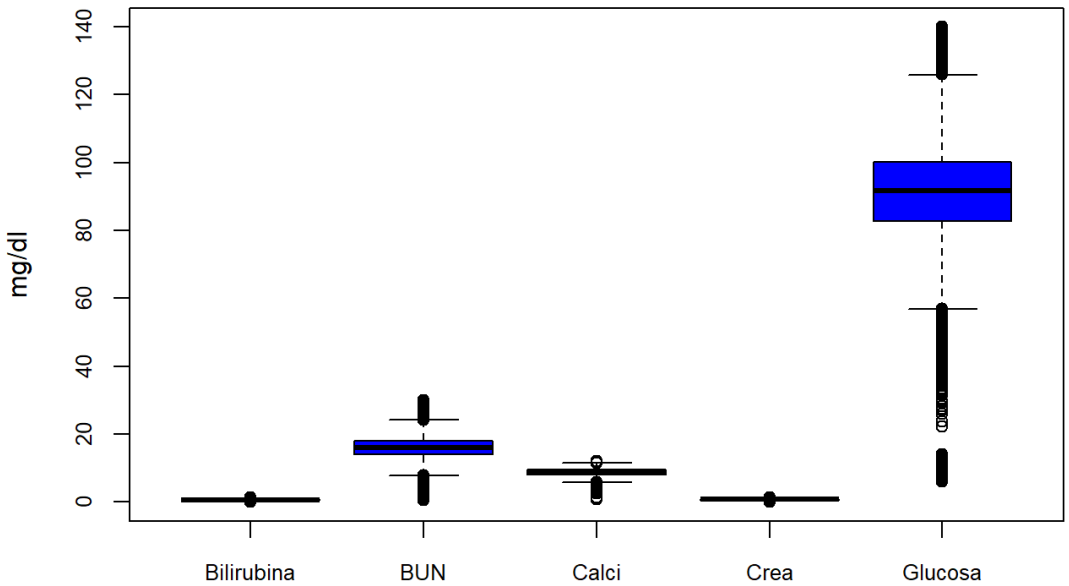


Figura nº 11. Distribució dels paramètres de laboratori mesurats amb mg/dl.

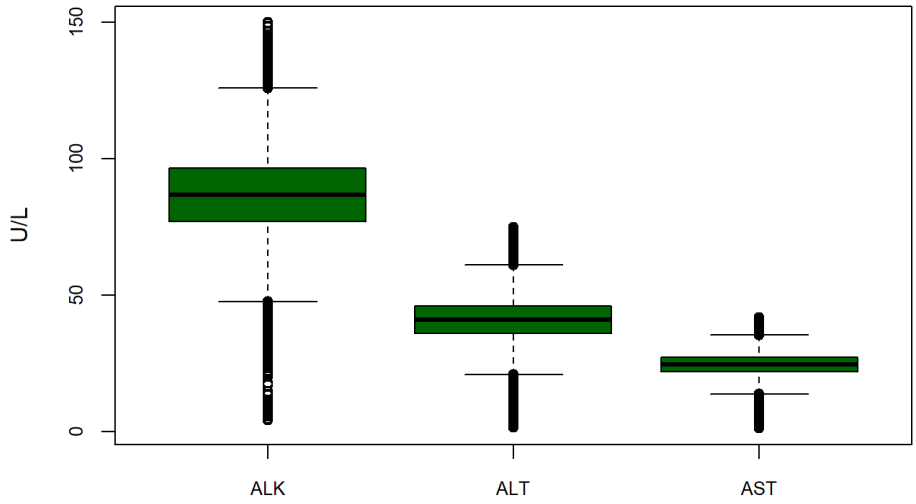


Figura nº 12. Distribució dels paramètres de laboratori mesurats amb U/L.

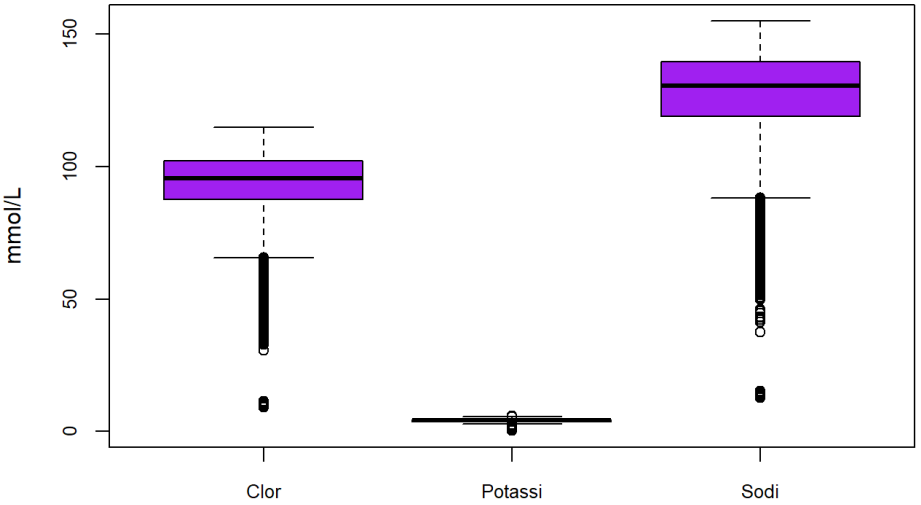


Figura nº 13. Distribució dels paramètres de laboratori mesurats amb mmol/L.

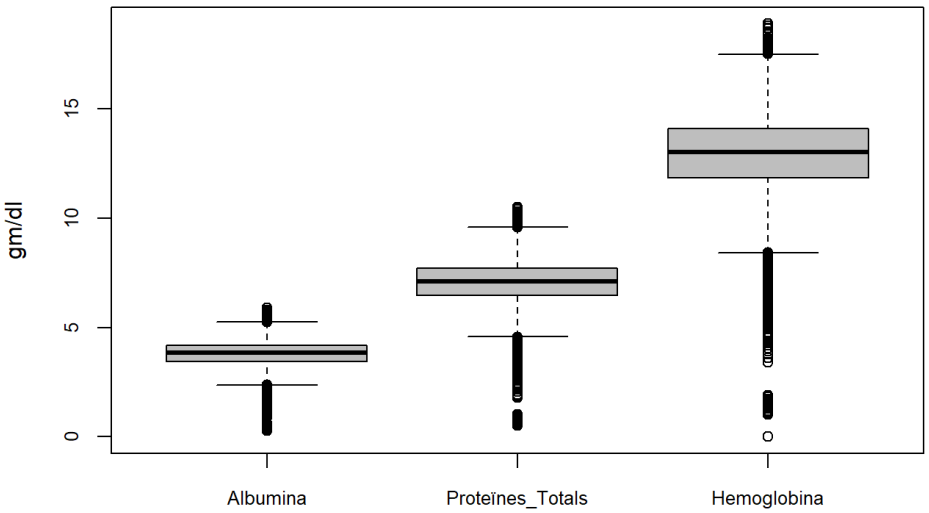


Figura nº 14. Distribució dels paramètres de laboratori mesurats amb gm/dl.

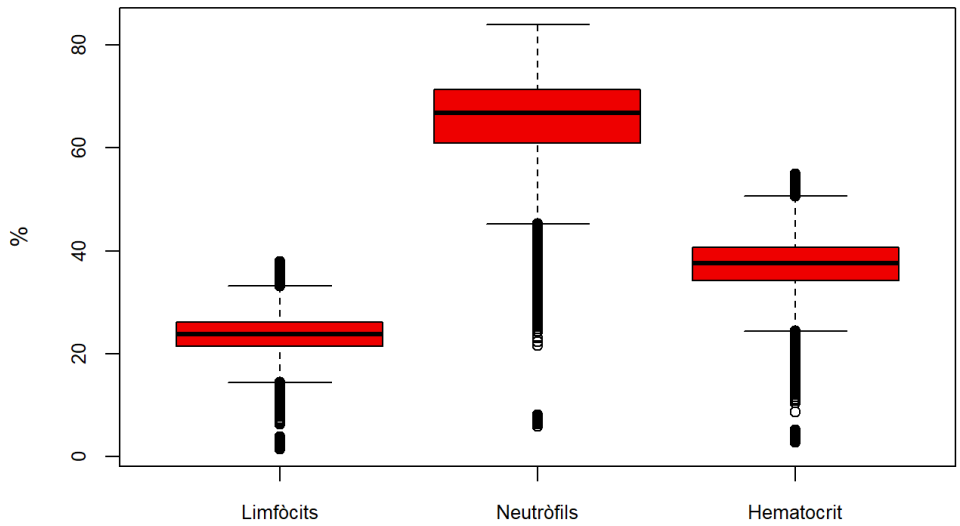


Figura nº 15. Distribució dels paràmetres de laboratori mesurats unitats de percentatge.

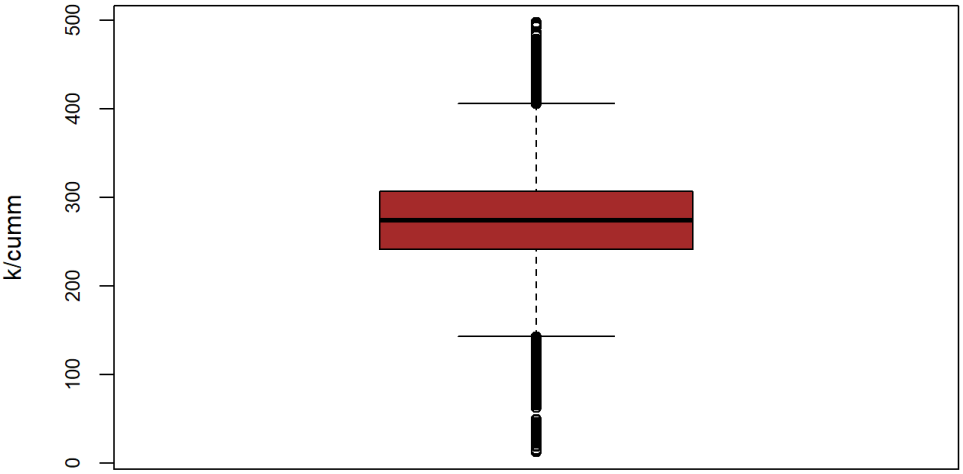


Figura nº 16. Distribució del recompte de plaquetes, expressat en k/cumm.

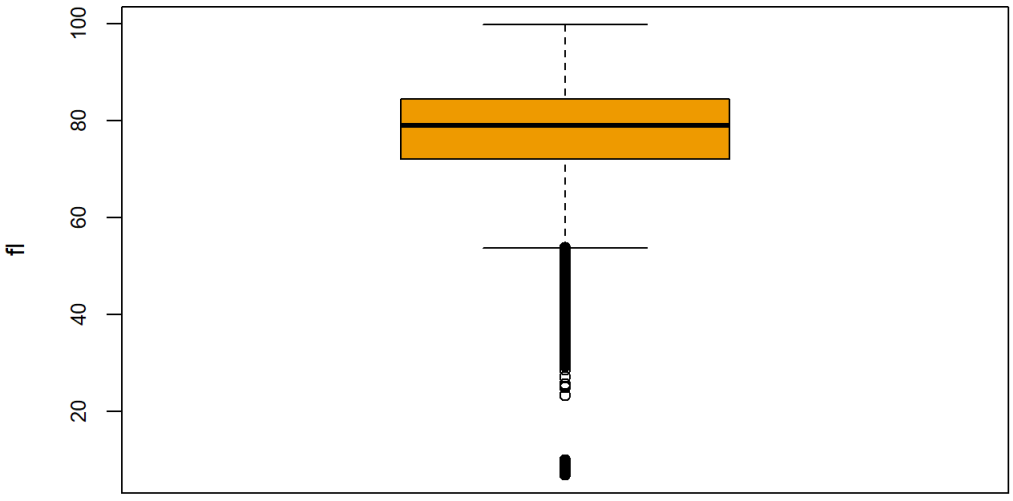


Figura nº 17. Distribució del VCM, expressat en fl.

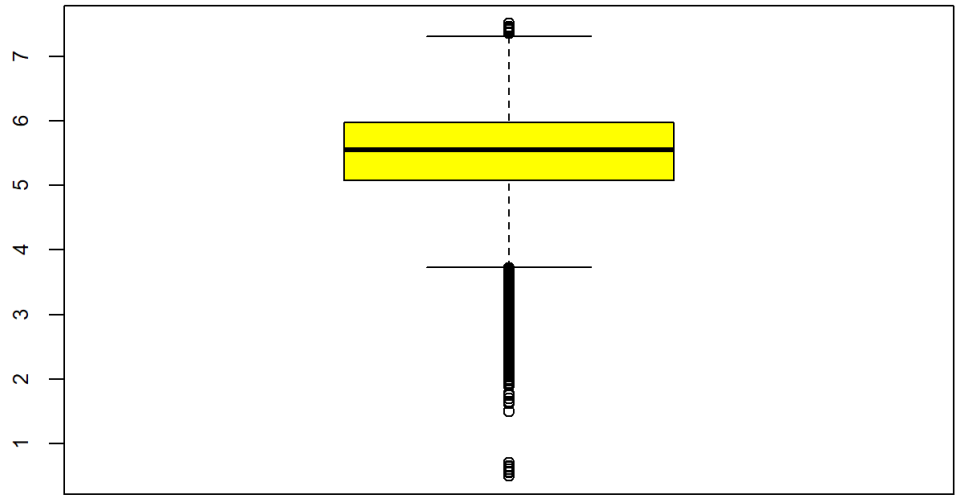


Figura nº 18. Distribució del pH de l'orina.

En general, s'observa que les distribucions es corresponen amb els rangs de referència de cada paràmetre.

Un cop s'ha obtingut informació de totes les variables per separat, s'analitza com es relacionen entre elles. Es comença per graficar la relació entre el diagnòstic d'AR, el gènere i la raça. El gràfic es realitza segons admissió (Figura nº19).

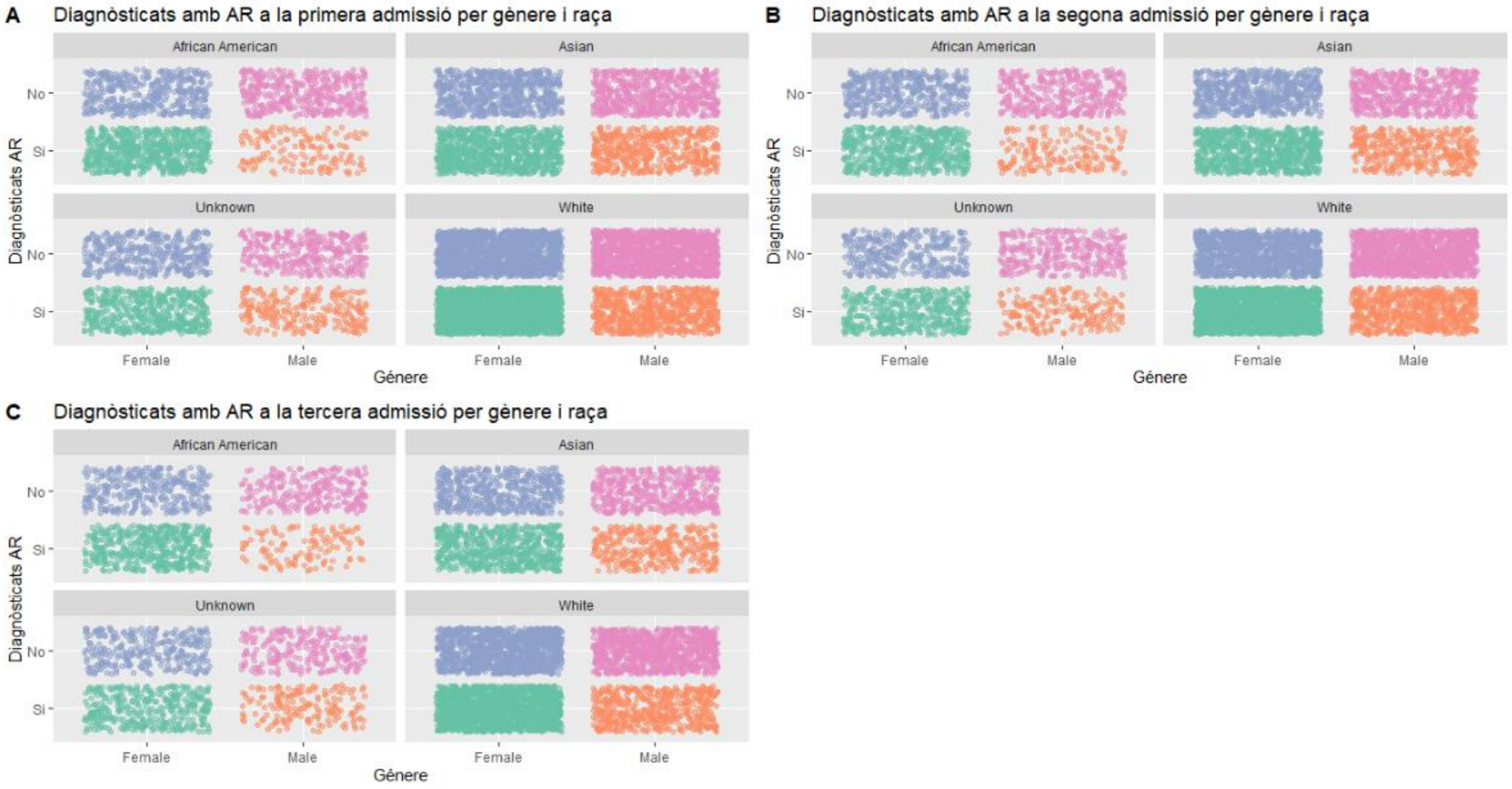


Figura nº 19. Relació diagnòstic amb gènere i raça dels pacients segons admissió.

Aparentment, sembla que hi ha més dones que homes diagnòsticades d'AR. Per altra banda, no sembla que hi hagi diferències entre raçes.

Tot seguit, s'explora la relació entre l'edat i el diagnòstic d'AR (Figura nº20).

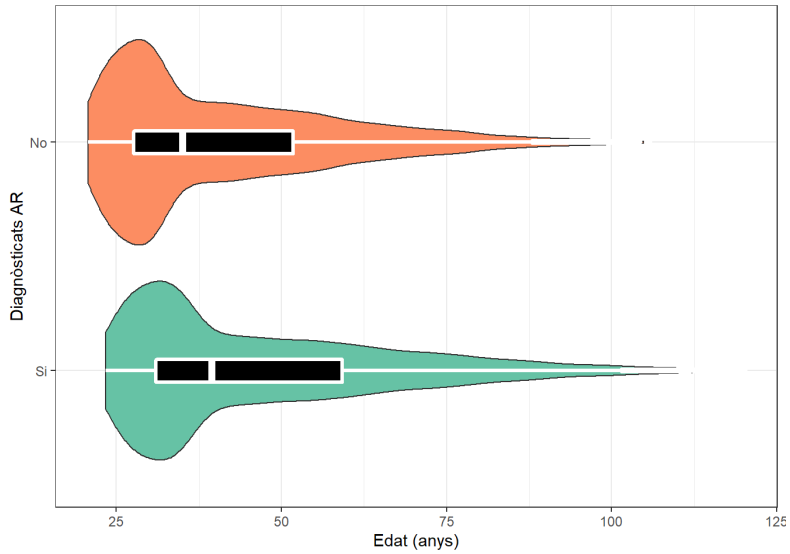


Figura nº 20. Gràfic que relaciona el diagnòstic i l'edat dels pacients.

A primera vista, sembla que l'Artritis reumatoide s'ha diagnòsticat en pacients d'edat una mica més avançada que la resta.

Finalment, s'explora la relació entre el diagnòstic i les variables de laboratori (Figures nº 21-24).

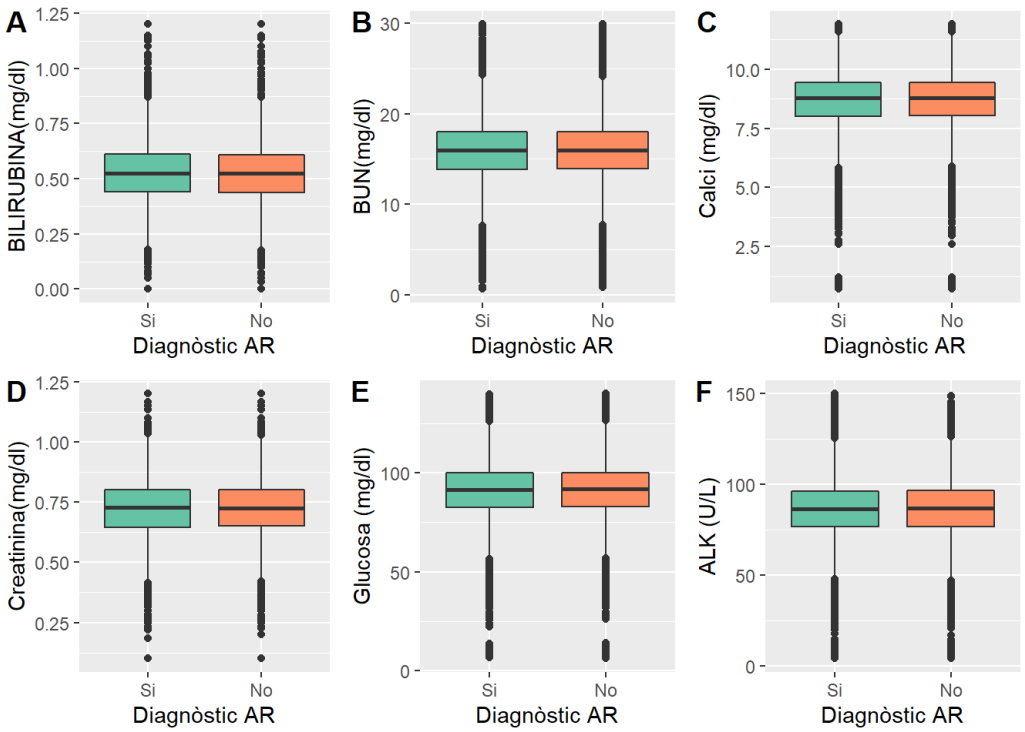


Figura nº 21. Relació entre diagnòstic i paràmetres de laboratori(1): Bilirubina (A), BUN (B), Calci (C), Creatinina (D), Glucosa (E) i ALK (F).

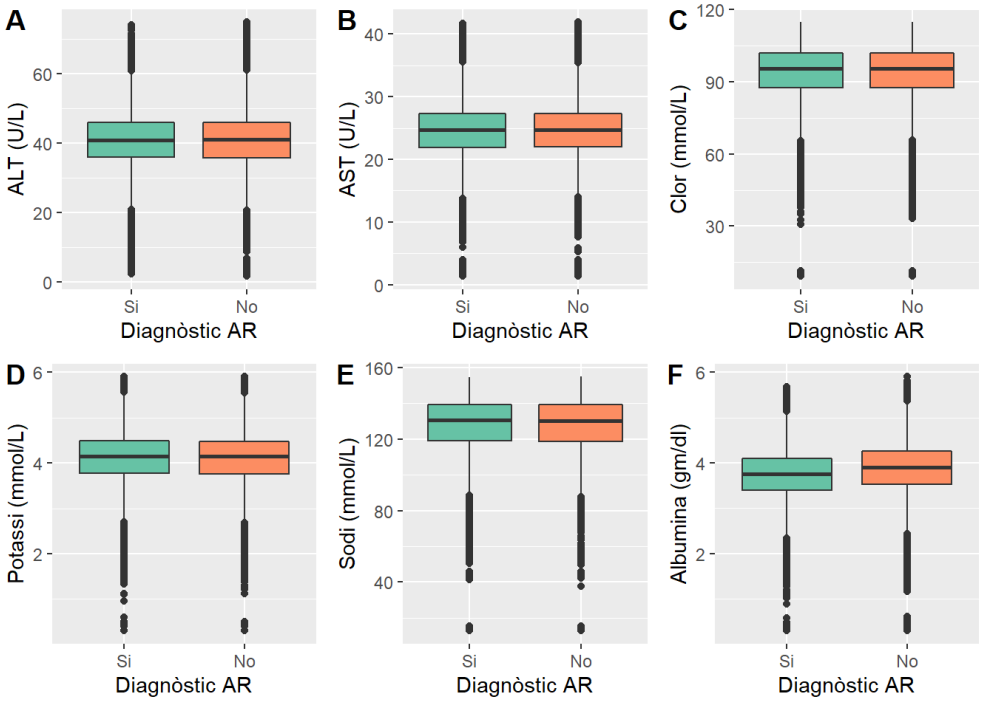


Figura nº 22. Relació entre diagnòstic i paràmetres de laboratori(2): ALT (A), AST (B), Clor (C), Potassi (D), Sodi (E) i Albúmina (F).

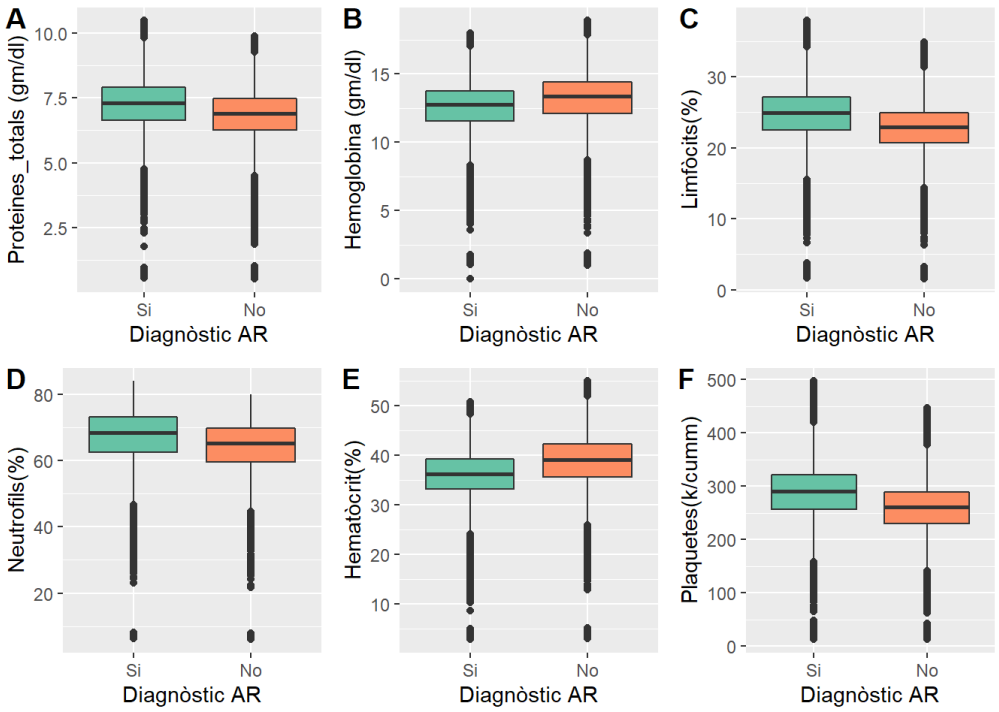


Figura nº 23. Relació entre diagnòstic i paràmetres de laboratori(3): Proteïnes totals (A), Hemoglobina (B), Limfòcits(C), Neutròfils (D), Hematòcrit (E), Plaquetes (F).

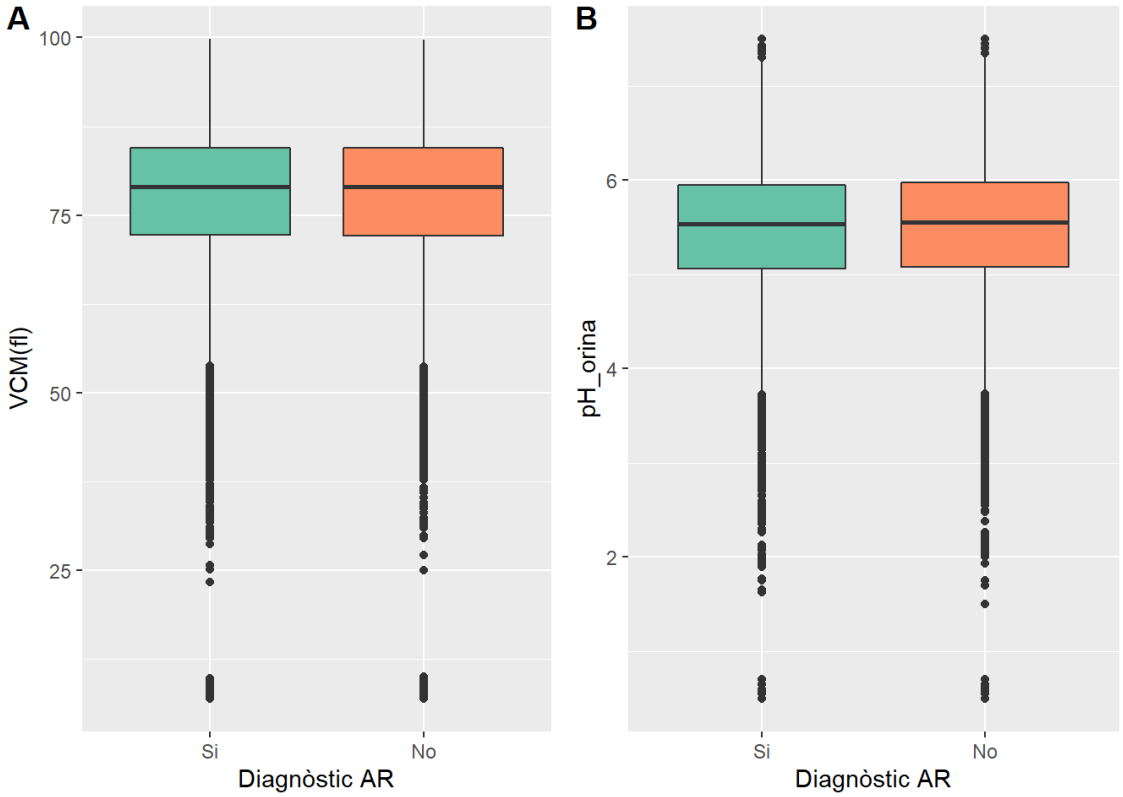


Figura nº 24. Relació entre diagnòstic i paràmetres de laboratori(4): VCM (A) i pH orina (B).

S'observa que hi ha variables que presenten distribucions de valors molt semblants entre grups de diagnòstic (variables de les Figures nº21 o 22), mentres que altres presenten distribucions lleugerament diferents (variables de la Figura nº23).

2.3. ANÀLISI ESTADÍSTIC DE LES VARIABLES

2.3.1. Anàlisi de les variables continues

1) Edat pacients

En primer lloc, s'explora si les dades que conformen la variable segueixen una distribució normal. Anteriorment, s'ha representat un histograma de les dades (Figura nº10-B) que suggereix que no hi ha normalitat en les dades, ja que sembla que segueixen una distribució de "cua dreta". Tot i així, per tal de confirmar-ho, es representa un QQ-plot (Figura nº 25) i s'executa un test de normalitat. Com que la n total de la base de dades és de 28.572, el Shapiro test no és adient ja que només té en compte 5.000 observacions. Per això mateix, s'executa un test Anderson-Darling.

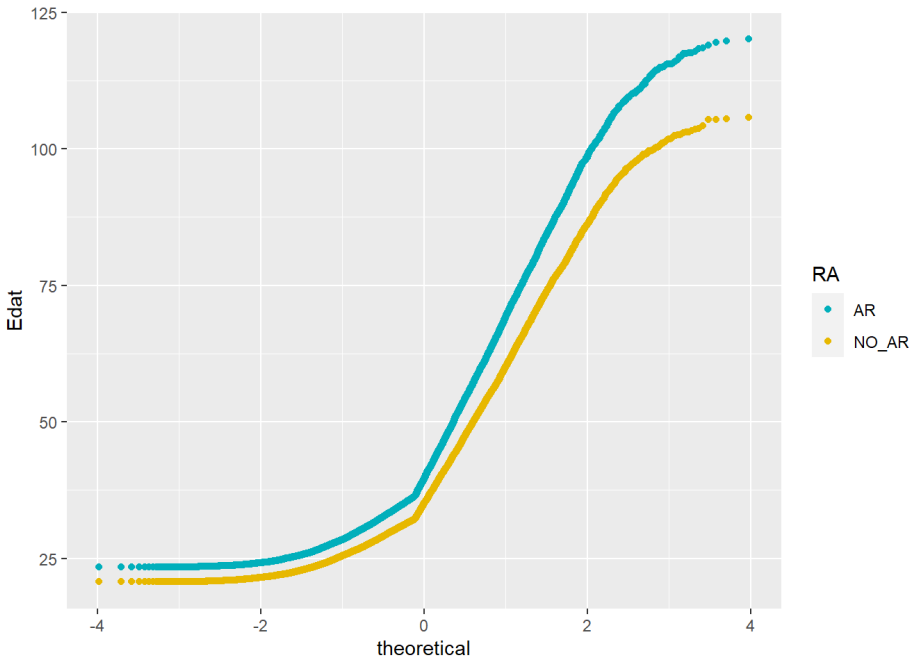


Figura nº 25. QQ-plot de l'edat dels pacients segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la , ja que s'obté un p-valor de < $2.2e^{-16}$. D'aquesta manera, tant el qq-plot com el test de normalitat confirmen que les edats dels pacients no segueixen una distribució normal. Per això, el test estadístic que es realitza per comparar l'edat entre els dos grups de pacients és un Mann-Whitney. El resum resultant de la comparació estadística de les edats dels pacients segons diagnòstic es mostra a la taula nº 5.

Grups	Mitjana (DE)	p-valor
Diagnosticats amb AR (n= 14.282)	46,90 (20,39)	**<0,001***
No diagnosticats amb AR (n=14.290)	41,27 (17,65)	

Taula nº 5. Resum de la comparació entre grups de l'edat dels pacients.

Efectivament, es troben diferències significatives entre grups per les edats dels pacients.

2) Durada ingrés hospital

Anteriorment s'ha representat un histograma de les dades (Figura nº10-C) que suggereix que no hi ha normalitat en les dades. Tot i així, es representa un QQ-plot (Figura nº 26), i s'executa el test de normalitat d'Anderson-Darling.

La durada de l'ingrès hospitalari es troba en format Data, pel que prèviament a l'anàlisi s'ha de transformar en format numèric.

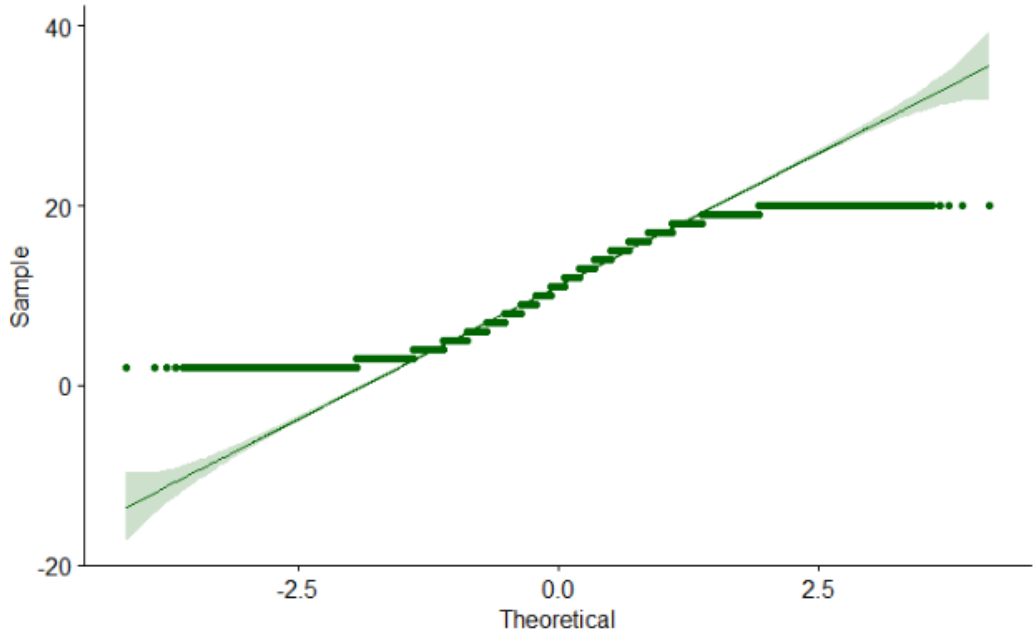


Figura nº 26. QQ-plot de la durada de l'ingrès hospitalari dels pacients.

El test d'Anderson-Darling rebutja l'hipotesis nul·la , ja que s'obté un p-valor de < $2.2e^{-16}$. D'aquesta manera, tant el qq-plot com el test de normalitat confirmen que la durada de l'ingrès hospitalari no segueix una distribució normal. Així mateix, el test estadístic utilitzat és és un Mann-Whitney. El resum resultant de la comparació estadística de les durades dels ingresos segons diagnòstic es mostra a la taula nº 6.

Grups	Mitjana (DE)	p-valor
Diagnosticats amb AR (n= 14.282)	11,02 (5,18)	0,642
No diagnosticats amb AR (n=14.290)	10,99 (5,20)	

Taula nº 6. Resum de la comparació entre grups de la durada dels ingressos hospitalaris.

Tal i com es mostra, no hi ha diferències significatives entre grups per les durades dels ingressos hospitalaris.

3) Paràmetres expresats en mg/dl (Bilirubina-BUN-Calci-Creatinina-Glucosa)

Per cada paràmetre, es representa un QQ-plot (Figura nº 27) i s'executa el test d'Anderson-Darling.

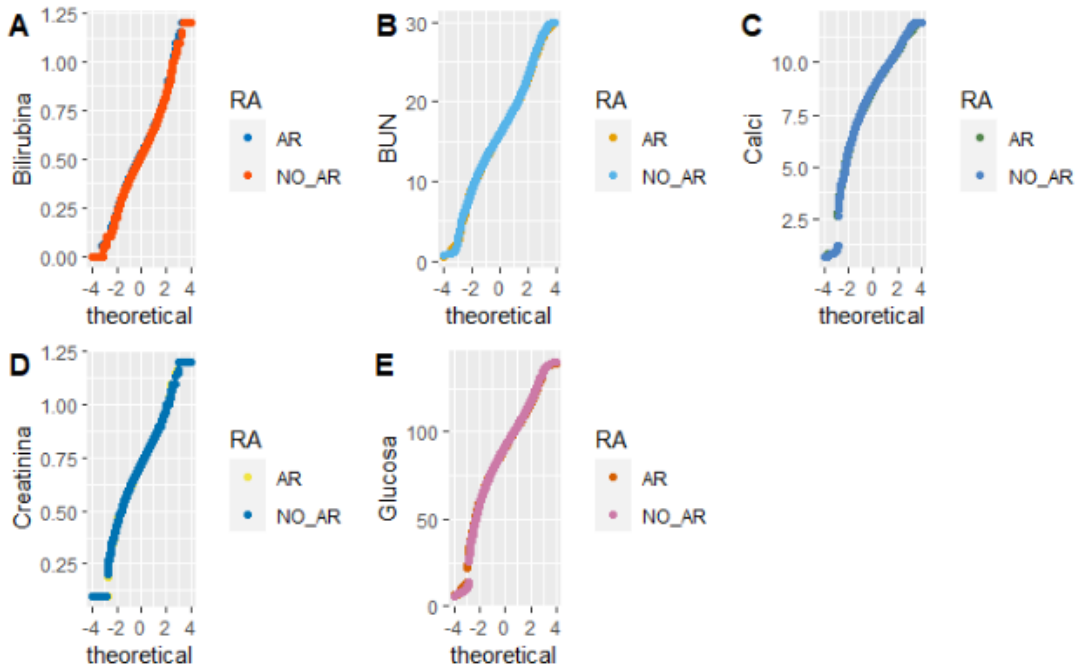


Figura nº 27. QQ-plots dels diferents paràmetres expressats en mg/dl segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la per tots els paràmetres , ja que s'obté un p-valor de $< 2.2e^{-16}$. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups per tots els paràmetres. El resum resultant es mostra a la taula nº 7.

Grups	**Mitjana (DE)**				
	Bilirubina	**BUN**	**Calci**	**Crea**	**Glucosa**
Diagnosticats amb AR (n= 14.282)	0,528 (0,145)	15,95 (3,41)	8,64 (1,19)	0,723 (0,131)	90,75 (14,19)
No diagnosticats amb AR (n=14.290)	0,525 (0,145)	15,98 (3,39)	8,66 (1,20)	0,722 (1,130)	90,99 (14,54)
p-valor	0,084	0,879	0,113	0,912	0,056

Taula nº 7. Resum de la comparació entre grups dels diferents paràmetres expressats en mg/dL.

S'observa que no hi ha diferències significatives entre grups per cap dels paràmetres expressats en mg/dl.

4) Paràmetres expressats en U/L (ALK-ALT-AST)

Per cada paràmetre, es representa un QQ-plot (Figura nº 28) i s'executa el test d'Anderson-Darling.

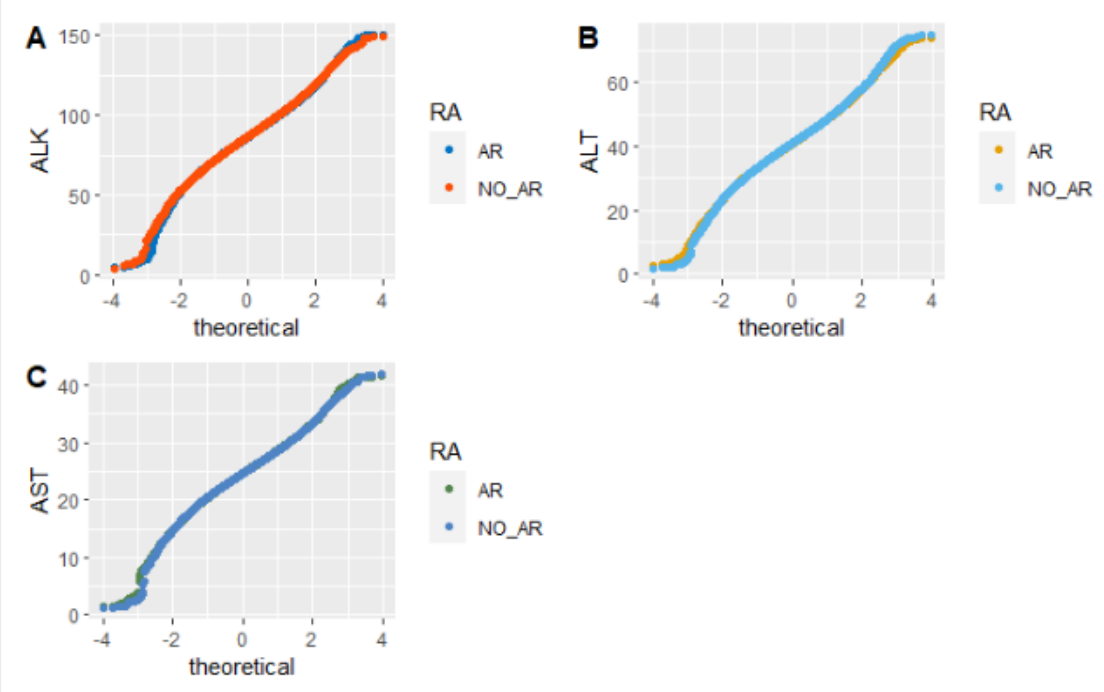


Figura nº 28. QQ-plots dels diferents paràmetres expressats en U/L segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la per tots els paràmetres expressats amb U/L, ja que s'obté un p-valor de < $2.2e^{-16}$. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups per tots els paràmetres expressats amb U/L. El resum resultant es mostra a la taula nº 8.

Grups	**Mitjana (DE)**		
	ALK	**ALT**	**AST**
Diagnosticats amb AR (n= 14.282)	86,38 (16,18)	40,92 (8,33)	24,574 (4,515)
No diagnosticats amb AR (n=14.290)	86,63 (16,17)	40,99 (8,48)	24,566 (4,526)
p-valor	0,300	0,423	0,965

Taula nº 8. Resum de la comparació entre grups dels diferents paràmetres expressats en U/L.

S'observa que no hi ha diferències significatives entre grups per cap dels paràmetres expressats en U/L.

5) Paràmetres expressats en mmol/L (Clor-Potassi-Sodi)

Per cada paràmetre, es representa un QQ-plot (Figura nº 29) i s'executa el test d'Anderson-Darling.

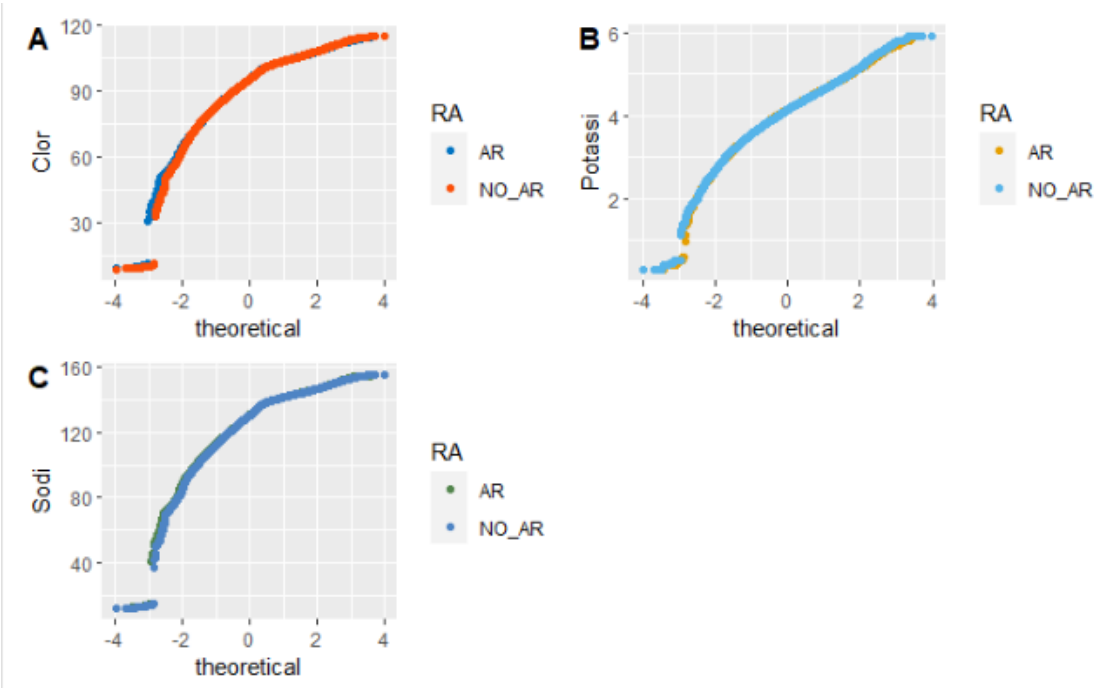


Figura nº 29. QQ-plots dels diferents paràmetres expressats en mmol/L segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la per tots els paràmetres expressats amb mmol/L , ja que s'obté un p-valor de $< 2.2e^{-16}$. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups per tots els paràmetres expressats amb mmol/L. El resum resultant es mostra a la taula nº 9.

Grups	**Mitjana (DE)**		
	Clor	**Potassi**	**Sodi**
Diagnosticats amb AR (n= 14.282)	93,43 (11,43)	4,088 (0,607)	127,44 (15,79)
No diagnosticats amb AR (n=14.290)	93,30 (11,85)	4,086 (0,604)	127,10 (16,30)
p-valor	0,724	0,336	0,295

Taula nº 9. Resum de la comparació entre grups dels diferents paràmetres expressats en mmol/L.

S'observa que no hi ha diferències significatives entre grups per cap dels paràmetres expressats en mmol/L.

6) Paràmetres expressats en gm/dl (Albúmina-Proteïnes totals-Hemoglobina)

Per cada paràmetre, es representa un QQ-plot (Figura nº 30) i s'executa el test d'Anderson-Darling.

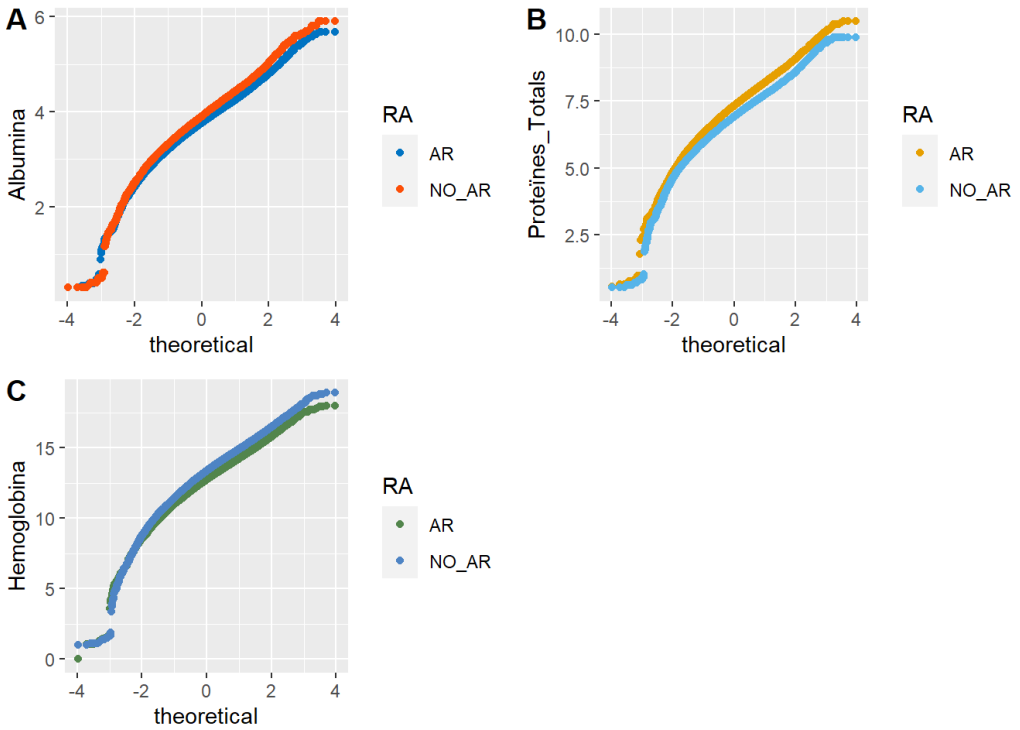


Figura nº 30. QQ-plots dels diferents paràmetres expressats en gm/dl segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la per tots els paràmetres expressats amb gm/dl , ja que s'obté un p-valor de < $2.2e^{-16}$. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups per tots els paràmetres expressats amb gm/dl. El resum resultant es mostra a la taula nº 10.

Grups	**Mitjana (DE)**		
	Albúmina	**Proteïnes totals**	**Hemoglobina**
Diagnosticats amb AR (n= 14.282)	3,72 (0,58)	7,23 (1,05)	12,59 (1,80)
No diagnosticats amb AR (n=14.290)	3,86 (0,62)	6,82 (1,00)	13,17 (1,91)
p-valor	**<0,001***	**<0,001***	**<0,001***

Taula nº 10. Resum de la comparació entre grups dels diferents paràmetres expressats en gm/dL.

S'observa que **hi ha diferències significatives entre grups** per tots els paràmetres expressats en gm/dL.

7) Paràmetres expressats en % (Limfòcits-Neutròfils-Hematòcrit)

Per cada paràmetre, es representa un QQ-plot (Figura nº 31) i s'executa el test d'Anderson-Darling.

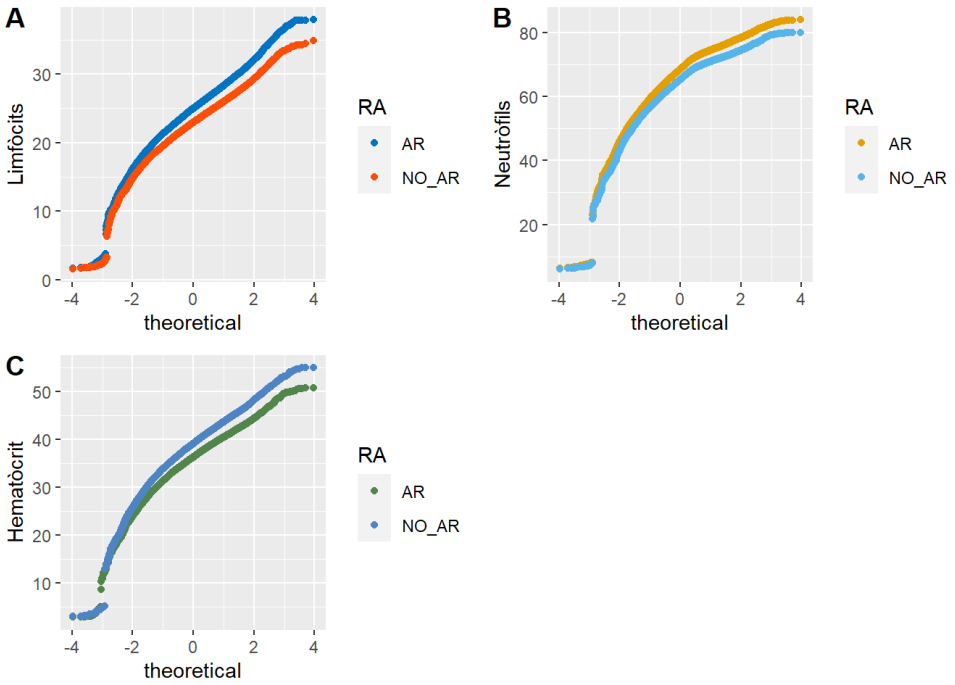


Figura nº 31. QQ-plots dels diferents paràmetres expressats en % segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la per tots els paràmetres expressats amb % , ja que s'obté un p-valor de < $2.2e^{-16}$. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups per tots els paràmetres expressats amb %. El resum resultant es mostra a la taula nº11.

Grups	**Mitjana (DE)**		
	Limfòcits	**Neutròfils**	**Hematòcrit**
Diagnosticats amb AR (n= 14.282)	24,75 (3,92)	66,85 (8,57)	35,80 (5,06)
No diagnosticats amb AR (n=14.290)	22,71 (3,58)	63,70 (8,15)	38,61 (5,51)
p-valor	**<0,001***	**<0,001***	**<0,001***

Taula nº 11. Resum de la comparació entre grups dels diferents paràmetres expressats en %.

S'observa que **hi ha diferències significatives entre grups** per tots els paràmetres expressats en %.

8) Plaquetes

Es representa un QQ-plot del recompte de plaquetes (Figura nº 32) i s'executa el test d'Anderson-Darling per comprovar la normalitat de les dades.

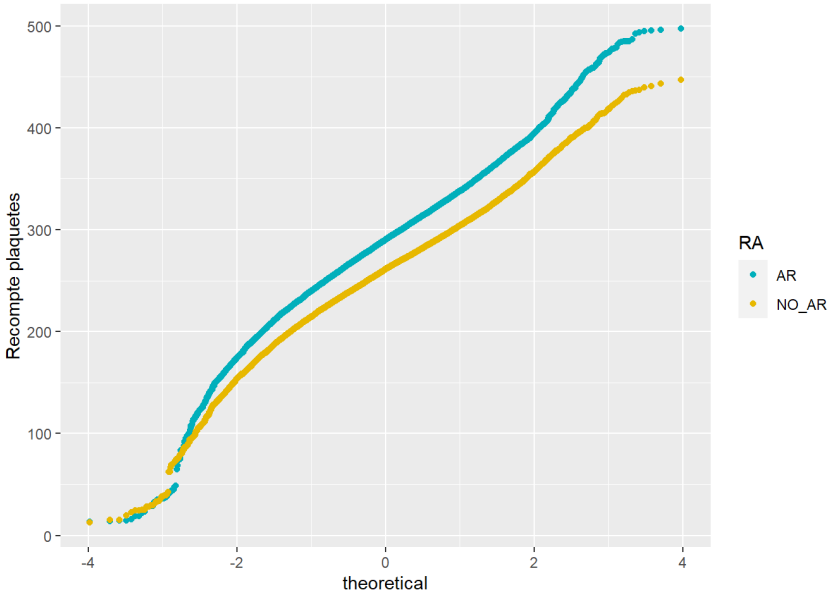


Figura nº 32. QQ-plot del recompte de plaquetes segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la, ja que s'obté un p-valor de < 2.2e-16. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups del recompte de plaquetes. El resum resultant es mostra a la taula nº 12.

Grups	Mitjana (DE)	p-valor
Diagnosticats amb AR (n= 14.282)	288,55 (54,12)	**<0,001***
No diagnosticats amb AR (n=14.290)	259,00 (49,13)	

Taula nº 12. Resum de la comparació entre grups del recompte de plaquetes .

S'observa que **hi ha diferències significatives entre grups** pel recompte de plaquetes.

9) VCM

Es representa un QQ-plot del VCM (Figura nº 33) i s'executa el test d'Anderson-Darling per comprovar la normalitat de les dades.

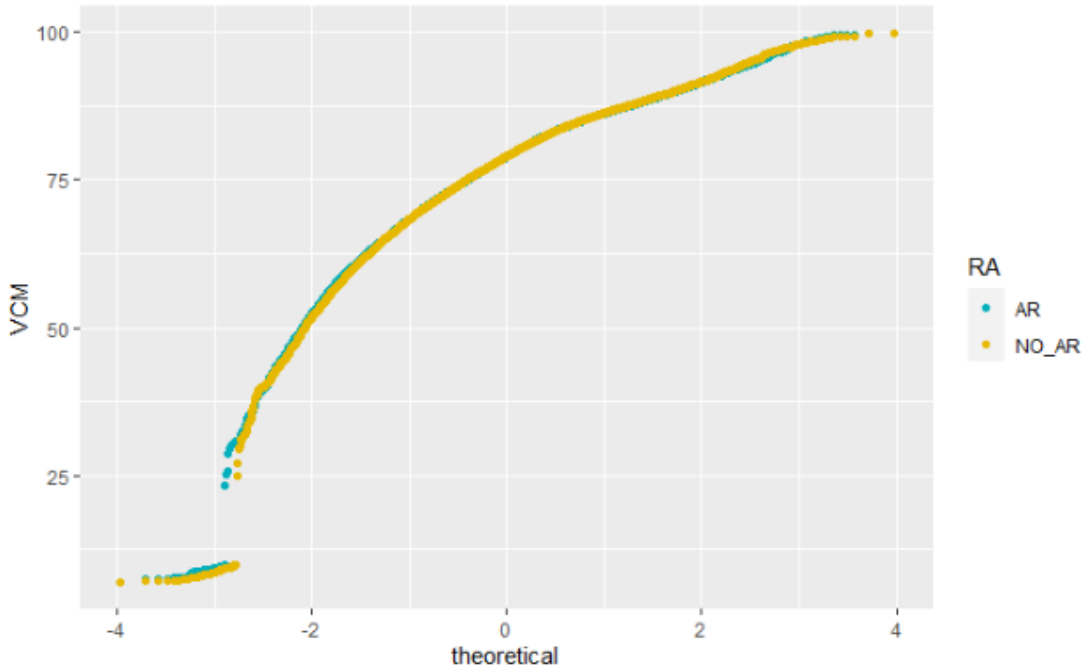


Figura nº 33. QQ-plot del VCM segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la, ja que s'obté un p-valor de < 2.2e-16. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups del VCM. Els resultats es mostren a la taula nº 13.

Grups	Mitjana (DE)	p-valor
Diagnosticats amb AR (n= 14.282)	77,28 (10,13)	0,943
No diagnosticats amb AR (n=14.290)	77,23 (10,34)	

Taula nº 13. Resum de la comparació entre grups del VCM.

S'observa que no hi ha diferències significatives entre grups pel VCM.

10) pH orina

Es representa un QQ-plot del pH de l'orina (Figura nº 34) i s'executa el test d'Anderson-Darling per comprovar la normalitat de les dades.

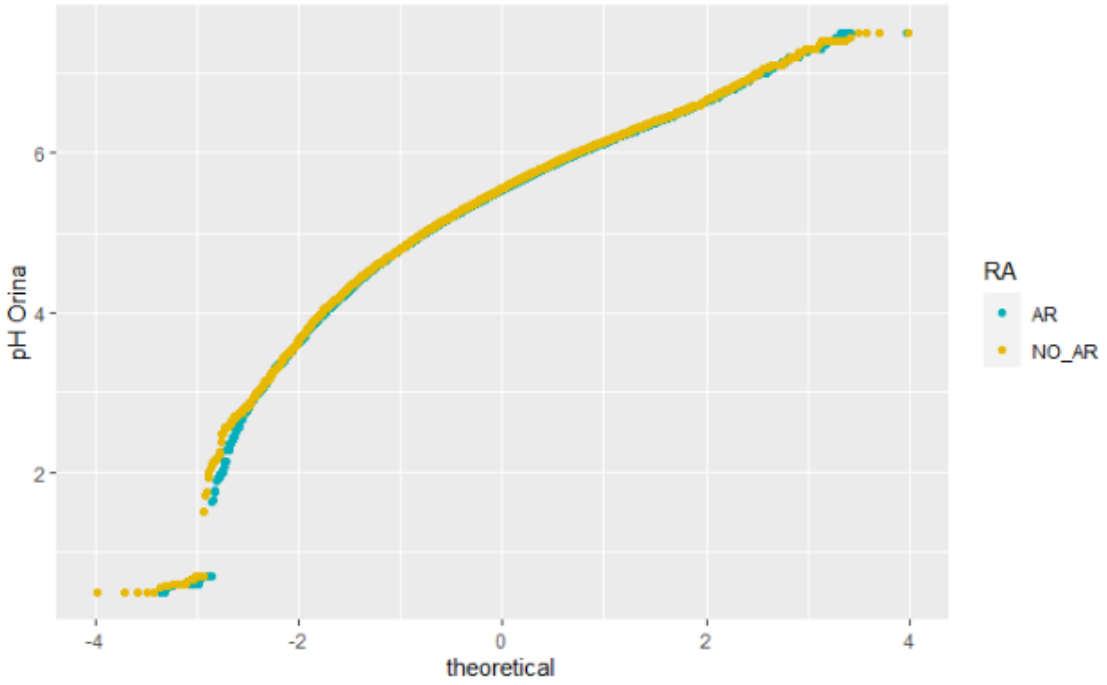


Figura nº 34. QQ-plot del pH de l'orina segons grup de diagnòstic.

El test d'Anderson-Darling rebutja l'hipotesis nul·la, ja que s'obté un p-valor de < $2.2e^{-16}$. D'aquesta manera, es rebutja que les dades segueixin una distribució normal. Així doncs, s'utilitza un Mann-Whitney per les comparacions entre grups del VCM. Els resultats es mostren a la taula nº 14.

Grups	Mitjana (DE)	p-valor
Diagnosticats amb AR (n= 14.282)	5,45 (0,76)	0,051
No diagnosticats amb AR (n=14.290)	5,46 (0,75)	

Taula nº 14. Resum de la comparació entre grups del pH de l'orina.

S'observa que no hi ha diferències significatives entre grups pel pH de l'Orina.

2.3.2. Anàlisi de les variables categòriques

1) Génere :

Es construeix una taula de contingència segons gènere i diagnòstic i, tot seguit, es realitza un test de chi quadrat per tal d'analitzar si es veuen diferències significatives (Taula nº 15).

	Gènere	
Diagnòstic	**Dones**	**Homes**
Diagnosticats amb AR	10.350	3.932
No diagnosticats amb AR	7.448	6.842
p-valor	**<0,001***	

Taula nº 15. Taula de contingència segons gènere i diagnòstic.

Efectivament, hi ha diferències significatives que demostren que el diagnòstic d'Artritis Reumatoide està relacionat amb el gènere, en aquest cas, amb el gènere femení.

2) Raça:

La taula de contingència segons raça disposa de quatre categòries (Taula nº 16).

	Raça			
Diagnòstic	**Africà**	**Asiàtic**	**Desconegut**	**Blanc**
Diagnosticats amb AR	2.067	3.286	1.881	7.048
No diagnosticats amb AR	2.129	3.288	1.897	6.976

Taula nº 16. Taula de contingència segons diagnòstic i raça.

En aquest cas, també es realitza un test chi-quadrat, però al tenir quatre categòries també es realitza un test post-hoc de comparacions 2 a 2. Per això, s'utilitza la funció *chisq.multcomp()* del paquet *RVAdeMemoire*. Els resultats es mostren a la taula nº17.

Comparació	**p-valor**
Chi-quadrat global	0,717
Desconegut no AR vs Desconegut AR	0,795
Africà AR vs Africà no AR	0,339
Asiàtic AR vs Asiàtic no AR	0,980
Blanc AR vs Blanc no AR	0,543

Taula nº 17. Chi quadrat i comparacions post-hoc de l'associació raça -diagnostic.

En aquest cas no hi ha diferències significatives, pel que no es demostra l'associació de la raça amb el diagnòstic.

2.3.3. Compilació dels factors de risc associats a la malaltia d'interès

Finalment, es conforma una taula (Taula nº18) on s'engloba, de manera resumida, totes les comparacions de variables realitzades en els apartats anteriors i les conclusions derivades.

Variable	p-valor obtingut (comparació entre grups diagnòstic)	Associació amb AR
Edat	<0,001	SÍ
Ingrés hospitalari	>0,05	NO
Bilirubina	>0,05	NO
BUN	>0,05	NO
Calci	>0,05	NO
Creatinina	>0,05	NO
Glucosa	>0,05	NO
ALK	>0,05	NO
ALT	>0,05	NO
AST	>0,05	NO
Clor	>0,05	NO
Potassi	>0,05	NO
Sodi	>0,05	NO
Albúmina	<0,001	SÍ
Proteïnes totals	<0,001	SÍ
Hemoglobina	<0,001	SÍ
Limfòcits	<0,001	SÍ
Neutròfils	<0,001	SÍ
Hematòcrit	<0,001	SÍ
Plaquetes	<0,001	SÍ
VCM	>0,05	NO
pH orina	>0,05	NO
Gènere	<0,001	SÍ
Raça	>0,05	NO

Taula nº 18. Resum de la comparació de variables segons grup de diagnòstic.

D'aquesta manera, totes les variables marcades en verd s'utilitzen en el següent apartat del cas pràctic on es posen a prova diferents models de predicció.

3. ALGORITMES I MODELS DE PREDICCIÓ DE MALALTIA

3.1 DESCRIPCIÓ DELS MODELS I ALGORITMES UTILITZATS EN ANÀLISI DE PREDICCIÓ DE RISC

3.1.1. Algoritmes més utilitzats en predicció de risc utilitzant dades d'EMR

A la revisió sistemàtica de Goldstein et al.[4], s'analitzen publicacions referents al desenvolupament de models de predicció de risc utilitzant dades EMR. En aquesta revisió es descriu que els GLM (regressió logística, regressió de Cox…) són els algoritmes més utilitzats en predicció de risc. No obstant, algoritmes de *Machine Learning* (mètodes bayesians, *Random Forest*) i les regressions regularitzades (LASSO, regressió de *Rigde*) també són mètodes bastant populars.
A més a més, en altres publicacions [2,7], es troba descrit que algoritmes de SVM i ANN també són útils per aquest tipus d'anàlisi.

3.1.2. Selecció i justificació d'algoritmes per l'estudi

La variable resposta, en aquest cas, es tracta d'una variable categòrica (Diagnòstic d'artritis reumatoide) amb dos nivells: SI / NO. Per tant, es necessita un algoritme que permeti classificar els pacients en les dos categories mencionades anteriorment segons els valors d'una sèrie de variables, que són les que s'han seleccionat a l'apartat anterior.
Així doncs, es tracta d'un problema de classificació, pel que es descarten tots els algoritmes o models que no són adients en aquest cas: totes les regressions lineals (incloses LASSO o la regressió de Ridge) i els models de supervivència, com és la regressió de Cox.
D'aquesta manera, els algoritmes i models de predicció de risc seleccionats es llisten a continuació:

- **Regressió logística múltiple:** és un mètode de classificació de variables categòriques binàries molt popular. En aquest model normalment no s'utilitzen masses variables predictores, ja que pot arribar a ser complicat, però tot i així, s'intentarà realitzar una predicció de malaltia utilitzant aquest mètode.
- **Algoritme de Naïve-Bayes:** és el model de classificació més simple dins dels algoritmes de *Machine Learning*, però tot i així, sol funcionar bastant bé. Varis autors utilitzen mètodes Bayesians quan tenen un alt nombre de variables per crear el model. Encara que no sol funcionar massa bé amb variable numèriques, s'intentarà.
- **Random forest:** és un algoritme de *Machine Learning* flexible i fàcil, que produeix bons resultats la major part del temps. Es pot utilitzar tant en classificació com en regressió.
- **SVM**: és un mètode molt eficient quan es vol realitzar classificació de dues categories. Funciona bé amb moltes variables predictores.
- **ANN:** Funciona molt bé tant en classificacions binàries com múltiples. És un model que treballa bé quan es disposa d'un volum de dades elevat.

3.1.3. Descripció, fortaleses i debilitats

1) <u>Regressió logística múltiple</u>

La regressió logística simple és un mètode de regressió que permet estimar la probabilitat d'una variable qualitativa binària en funció d'una variable quantitativa. Una de les principals aplicacions de la regressió logística és la de classificació binària, en el que les observacions es classifiquen en un grup o altre depenent del valor del predictor. La regressió logística múltiple és una extensió de la regressió logística simple. Es basa en els mateixos principis que la regressió logística simple però ampliant el número de predictors, que poden ser tant continus com categòrics [10,16].

A l'hora d'avaluar la validesa i la qualitat d'aquest model, s'analitza tant el model en el seu conjunt com els predictors que el formen. Es considera que el model és útil si es capaç de mostrar una millora respecte al model nul (sense predictors). Existeixen 3 tests estadístics que quantifiquen la millora: *likelihood ratio, score i Wald test*. Si els tres no arriben a la mateixa conclusió, es recomana basar-se en el *likelihood ratio*.

Es presenta una taula de fortaleses i debilitats d'aquest algoritme:

Fortaleses	Debilitats
* Útil en problemes de classificació de variables binàries.	* Selecció característiques prèvia entrenament és clau.
* Predictors poden ser tant variables continues com categòriques.	* Impossibilitat de resoldre directament problemes no lineals.
* Eficaç i simple, no requereix de grans recursos computacionals.	* Variable objectiu ha de ser linealment separable, sinó no classificarà correctament.
* El pes de cada característica determina la importància que té en la decisió final.	* Dependència en les característiques.

Taula nº 19. Taula de fortaleses i debilitats de la regressió logística múltiple.

2) Algoritme de Naïve Bayes

L'algoritme de *Naïve Bayes* és un dels models de classificació de *Machine Learning* més senzills. Aplica els principis del teorema de *Bayes.*

Es presenta una taula de fortaleses i debilitats d'aquest algoritme:

Fortaleses	Debilitats
* Simple, ràpid i efectiu.	* Assumeix que totes les variables són igual d' importants i independents.
* Funciona bé amb dades sorolloses i valors mancants.	* No funciona del tot bé amb bases de dades amb moltes variables numèriques.
* Necessita poques mostres a la fase d'entrenament, tot i que també treballa bé amb un número de mostres elevat.	* Les probabilitats estimades són menys fiables que les classes predites.
* És fàcil d'obtenir una estimació de probabilitat per una de predicció.	

Taula nº 20. Taula de fortaleses i debilitats de l'algoritme de *Naive Bayes* .

3) Algoritme de Random Forest

Es basa en fer una selecció al atzar de part de les dades i part de les característiques per crear un arbre de decisions. El procés es repeteix de forma independent per un número prefixat d'arbres, fins a crear un bosc. La decisió es pren per votació del conjunt d'arbres. Permet treballar bé amb grans conjunts de dades.

La taula de fortaleses i debilitats és la següent:

Fortaleses	Debilitats
* Classificador que actua bé en la majoria de problemes.	* Els models no són fàcilment interpretables.
* Permet manejar variables numèriques o qualitatives, incloent dades incompletes o amb soroll.	* Necessita dedicar temps al ajustament del model a les dades.
* Elimina característiques poc importants, seleccionant només les més importants.	
* Pot utilitzar-se en conjunts de dades molt grans o amb moltes característiques.	

Taula nº 21. Taula de fortaleses i debilitats de l'algoritme de *Random Forest*.

4) Algoritme de SVM

És un mètode basat en aprenentatge que s'utilitza per la resolució de problemes de regressió i classificació. En cas de tenir dades linealment separables, l'algoritme es basa en trobar la millor separació possible entre classes (hiperplà), que maximitza el marge de separació entre classes (MMH).

En cas de no tenir dades linealment separables, utilitza les funcions *Kernel*, les quals resolen els problemes de classificació traslladant dades a un espai a on l'hiperplà és lineal, és a dir, més fàcil d'obtenir.

La taula de fortaleses i debilitats és la següent:

Fortaleses	Debilitats
* Pot utilitzar-se en classificació o problemes de predicció numèrica.	* Trobar el millor model requereix provar varies combinacions entre les funcions *Kernel* i els paràmetres del model.
* No s'influencia per les dades sorolloses i no té tendència a sobre ajustar	* Pot ser lent d'entrenar, sobretot si les dades d'entrada tenen moltes característiques.
* És més fàcil d'utilitzar que les xarxes neuronals, gràcies a l'existència d'algoritmes SVM ben suportats.	* És un model complex, molt difícil d'interpretar.
* Gran precisió i perfil alt d'èxits.	

Taula nº 22. Taula de fortaleses i debilitats de l'algoritme de l'algoritme SVM.

5) Algoritme de ANN

L'algoritme ANN es basa en una xarxa de neurones artificials per resoldre problemes d'aprenentatge, basat en el comportament de les neurones biològiques.

Es crea una xarxa amb diferents capes interconnectades per processar la informació. Cada capa esta formada per un grup de nodes que transmeten informació als nodes de les següents capes.

És un sistema que aprèn i es forma a sí mateix, en lloc de ser programat de forma explícita.

La taula de fortaleses i debilitats és la següent:

Fortaleses	Debilitats
* Pot utilitzar-se en classificació o problemes de predicció numèrica.	* Computacionalment és molt intens i lent d'entrenar, sobretot si la topologia de la xarxa és complexa.
* Capaç de modelar patrons més complexos que altres algoritmes.	* Té tendència a sobre-ajustar les dades d'entrenament.
* Realitza poques assumpcions de les relacions subjacents de les dades.	* És un model complex, molt difícil d'interpretar.

Taula nº 23. Taula de fortaleses i debilitats de l'algoritme de l'algoritme ANN.

3.2. FORMULACIÓ I EXECUCIÓ DELS MODELS

3.2.1. Formulació, generació i entrenament dels models

El primer pas en la formulació dels models és definir què es vol predir/classificar i quines variables es volen utilitzar com a predictores per tal d'aconseguir-ho.

L'objectiu és classificar un pacient en si té risc de desenvolupar Artritis Reumatoide o no a partir de la informació que aporten les variables escollides a la primera part del treball, que són: l'edat, el gènere, els valor d'albúmina, valor de proteïnes totals, l'hemoglobina, els limfòcits, els neutròfils, l'hematòcrit i les plaquetes.

A continuació, es redueix la base de dades EMR utilitzada fins aquest punt per tal de que només contingui les variables d'interès. De la mateixa manera, es canvien els nom d'algunes de les variables per facilitar l'anàlisi i es re-codifica el diagnòstic d'Artritis Reumatoide en forma binària (1- Artritis Reumatoide / 0- No Artritis Reumatoide). D'aquesta manera, la base de dades queda configurada amb 9 variables:

Variable	Descripció	Tipus	Rang / Nivells
RA	Diagnòstic Artritis	Factor	2 nivells (0/1)
Patient.Gender	Gènere	Factor	2 nivells (Home/Dona)
PatientAge	Edat	Contínua	20,70 – 120,09
Alb	Valors d'albumina	Contínua	0,29 – 5,90
Prot	Valors proteïnes totals	Contínua	0,50-10,49
Neutros	Neutròfils absoluts	Contínua	6,00-83,89
Limfos	Limfòcits absoluts	Contínua	1,50-37,93
Hematocrit	Valors hematòcrit	Contínua	2,87-54,90
Hemoglobina	Valors hemoglobina	Contínua	0,00-18,90
Plaquetes	Valors plaquetes	Contínua	12,50-497,20

Taula nº 24. Variables seleccionades per la construcció de models de predicció.

Tot seguit, es procedeix a formular els diferents models.

3.2.1.1. Regressió logística múltiple

En primer lloc, es divideix la base de dades en un conjunt de dades d'entrenament i un conjunt de dades test. Per les dades d'entrenament, s'utilitzen 2/3 parts de les dades, mentre que per les dades test, s'utilitza l'1/3 restant. Per la formulació de la regressió logística, s'utilitzen les dades d'entrenament:

```
model_logistic <- glm (RA ~ Patient.Gender + PatientAge + Alb + Prot + Hemoglo
bina + Limfos + Neutros + Hematòcrit + Plaquetes, data = AR_LR_train , family =
 "binomial")
```

Tal i com s'observa a la Taula nº25, totes les variables estan relacionades de forma significativa amb el diagnòstic d'Artritis Reumatoide. Segons el model, el logaritme *odds* que un pacient desenvolupi Artritis Reumatoide es troba negativament relacionat amb:

- El gènere masculí (L'*Odds* que una dona desenvolupi AR és $e^{0.84}$ =2,32 vegades major que un home).
- Els valors d'albúmina, d'hemoglobina i l'hematòcrit.

Per altre banda, es troba positivament relacionat amb l'edat, les proteïnes totals, els limfòcits, els neutròfils i les plaquetes.

Variable	Coeficient	p-valor
Patient.Gender	-0,844	**<0,001***
PatientAge	0,016	**<0,001***
Alb	-0,437	**<0,001***
Prot	0,398	**<0,001***
Hemoglobina	-0,178	**<0,001***
Limfos	0,157	**<0,001***
Neutros	0,047	**<0,001***
Hematocrit	-0,109	**<0,001***
Plaquetes	0,012	**<0,001***

Taula nº25. Coeficients i p-valor obtinguts amb la regressió logística múltiple.

3.2.1.2. Algoritme de *Naïve Bayes*

En aquest cas, es factoritzen totes les variables, ja que aquest algoritme treballa millor amb variables factoritzades que amb variables numèriques. La factorització es realitza seguint les següents directrius:

- Edat: Jove (Fins a 35 anys) / Adult (Dels 35 a 75 anys) /Ancià (Superior a 75 anys).
- Albúmina: Baix (Valors < 3) / Normal (Entre 3 i 5,4) / Alt (Valors > 5,4).
- Proteïnes: Baix (Valors < 6) / Normal (Entre 6 i 8,3) / Alt (Valors > 8,3).
- Hemoglobina: Baix (Valors < 11,6) / Normal (Entre 11,6 i 16,5) / Alt (Valors > 16,5).
- Limfòcits: Baix (Valors < 20) / Normal (Entre 20 i 40) / Alt (Valors > 40).
- Neutròfils: Baix (Valors < 54) / Normal (Entre 54 i 70) / Alt (Valors > 70).
- Hematòcrit: Baix (Valors < 36) / Normal (Entre 36 i 50) / Alt (Valors > 50).
- Plaquetes: Baix (Valors < 150) / Normal (Entre 150 i 400) / Alt (Valors > 400).

Tot seguit es procedeix en la partició de les dades, tal i com s'ha realitzat amb el model anterior. Així doncs, 19.143 pacients s'utilitzen per l'entrenament del model i 9.429 pacients s'utilitzen per la posterior avaluació del rendiment d'aquest.

Un cop realitzat aquest pas, es procedeix a generar el model *Naïve Bayes* amb *Laplace* = 0, que seria el model per defecte. El contingut de l'objecte R resultant conté les probabilitats condicionades de cada categoria segons el risc del pacient per desenvolupar AR.

```
model_NaiveBayes<- naiveBayes(AR_NB_train[2:10],AR_NB_train$RA, laplace = 0)
```

3.2.1.3. Algoritme de Random Forest

Per aquest algoritme no hi ha problema en utilitzar directament les variables numèriques, pel que s'utilitza la base de dades original.

Un cop feta la divisió de les dades en els dos conjunts, es procedeix a formular el model amb els paràmetres per defecte (*ntree = 500*) i *mtry = sqrt(*p):

```
model_RF <- randomForest(RA ~., data=AR_RF_train, ntree = 500)
```

El resultat de l'objecte anterior mostra que el bosc inclou 500 arbres, que es proven 3 variables en cada divisió i que la *ratio* d'error és de 1,71%. Per altra banda, la Figura nº35 mostra la importància que tenen les variables dins el model, visualitzant així que la variable neutròfils és la que pren més valor en aquest cas.

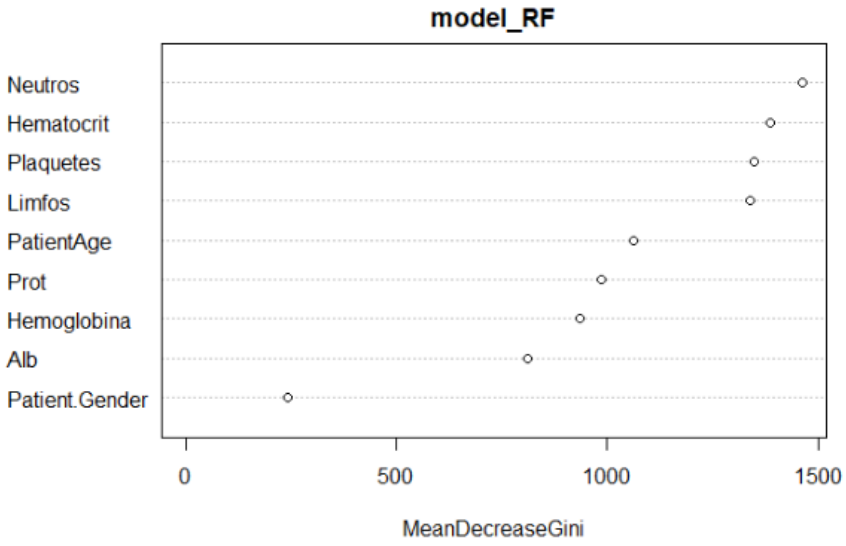


Figura nº35. Importància de les variables dins el model de *Random Forest*.

3.2.1.4. Algoritme de SVM

No és necessari tornar a dividir les dades en dos parts, ja que l'algoritme SVM permet utilitzar les particions utilitzades amb *Random Forest*. Així doncs, es procedeix directament a construir un primer model SVM lineal:

```
model_SVM_lineal <- ksvm(RA ~., data=AR_RF_train, kernel='vanilladot')
```

3.2.1.5. Algoritme de ANN

En aquest cas, la funció que s'utilitza per la construcció de l'algoritme ANN treballa millor si totes les variables numèriques es troben normalitzades en el rang (0,1). Per tant, la base de dades queda conformada de la següent forma:

Variable	Mitjana	Rang
RA	NA	NA
Patient.Gender	NA	NA
PatientAge	0,235	0,00 – 1,00
Alb	0,624	0,00 – 1,00
Prot	0,654	0,00 – 1,00
Neutros	0,761	0,00 – 1,00
Limfos	0,610	0,00 – 1,00
Hematocrit	0,660	0,00 – 1,00
Hemoglobina	0,681	0,00 – 1,00
Plaquetes	0,539	0,00 – 1,00

Taula nº26. Variables normalitzades i transformades per la construcció de l'algoritme ANN.

Igual que la resta d'algoritmes, s'utilitzen 2/3 parts de les dades per l'entrenament del model i 1/3 per la posterior avaluació del rendiment.

Finalment, es genera el model ANN amb 1 sol node a la capa intermèdia (Figura nº36).

```
model_ANN_h1 <- train(RA ~ ., train.set, method='nnet',
                 trControl= trainControl(method='none'),
                 tuneGrid= NULL, tuneLength=1 ,
                  trace = FALSE)
```

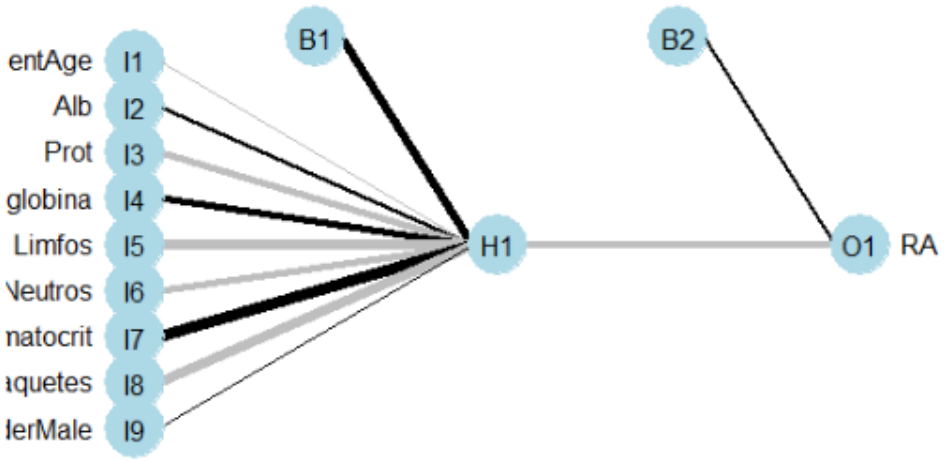


Figura nº36. Representació del model ANN amb 1 node a la capa oculta.

3.2.2. Predicció i avaluació dels models

A continuació s'avalua el rendiment dels models utilitzant el conjunt de dades test. S'utilitzen diferents paràmetres de mesura:

- Precisió (*Accuracy*): Paràmetre bàsic per mesurar el rendiment. Indica el % d'observacions classificades correctament.
- Valor kappa: paràmetre estadístic que mesura l'acord entre dos observadors. Té en compte la distribució dels marginals, que utilitza per corregir l'índex de concordança, excloent així la concordança produïda per atzar [10].

Valor kappa	Estimació concordança
<0,20	Pobre
0,21-0,40	Dèbil
0,41-0,60	Moderada
0,61-0,80	Bona
0,81-1,00	Molt bona

Taula nº 27 . Taula de classificació de l'acord segons el valor kappa.

- Sensibilitat: % de positius reals.
- Especificitat: % de negatius reals.
- Corba ROC: És un gràfic que mostra l'actuació dels models de classificació en tots els llindars de classificació. Representa dos paràmetres: La taxa de positius reals enfront la taxa de falsos positius.
- AUC: Àrea sota la corba. Mesura l'àrea bidimensional per sota de la corba ROC. La interpretació de l'AUC seria la probabilitat que un model classifiqui un positiu aleatori més alt que un negatiu aleatori. Es mou en un rang de valors entre 0 i 1 [10].

Interval de valors AUC	Actuació del model
[0,97-1)	Excel·lent
[0,9-0,97)	Molt bona
[0,75-0,90)	Bona
[0,6-0,75)	Regular
[0,5-0,6)	Dolent
[0,5]	Nul·la

Taula nº 28. Taula d'interpretació de l'actuació del model segons el valor AUC.

3.2.2.1. Regressió logística múltiple

Es diu que una regressió logística proporciona un bon ajustament si demostra una millora respecte a un model amb menys predictors. Això es verifica mitjançant *el test de Likelihood*, que compara la bondat d'ajustament del model complet amb la del model amb menys predictors. Per tal de realitzar aquesta verificació, es formula un model que inclou tots els predictors anteriors menys la variable Gènere i es realitza el test de *Likelihood*.

El p-valor que s'obté del test anterior és < 0.001, pel que es confirma que és millor utilitzar el model més complex, incloent tots els predictors.

Per altra banda, els paràmetres obtinguts que mesuren l'actuació del model és resumeixen en la taula nº 29. La corba ROC obtinguda es mostra a la Figura nº37.

Paràmetre	Valor	Interpretació
Precisió	0,755	Bona
Valor kappa	0,511	Moderada
Sensibilitat	0,761	Bona
Especificitat	0,750	Bona
AUC	0,827	Bona

Taula nº 29. Paràmetres mesura actuació del model regressió logística múltiple.

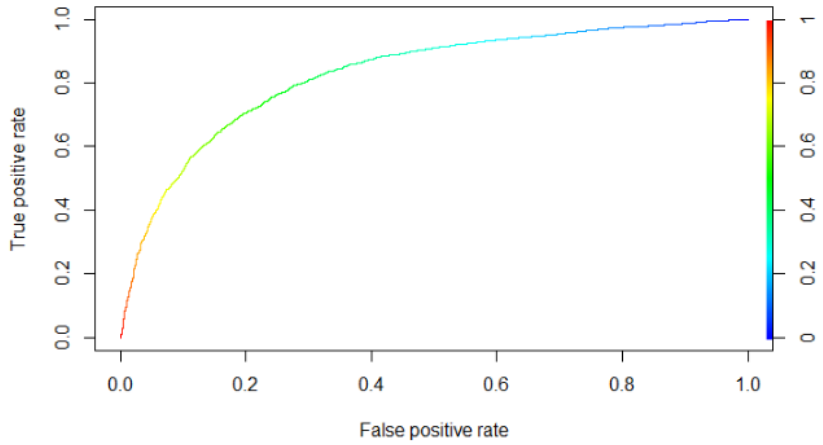


Figura nº37. Corba ROC del model de regressió logística múltiple.

3.2.2.2. Algoritme de *Naive Bayes*

Els paràmetres obtinguts que mesuren l'actuació del model de *Naive Bayes* amb *Laplace*=0 es resumeixen en la taula nº 30. La corba ROC obtinguda es mostra a la Figura nº38.

Paràmetre	Valor	Interpretació
Precisió	0,688	Regular
Valor kappa	0,376	Dèbil
Sensibilitat	0,725	Regular
Especificitat	0,651	Regular
AUC	0,688	Regular

Taula nº 30. Paràmetres mesura actuació del model *Naïve Bayes* amb *Laplace* =0.

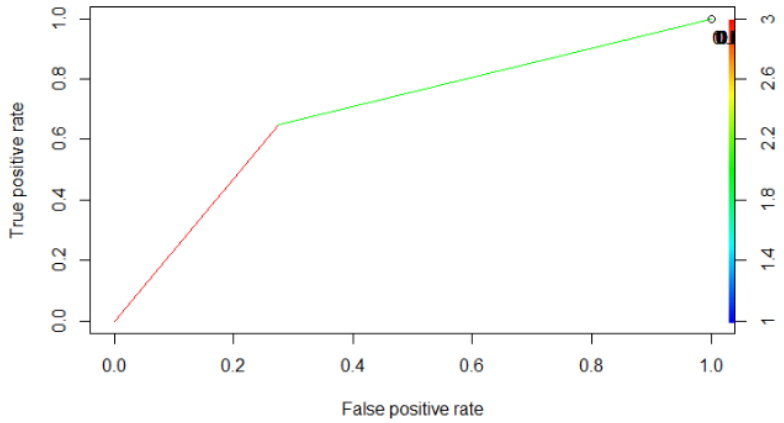


Figura nº38. Corba ROC del model *Naïve Bayes* amb *Laplace* =0.

3.2.2.3. Random forest

Els paràmetres obtinguts que mesuren l'actuació del model de *Random Forest* amb *ntrees*=500 es resumeixen en la taula nº 31. La corba ROC obtinguda es mostra a la Figura nº39.

Paràmetre	Valor	Interpretació
Precisió	0,788	Bona
Valor kappa	0,576	Moderada
Sensibilitat	0,797	Bona
Especificitat	0,779	Bona
AUC	0,787	Bona

Taula nº 31. Paràmetres mesura actuació de l'algoritme *Random Forest* amb *ntrees*=500.

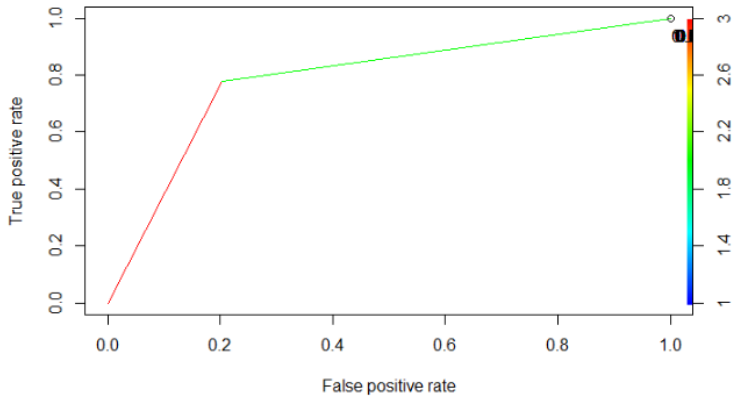


Figura nº39. Corba ROC de l'algoritme *Random Forest* amb *ntrees*=500.

3.2.2.4. Support Vector Machines

Els paràmetres obtinguts que mesuren l'actuació del model SVM lineal es resumeixen en la taula nº 32. La corba ROC obtinguda es mostra a la Figura nº40.

Paràmetre	Valor	Interpretació
Precisió	0,756	Bona
Valor kappa	0,511	Moderada
Sensibilitat	0,766	Bona
Especificitat	0,745	Regular
AUC	0,756	Bona

Taula nº 32. Paràmetres mesura actuació de l'algoritme SVM lineal.

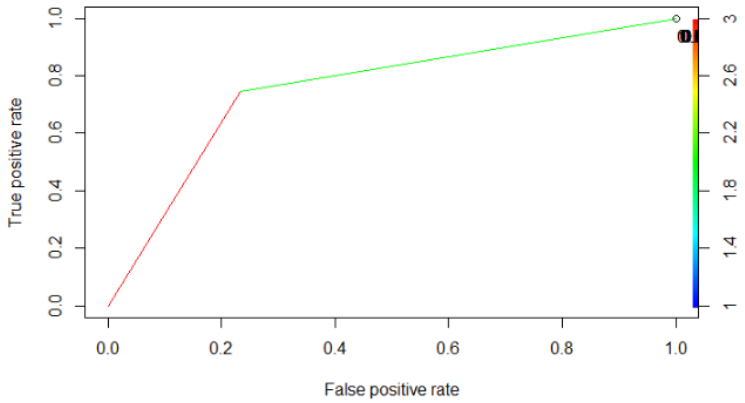


Figura nº40. Corba ROC de l'algoritme SVM lineal.

3.2.2.5. Artificial Neural Network

Els paràmetres obtinguts que mesuren l'actuació del model ANN amb una neurona a la capa oculta es resumeixen en la taula nº 33. La corba ROC obtinguda es mostra a la Figura nº 41.

Paràmetre	Valor	Interpretació
Precisió	0,761	Bona
Valor kappa	0,521	Moderada
Sensibilitat	0,781	Bona
Especificitat	0,740	Regular
AUC	0,760	Bona

Taula nº 33. Paràmetres mesura actuació de l'algoritme ANN amb un 1 node capa oculta.

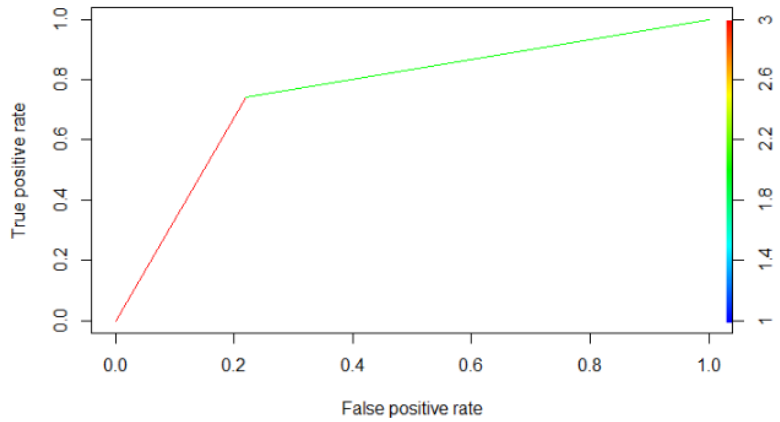


Figura nº41. Corba ROC de l'algoritme ANN amb 1 node a la capa oculta.

3.2.3. Millora del rendiment dels models

3.2.3.1. Regressió logística

Referent a la regressió logística, s'ha demostrat que el model que inclou totes les variables significatives és el que presenta un *Likelihood ratio* superior, pel que no hi ha cap model alternatiu amb menys variables predictores per intentar millorar el rendiment.

3.2.3.2. Naïve Bayes

En el cas de l'algoritme de *Naïve Bayes*, es pot intentar millorar el rendiment del model utilitzant *Laplace* =1:

```
model_NaiveBayesL1<-naiveBayes(AR_NB_train[2:10],AR_NB_train$RA, laplace = 1
```

Els paràmetres de mesura de l'actuació del model de *Naive Bayes* amb *Laplace*=1 es resumeixen en la taula nº 34. La corba ROC obtinguda es mostra a la Figura nº42.

Paràmetre	Valor	Interpretació
Precisió	0,688	Regular
Valor kappa	0,376	Dèbil
Sensibilitat	0,725	Regular
Especificitat	0,651	Regular
AUC	0,688	Regular

Taula nº 34. Paràmetres mesura actuació del model de *Naive Bayes* amb *Laplace*=1

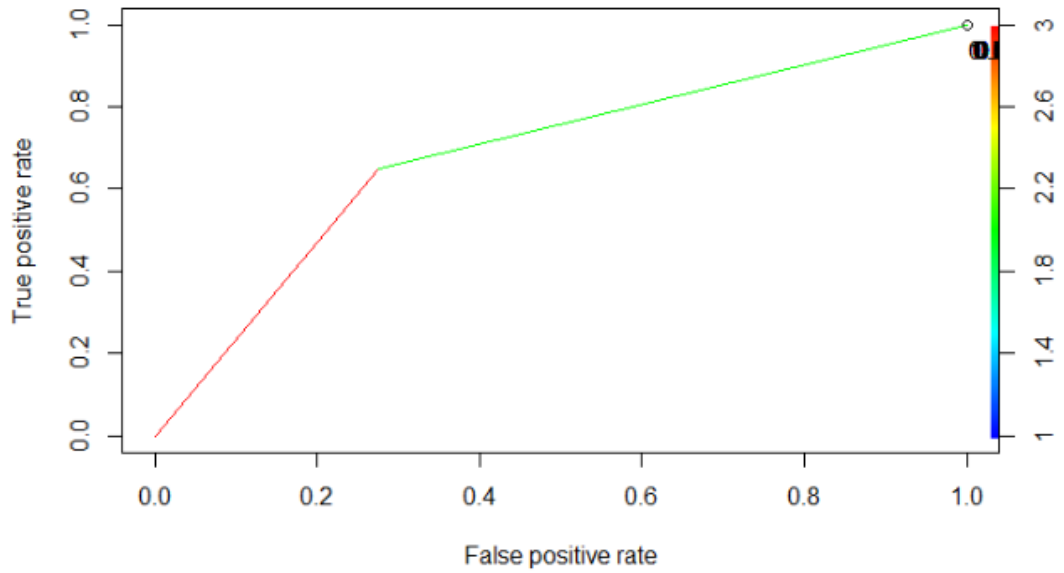


Figura nº42. Corba ROC del model de *Naïve Bayes* amb *Laplace* =1.

3.2.3.3. Random Forest

En el cas de l'algoritme de *Random Forest*, es pot intentar millorar el rendiment del model incrementant el número d'arbres a 1.000:

```
model_RF1000 <- randomForest(RA ~., data=AR_RF_train, ntree = 1000)
```

Els paràmetres de mesura de l'actuació del model de *Random Forest* amb *ntrees*=1000 es resumeixen en la taula nº 35. La corba ROC obtinguda es mostra a la Figura nº43.

Paràmetre	Valor	Interpretació
Precisió	0,789	Bona
Valor kappa	0,577	Moderada
Sensibilitat	0,797	Bona
Especificitat	0,780	Bona
AUC	0,788	Bona

Taula nº 35. Paràmetres mesura actuació del model *Random Forest* amb *ntrees*=1000.

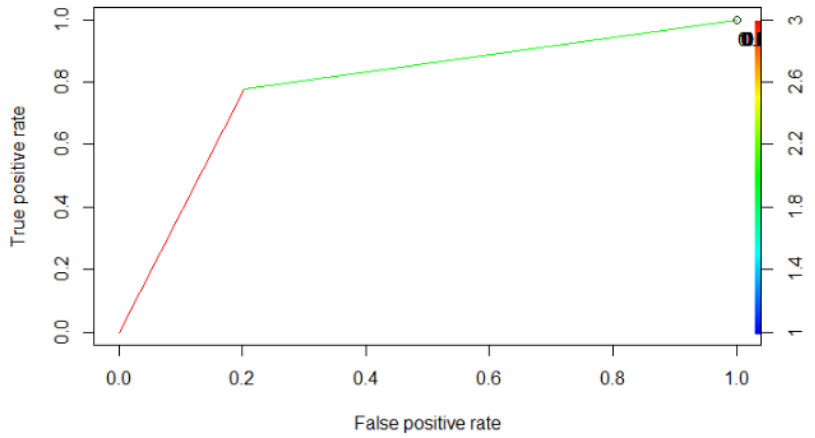


Figura nº43. Corba ROC del model de *Random Forest* amb *ntrees*=1000.

3.2.3.4. SVM

En el cas de l'algoritme SVM, es pot intentar millorar el rendiment del model utilitzant un SVM radial o gaussià:

```
model_SVM_gaussia <- ksvm(RA ~., data=AR_RF_train, kernel='rbfdot')
```

Els paràmetres que mesuren l'actuació del model SVM radial es resumeixen en la taula nº 36. La corba ROC obtinguda es mostra a la Figura nº44.

Paràmetre	Valor	Interpretació
Precisió	0,781	Bona
Valor kappa	0,561	Moderada
Sensibilitat	0,794	Bona
Especificitat	0,767	Bona
AUC	0,781	Bona

Taula nº 36. Paràmetres mesura actuació de l'algoritme SVM radial.

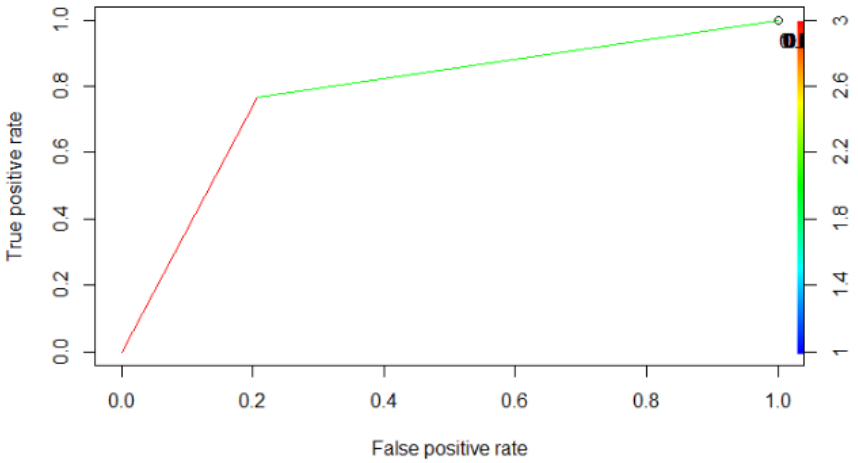
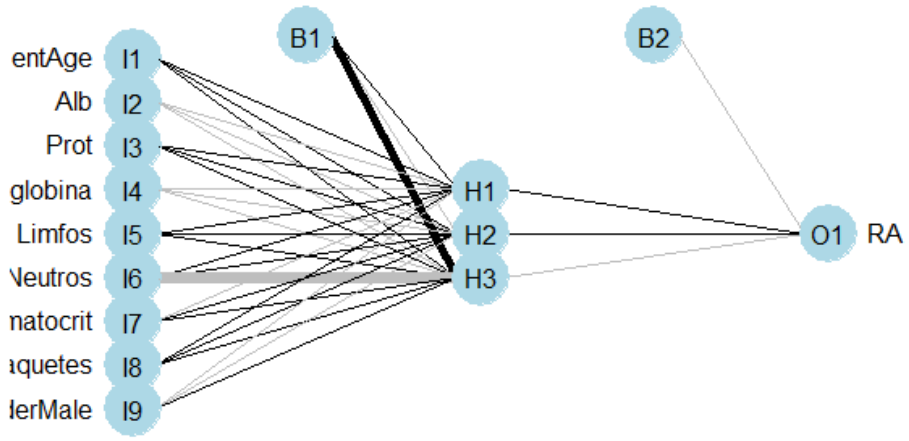


Figura nº44. Corba ROC de l'algoritme SVM lineal.

3.2.3.5. ANN

En el cas de l'algoritme ANN, es pot intentar millorar el rendiment del model augmentant el número de nodes a la capa oculta. Per això, es formulen i executen dos models alternatius, amb 3 i 5 nodes a la capa oculta, respectivament (Figura nº 45):

```
model_ANN_h3 <- train(RA ~ ., train.set, method='nnet',
                 trControl= trainControl(method='none'),
                 tuneGrid= data.frame(size=3,decay=0), tuneLength=1 ,
                 trace = FALSE)
```




```
model_ANN_h5 <- train(RA ~ ., train.set, method='nnet',
                 trControl= trainControl(method='none'),
                 tuneGrid= data.frame(size=5,decay=0), tuneLength=1 ,
                 trace = FALSE)
```

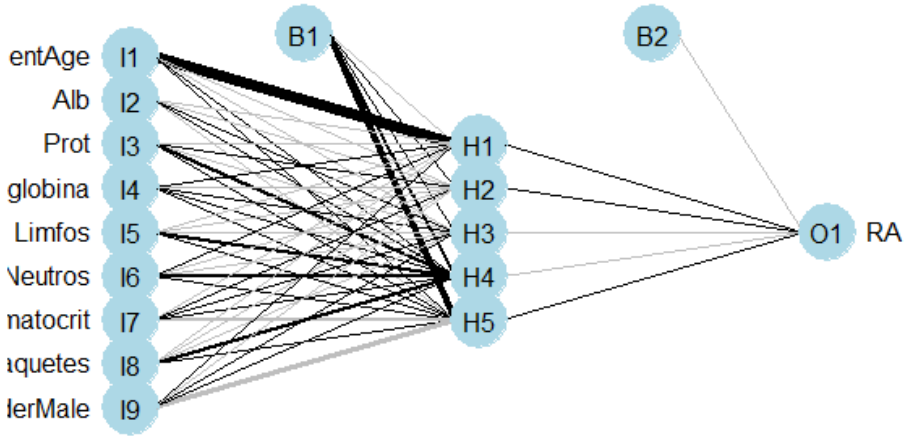


Figura nº45. Representació dels model ANN amb 3 i 5 nodes a la capa oculta.

Els paràmetres que mesuren l'actuació d'aquests dos models ANN es resumeixen a la taula nº 37. La corba ROC d'ambdós models es visualitza a la Figura nº 46.

Paràmetre	3 nodes capa oculta		5 nodes capa oculta	
	Valor	Interpretació	Valor	Interpretació
Precisió	0,768	Bona	0,768	Bona
Valor kappa	0,536	Moderada	0,536	Moderada
Sensibilitat	0,761	Bona	0,799	Bona
Especificitat	0,775	Bona	0,737	Regular
AUC	0,768	Bona	0,768	Bona

Taula nº 37. Paràmetres mesura actuació de l'algoritme ANN amb 3 i 5 nodes a la capa oculta

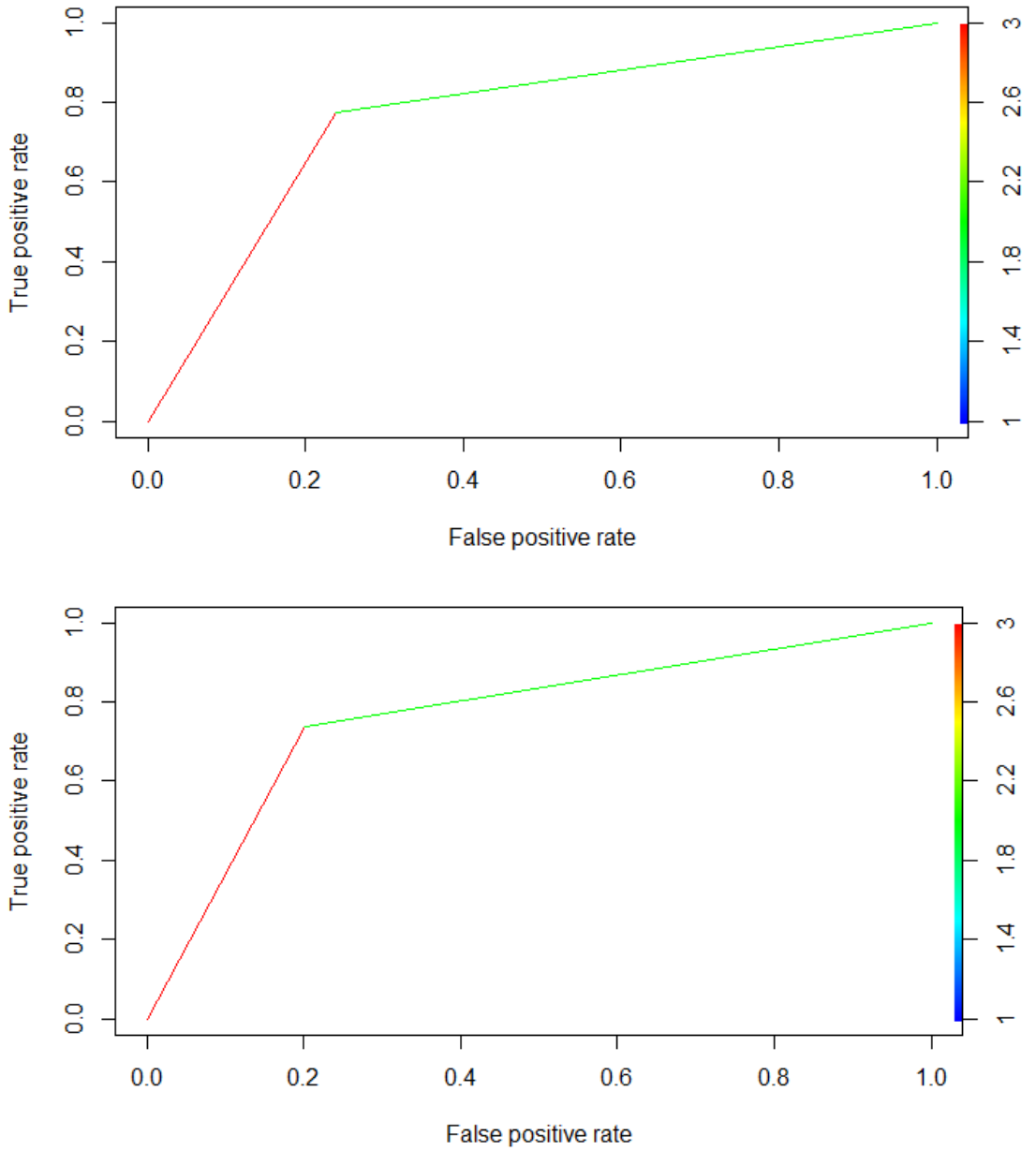


Figura nº46. Corbes ROC de l'algoritme ANN amb 3 i 5 nodes a la capa oculta, respectivament.

3.2.4. Validació dels models

Els mètodes de validació, també coneguts com a *resampling,* són estratègies que permeten estimar la capacitat predictiva dels models quan s'apliquen a noves observacions. La idea bàsica és que els models s'ajusten utilitzant un conjunt d'entrenament i s'avaluen amb la resta d'observacions. Aquest procés es repeteix varies vegades, de tal forma que el resultat final és una mitjana de totes les repeticions. D'aquesta manera, es compensen les possibles desviacions que sorgeixin del repartiment aleatori de les observacions.
Els mètodes més comuns per validar models de predicció són *K-Fold Cross Validation* i *Bootstrap* [17]:

- El *K-Fold Cross Validation* és un procés iteratiu. Consisteix en dividir les dades de forma aleatòria en k grups, a on k-1 grups s'utilitzen per entrenar el model i un dels grups s'utilitza com a validació. El procés es repeteix k vegades utilitzant un grup diferent com a validació en cada iteració. Per tant, hi ha k estimacions de l 'error, i la mitjana de totes aquestes s'utilitza com a estimació final [17] .
- *Bootstrap* també és un procés iteratiu, que genera una mostra (conjunt de dades entrenament) per mostreig aleatori amb reposició. Com a resultat, algunes observacions apareixen múltiples vegades a la mostra i d'altres, cap. Les observacions no seleccionades reben el nom de *out-of-bag (OOB)*. Per cada iteració es genera una nova mostra bootstrap que s'utilitza per entrenar el model, mentre que l'avaluació es realitza amb les observacions *OOB*. Igualment, es calcula la mitjana d'error de totes les iteracions [17].

Per validar i comparar els models anteriors s'utilitza el *Bootstrapping* per defecte (25 repeticions i 3 possibles *decay*) i *10-Fold Cross Validation*.
Totes les validacions es realitzen amb les funcions incloses al paquet *caret.*

3.2.4.1. Regressió logística

Els paràmetres obtinguts que mesuren l'actuació de la regressió logística múltiple amb *10-Fold Cross Validation* i *Bootstrapping* es resumeixen en la taula nº 38. Tal i com s'observa, ambdós mètodes donen els mateixos resultats.

Paràmetre	*10-Fold Cross Validation*		*Bootstrapping*	
	Valor	**Interpretació**	**Interpretació**	**Interpretació**
Precisió	0,750	Bona	0,750	Bona
Valor kappa	0,500	Moderada	0,500	Moderada
Sensibilitat	0,761	Bona	0,761	Bona
Especificitat	0,738	Regular	0,738	Regular

Taula nº 38. Mesures actuació del model de regressió logística múltiple validat.

3.2.4.2. Naïve Bayes

Els paràmetres obtinguts que mesuren l'actuació del model *Naïve Bayes* amb *10-Fold Cross Validation* i *Bootstrapping* es resumeixen en la taula nº 39. Tal i com s'observa, ambdós mètodes donen els mateixos resultats.

Paràmetre	*10-Fold Cross Validation*		*Bootstrapping*	
	Valor	**Interpretació**	**Interpretació**	**Interpretació**
Precisió	0,638	Regular	0,638	Regular
Valor kappa	0,277	Dèbil	0,277	Dèbil
Sensibilitat	0,659	Regular	0,659	Regular
Especificitat	0,618	Regular	0,618	Regular

Taula nº 39. Mesures actuació del model *Naïve Bayes* validat.

3.2.4.3. Random Forest

Els paràmetres obtinguts que mesuren l'actuació del model *Random Forest* amb *10-Fold Cross Validation* i *Bootstrapping* es resumeixen en la taula nº 40. Sembla que per *Booststrapping* s'aconsegueix una precisió lleugerament superior.

Paràmetre	*10-Fold Cross Validation*		*Bootstrapping*	
	Valor	**Interpretació**	**Interpretació**	**Interpretació**
Precisió	0,782	Bona	0,784	Bona
Valor kappa	0,564	Moderada	0,567	Moderada
Sensibilitat	0,784	Bona	0,788	Bona
Especificitat	0,780	Bona	0,778	Bona

Taula nº 40. Mesures actuació del model *Random Forest* validat.

3.2.4.4. SVM radial

Els paràmetres obtinguts que mesuren l'actuació del model SVM radial amb *10-Fold Cross Validation* i *Bootstrapping* es resumeixen en la taula nº 41. S'observa que la precisió és la mateixa per ambdós mètodes però que hi ha petites diferències amb la resta de paràmetres.

Paràmetre	*10-Fold Cross Validation*		*Bootstrapping*	
	Valor	**Interpretació**	**Interpretació**	**Interpretació**
Precisió	0,774	Bona	0,774	Bona
Valor kappa	0,549	Moderada	0,547	Moderada
Sensibilitat	0,783	Bona	0,779	Bona
Especificitat	0,765	Bona	0,768	Bona

Taula nº 41. Mesures actuació del model SVM validat.

3.2.4.5. ANN

Els paràmetres obtinguts que mesuren l'actuació del model ANN amb *10-Fold Cross Validation* i *Bootstrapping* es resumeixen en la taula nº 43. En la validació s'exploren 3 valors diferents de *size* i *decay*. Els valors dels paràmetres d'actuació són seleccionats en base a l'actuació més òptima, que en aquest cas es tracta del model que utilitza 5 nodes a la capa oculta i un *decay* de 0,1:

Size	Decay	Accuray	Kappa
1	0,00	0,7510	0,5020
1	0,0001	0,7511	0,5022
1	0,1	0,7415	0,4830
3	0,00	0,7510	0,5238
3	0,0001	0,7619	0,5308
3	0,1	0,7654	0,5216
5	0,00	0,7707	0,5415
5	0,0001	0,7708	0,5416
5	**0,1**	**0,7727**	**0,5454**

Taula nº 42. Precisió i valor kappa segons diferents valors de size i *decay*.

Paràmetre	*10-Fold Cross Validation*		*Bootstrapping*	
	Valor	**Interpretació**	**Interpretació**	**Interpretació**
Precisió	0,771	Bona	0,772	Bona
Valor kappa	0,541	Moderada	0,545	Moderada
Sensibilitat	0,776	Bona	0,790	Bona
Especificitat	0,764	Bona	0,755	Bona

Taula nº 43. Mesures actuació del model ANN validat.

Sembla que per *Booststrapping* s'aconsegueix una precisió lleugerament superior.

3.3. COMPARACIÓ DE L'ACTUACIÓ DELS MODELS

La taula nº 44 resumeix l'actuació de tots els models anteriors. Les validacions s'agrupen en una sola columna i es presenten com a "model validat", indicant els resultats del mètode amb millor rendiment.

Per altra banda, la taula també inclou el temps computacional dels diferents algoritmes provats, calculat a partir del paquet *tictoc* [14].

Paràmetre	Models predicció														
	Regressió logística		Naïve Bayes			Random Forest			SVM			ANN			
	Tots predictors	Validat	Laplace =0	Laplace =1	Validat	Ntrees = 500	Ntrees = 1000	Validat (Bootstrap)	Lineal	Radial	Validat	1 node	3 nodes	5 nodes	Validat (Bootstrap)
Precisió	0,755	0,750	0,688	0,688	0,638	0,788	0,789	0,782	0,756	0,781	0,774	0,761	0,768	0,768	0,772
Valor kappa	0,511	0,500	0,376	0,376	0,277	0,576	0,577	0,564	0,511	0,561	0,549	0,521	0,536	0,536	0,545
Sensibilitat	0,761	0,761	0,725	0,725	0,659	0,797	0,797	0,784	0,766	0,794	0,783	0,781	0,761	0,799	0,790
Especificitat	0,750	0,738	0,651	0,651	0,618	0,779	0,780	0,780	0,745	0,767	0,765	0,740	0,775	0,737	0,755
AUC	0,827	*NA*	0,688	0,688	*NA*	0,787	0,788	*NA*	0,756	0,781	*NA*	0,760	0,768	0,768	*NA*
Temps computacional (seg)	**Tots predictors**	**Validat**	**Laplace =0**	**Laplace =1**	**Validat**	**Ntrees = 500**	**Ntrees = 1000**	**Validat (Bootstrap)**	**Lineal**	**Radial**	**Validat**	**1 node**	**3 nodes**	**5 nodes**	**Validat (Bootstrap)**
Entrenament	0,23	1,73	0,03	0,03	27,19	26,51	27,57	997,43	14,94	35,23	686,97	1,34	2,49	2,42	368,13
Predicció	0,01	0,05	1,59	1,71	2,45	0,67	1,32	0,86	0,36	6,25	5,54	0,06	0,07	0,05	0,06
Global	0,24	1,80	1,62	1,74	29,68	27,18	28,91	998,31	15,33	41,5	692,52	1,40	2,61	2,57	368,19

Taula nº 44. Mesures actuació i temps computacional de tots els models de predicció.

3.3.1. Selecció del millor model de predicció

Referent a la regressió logística múltiple, està descrit que el millor model a utilitzar és el que obté un *likelihood ratio* inferior. En aquest cas, es tracta del model que utilitza totes les variables seleccionades a la primera part del treball. S'observa que els valors dels paràmetres que mesuren l'actuació del model són una mica inferiors quan es realitza la validació.

En el cas de l'algoritme de *Naïve Bayes*, tant si s'utilitza *Laplace=0* com *Laplace=1*, els valors obtinguts dels paràmetres avaluats són els mateixos, pel que la validació es realitza amb el model *Laplace*=0.

Per altre banda, amb el model de *Random Forest* tampoc es visualitzen millores quan s'utilitza un nombre d'arbres més elevat. Per això mateix, la validació es duu a terme utilitzant el primer model entrenat.

SVM presenta una actuació molt semblant a la regressió logística múltiple. Malgrat això, es visualitzen millors resultats amb el SVM radial, pel que la validació es realitza seguint aquesta fórmula.

Finalment, respecte el model ANN , els models amb tres i cinc nodes a la capa oculta són els que presenten millor actuació. La diferència entre ambdós rau en que el model amb 3 nodes a la capa oculta presenta millor especificitat, mentre que el model amb 5 nodes a la capa oculta presenta millor sensibilitat. Segons la validació, el millor model és el que té 5 nodes a la capa oculta i 0,1 de *decay*.

Les actuacions de tots els models per la classificació de risc en dos categories es consideren bones en excepció de l'algoritme de *Naïve Bayes*, que es considera regular.

Si s'escull la precisió del model com a paràmetre primari per determinar l'actuació dels models, es conclou que el model amb millor rendiment és ***Random Forest***, indiferentment del nombre d'arbres utilitzats. El segueixen: SVM radial > ANN amb 5 i 3 nodes capa oculta > ANN amb 1 node capa oculta > SVM lineal > regressió logística múltiple> *Naïve Bayes.* S'obtenen els mateixos resultats si s'utilitza el valor kappa com a mesura primària.

En canvi, si s'utilitza l'AUC per mesurar el rendiment dels models, el model amb millor actuació és la **regressió logística múltiple**, seguit per *Random Forest*> SVM radial > ANN amb 5 i 3 nodes a la capa oculta > ANN amb 1 node capa oculta > SVM lineal >*Naïve Bayes*.

A nivell de temps computacional, s'observa que els models més lents són *Random Forest* i SVM radial, que tarden uns 28 i 41 segons, respectivament. No obstant això, es considera que tots són bastants àgils. Per altra banda, les validacions comporten un major temps computacional, degut a les diferents iteracions i repeticions. Per exemple, la validació del model *Random Forest* comporta esperar uns 15 minuts per a obtenir els resultats, mentre que les validacions de SVM i ANN comporten 12 i 6 minuts respectivament.

3.4. INFORME REPRODUÏBLE I APLICACIÓ *SHINY*

3.4.1. Informe reproduïble

L'Annex nº3 és un informe reproduïble en *PDF* que conté els resultats de l'anàlisi estadístic generats amb *Rmarkdown*. Igualment, també s'adjunta el codi utilitzat (Annex nº2).

3.4.2. Aplicacions Shiny

Shiny és un paquet d'R que facilita la construcció d'aplicacions web interactives [20]. Permet la manipulació de dades sense manipulació del codi R.

La taula nº 45 resumeix alguns dels avantatges i inconvenients de les aplicacions *Shiny:*

Avantatges	Inconvenients
* Temps de resposta ràpid i eficient.	* Requereix de les actualitzacions oportunes, ja que les funcions queden obsoletes.
* No requereix coneixement de programació com *HTML, JavaScript o CSS*.	* S'ha de guardar amb *l'encoding* adient per no tenir problemes amb caràcters especials.
* Automatització completa de l'aplicació.	* S'ha de tenir una bona base de programació amb R i conèixer bé l'eina.
* Codi obert i gratuït.	

Taula nº45. Avantatges i inconvenients d'una aplicació *Shiny.*

Les aplicacions *Shiny* tenen dos components principals[20]:

1) Seqüència d'ordres d'interfície d'usuari (*ui.R*): controlar el disseny i l'aspecte de l'aplicació. Proporciona la interactivitat de l'aplicació, ja que és on es defineix les possibles entrades per l'usuari i a quines sortides s'associarà.
   S'utilitza una funció anomenada *fluidPage* per crear una pantalla que s'ajusta automàticament a les dimensions de la finestra del navegador. La ui.R bàsicament consisteix en col·locar elements dins d'aquesta funció. Alguns dels elements per crear un disseny bàsic són:

   - *titlePanel:* títol de cada apartat
   - *sideBarLayout /selectLayout*: creació barra lateral amb rang de valors / creació llistat d'opcions a escollir.

2) Seqüencia d'ordres servidor (*server.R*): proporciona les instruccions que es necessiten per construir l'aplicació, és a dir, tots els passos per a convertir l'entrada de dades donada per l'usuari en la sortida desitjada.

Els dos components es comuniquen entre ells, ja que *ui.R* transmet els paràmetres a *server.R* i aquest, a la vegada, li transmet els resultats.

3.4.3. Utilització en el projecte

En aquest cas, es crea una aplicació *Shiny* molt senzilla que permet fer una exploració gràfica ràpida de la base de dades utilitzada per la construcció dels models de predicció (Secció 4.2.1). La generació del codi es basa en exemples semblants trobats a *Github i Rpubs* [18,19]. Aquest codi es troba com a Annex nº4.

En aquest cas, l'aplicació *Shiny* disposa de 3 pestanyes:

- Pestanya 1: *Boxplot* → Gràfic de bigotis de la variable numèrica seleccionada segons diagnòstic o gènere (Figura nº 47) .
- Pestanya 2: Histograma → Gràfic de densitat de la variables numèrica seleccionada. (Figura nº 48) .
- Pestanya 3: Taula de dades → Visualització de dades (Figura nº 49).

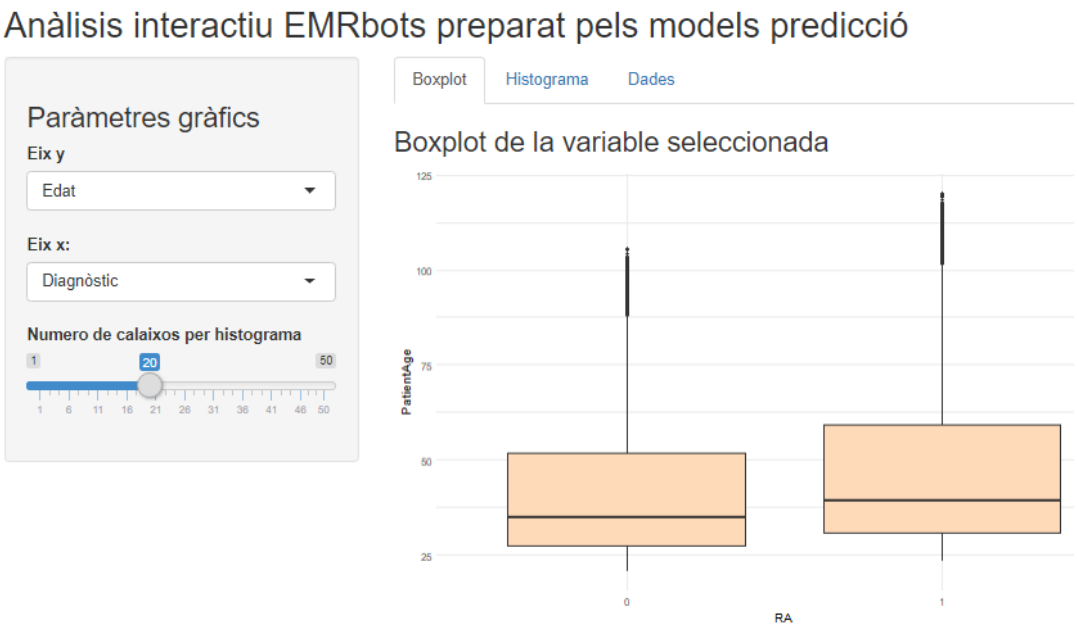


Figura nº 47. Pestanya 1 de l'aplicació Shiny creada.

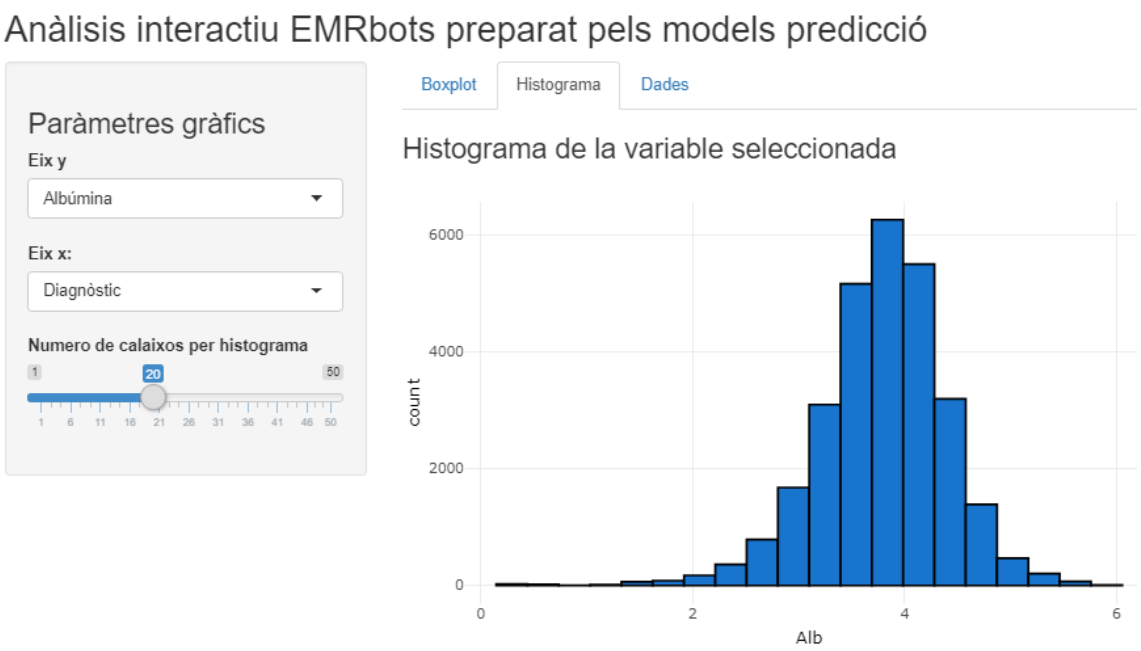
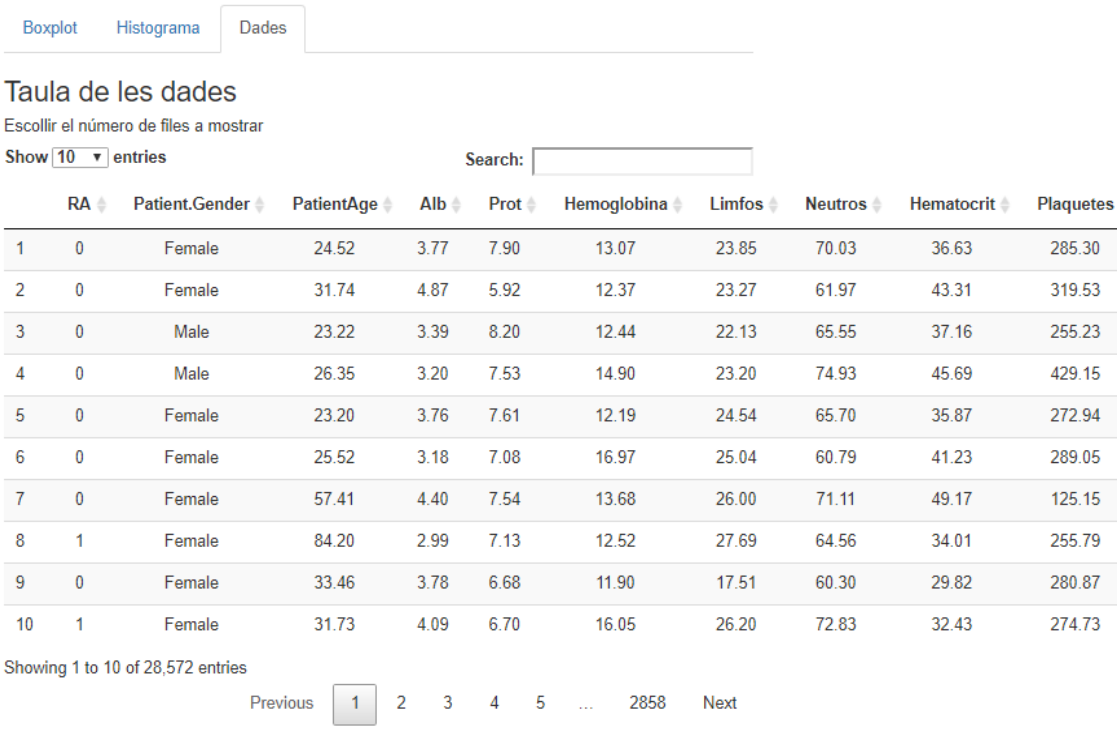


Figura nº 48. Pestanya 2 de l'aplicació Shiny creada .



	RA	Patient.Gender	PatientAge	Alb	Prot	Hemoglobina	Limfos	Neutros	Hematocrit	Plaquetes
1	0	Female	24.52	3.77	7.90	13.07	23.85	70.03	36.63	285.30
2	0	Female	31.74	4.87	5.92	12.37	23.27	61.97	43.31	319.53
3	0	Male	23.22	3.39	8.20	12.44	22.13	65.55	37.16	255.23
4	0	Male	26.35	3.20	7.53	14.90	23.20	74.93	45.69	429.15
5	0	Female	23.20	3.76	7.61	12.19	24.54	65.70	35.87	272.94
6	0	Female	25.52	3.18	7.08	16.97	25.04	60.79	41.23	289.05
7	0	Female	57.41	4.40	7.54	13.68	26.00	71.11	49.17	125.15
8	1	Female	84.20	2.99	7.13	12.52	27.69	64.56	34.01	255.79
9	0	Female	33.46	3.78	6.68	11.90	17.51	60.30	29.82	280.87
10	1	Female	31.73	4.09	6.70	16.05	26.20	72.83	32.43	274.73

Figura nº 49. Pestanya 3 de l'aplicació Shiny creada.

4. CONCLUSIONS

Els objectius d'aquest TFM consistien en entrar en el món dels EMR, aprendre a tractar les dades que se'n deriven i simular un estudi de predicció de risc. Degut a la confidencialitat de les històries clíniques dels pacients reals, s'ha treballat amb dades simulades (*EMRbot*s), pel que no s'ha pogut lidiar amb totes les problemàtiques derivades dels EMR reals. Per altra banda, els resultats obtinguts no són extrapolables als pacients reals. Tot i així, sí que ha servit per introduir-se en aquesta àrea i tenir una idea del que comporta la manipulació d'aquest tipus de dades.

Els EMR solen incorporar una gran quantitat d'observacions i variables. Per això mateix, s'ha escollit una base de dades amb un elevat número de pacients simulats. Tal quantitat de dades ha suposat haver de buscar alternatives a l'Excel tradicional per poder dur a terme el filtratge i la selecció de variables i observacions, amb el hàndicap de que les dades es trobaven en format.*txt*. Conseqüentment, s'ha utilitzat *Power Query*. Malgrat l'ajuda d'aquest complement, aquest pas ha requerit d'una gran dedicació de temps i recursos computacionals, pel que, en cas de disposar d'un altre tipus de dades, es recomana buscar alternatives per agilitzar el procés.

El fet de treballar amb dades simulades ha implicat que les dades fossin bastant completes (poca *missing data*) i que les variables estiguessin codificades i estandarditzades a les escales internacionals, pel que l'adequació de les dades pel posterior anàlisi no ha suposat un gran esforç. No obstant, el fet d'utilitzar dades simulades a partir d'algoritmes aleatòris, ha comportat que no es trobessin diferències en la distribució de variables entre els dos grups de pacients, pel que s'han hagut de realitzar una serie de modificacions. Un cop dut a terme l'anàlisi estadístic posterior a aquestes transformacions, s'ha identificat que els següents factors poden estar associats a l'aparició (o diagnòstic) d'Artritis Reumatoide: Hipoalbuminèmia, proteïnèmia, anèmia, leucocitosis, hiperplaquetosis, gènere femení i una edat a partir dels 45 anys. Tot i així, es reitera el fet que els resultats obtinguts no són aplicables a nivell real, ja que han estat totalment manipulats.

L'estudi de predicció de risc, tal i com es menciona al llarg del treball, s'ha tractat d'un problema de classificació en dos categories. D'aquesta forma, s'han seleccionat els algoritmes i models més adients per aquest tipus d'anàlisi, prèvia recerca bibliografia: Regressió logística múltiple, algoritme de *Naïve Bayes*, *Random Forest*, SVM i ANN. Els models s'han construït a partir dels factors mencionats al paràgraf anterior i la seva actuació s'ha estimat a partir de diferents paràmetres, tals i com són les corbes ROC, la precisió o l'AUC. Igualment, tots ells han sigut validats i testats per robustesa. Els resultats obtinguts indiquen que el model de *Random Forest* és el que ha funcionat millor, tot i que s'ha observat que és més lent que la resta en termes computacionals. No obstant, la seva actuació no ha sigut destacable per sobre d'altres models com SVM radial, ANN o regressió logística múltiple, pel que qualsevol d'ells es considera vàlid per problemes de classificació de malaltia a partir de predictors numèrics i categòrics. L'únic model que ha presentat una actuació significativament inferior que la resta ha sigut l'algoritme de *Naïve Bayes*, pel que no es recomanaria utilitzar-lo en estudis d'índole similar.

Referent a les habilitats obtingudes amb el TFM, he pogut desenvolupar i aprofundir en molts camps de la bioestadística que havia treballat de forma més superficial al llarg del Màster de Bioinformàtica i Bioestadística de la UOC. Es podria dir que he pogut completar la majoria d'objectius i tasques plantejades en un inici, tot i que els terminis marcats han suposat una limitació en el desenvolupament d'algunes d'elles.

El fet de treballar amb una quantitat tant gran de dades m'ha fet adonar de la importància del processos previs als anàlisis estadístics, els quals a vegades no se'ls reconeix la importància que tenen. Un bon tractament i adequació de les dades facilita molt tots els passos posteriors.

La limitació principal al llarg del treball ha sigut l'impossibilitat de poder treballar amb dades reals i haver de manipular les dades simulades per tal d'obtenir resultats. Això comporta no es puguin utilitzar en anàlisis futurs ni es puguin considerar rellevants dins el món científic-bioestadístic. Tot i així, a nivell personal m'ha aportat destresa i experiència en aquests tipus d'anàlisis.

5. GLOSSARI

ALK: Alkaline phosphatase

ALT: Alanine transaminase

AST: Aspartate aminotransferase

ANN: Artificial Neural Network

AR: Artritis Reumatoide

AUC: Area under the curve

BUN: Blood Urea Nitrogen

DE: Desviació estàndard

EMR /EHR: Electronic Medical Record/Electronic Health Record

GLM: General Linear Model

MCAR: Missing completely at random

OOB: Out of bag

ROC: Receiver operating characteristic curve

SVM: Support Vector Machine

6. BIBLIOGRAFIA

Publicacions, tesis:

[1]Armstrong, C., Swarbrick, C. M., Pye, S. R., & O'Neill, T. W. (2005). **Occurrence and risk factors for falls in rheumatoid arthritis.** *Annals of the Rheumatic Diseases,* 64(11), 1602 LP – 1604. https://doi.org/10.1136/ard.2004.031195.

[2]Carroll, R. J., Thompson, W. K., Eyler, A. E., Mandelin, A. M., Cai, T., Zink, R. M., Pacheco, J. A., Boomershine, C. S., Lasko, T. A., Xu, H., Karlson, E. W., Perez, R. G., Gainer, V. S., Murphy, S. N., Ruderman, E. M., Pope, R. M., Plenge, R. M., Kho, A. N., Liao, K. P., & Denny, J. C. (2012). **Portability of an algorithm to identify rheumatoid arthritis in electronic health records.** *Journal of the American Medical Informatics Association*, *19*(e1), e162–e169. https://doi.org/10.1136/amiajnl-2011-000583.

[3]Fernandez A.,Llorente M.J (2012). **Diagnóstico y seguimiento de la artritis reumatoide**. Ed Cont Lab Clín; 16: 70 – 81.

[4]Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). **Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review**. *Journal of the American Medical Informatics Association*, *24*(1), 198–208. https://doi.org/10.1093/jamia/ocw042.

[5]Glicksberg, B. S., Oskotsky, B., Giangreco, N., Thangaraj, P. M., Rudrapatna, V., Datta, D., Frazier, R., Lee, N., Larsen, R., Tatonetti, N. P., & Butte, A. J. (2019). **ROMOP: a light-weight R package for interfacing with OMOP-formatted electronic health record data**. *JAMIA Open*, *2*(1), 10–14. https://doi.org/10.1093/jamiaopen/ooy059

[6]Gutiérrez-Sacristán, A., Bravo, À., Giannoula, A., Mayer, M. A., Sanz, F., & Furlong, L. I. (2018). **comoRbidity: an R package for the systematic analysis of disease comorbidities**. Bioinformatics (Oxford, England), 34(18), 3228–3230. https://doi.org/10.1093/bioinformatics/bty315

[7]Himes, B. E., Dai, Y., Kohane, I. S., Weiss, S. T., & Ramoni, M. F. (2009). **Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records.** Journal of the American Medical Informatics Association : JAMIA, 16(3), 371–379. https://doi.org/10.1197/jamia.M2846

[8]Kartoun, U. (2019). **Advancing informatics with electronic medical records bots (EMRBots).** *Software Impacts*, *2*, 100006. https://doi.org/https://doi.org/10.1016/j.simpa.2019.100006.

[9]Kohane, I. S., Aronow, B. J., Avillach, P., Beaulieu-Jones, B. K., Bellazzi, R., Bradford, R. L., Brat, G. A., Cannataro, M., Cimino, J. J., García-Barrio, N., Gehlenborg, N., Ghassemi, M., Gutiérrez-Sacristán, A., Hanauer, D. A., Holmes, J. H., Hong, C., Klann, J. G., Loh, N. H. W., Luo, Y., … Cai, T. (2021). **What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask**. *J Med Internet Res*, *23*(3), e22219. https://doi.org/10.2196/22219

[10]Lizares, M. **Comparación de modelos de clasificación: Regresión logística y arboles de clasificación para evaluar el rendimiento académico**. Setiembre 2017. (PDF versió: https://core.ac.uk/download/pdf/323352959.pdf)

[11]Myasoedova, E., Davis, J. M. 3rd, Crowson, C. S., & Gabriel, S. E. (2010). **Epidemiology of rheumatoid arthritis: rheumatoid arthritis and mortality.** *Current Rheumatology Reports*, *12*(5), 379–385. https://doi.org/10.1007/s11926-010-0117-y

[12]Nash, A., Chang, T., Wan, B., & Cader, M. (2021). **rdrugtrajectory: An R Package for the Analysis of Drug Prescriptions in Electronic Health Care Records.** https://doi.org/10.1101/2021.01.08.425952

[13]Springate, D. A., Parisi, R., Olier, I., Reeves, D., & Kontopantelis, E. (2017). **rEHR: An R package for manipulating and analysing Electronic Health Record data.** *PloS One*, *12*(2), e0171784–e0171784. https://doi.org/10.1371/journal.pone.0171784

Pàgines web, Rpubs, Github:

[14] Alexej's blog. **5 ways to measure running time of R code**. Maig 2017. https://www.r-bloggers.com/2017/05/5-ways-to-measure

[15] Alice M. **Imputing Missing data with R; MICE package**. Octubre 2015 (Actualització Maig 2018). https://datascienceplus.com/imputing-missing-data-with-r-mice-package/

[16]Amat, J. **Regresión logística simple y múltiple.** Agost 2016. https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.html

[17]Amat, J. **Validación de modelos predictives: Cross-validation, OneLeaveOut, Bootstraping.** Novembre 2016 (Actualització Novembre 2020). https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap

[18]Aumatre, A. A Shiny app for visualizing Eurostat Statistics in Income and Living Conditions. Juny 2019. https://github.com/aaumaitre/eurostat

[19]Jozefek, P. (2020). **ShinyApp HSB2 Code**. https://rpubs.com/pjozefek/573335

[20] Shiny from *Rstudio*. https://shiny.*Rstudio*.com/

7. ANNEXOS

Els Annexos llistats a continuació s'adjunten com a complement de la memòria:

- Annex 1: Excel amb la base de dades *EMRbot* preparada per l'anàlisi.
- Annex 2: Codi R utilitzat per generar l'informe reproduïble en format *Rmarkdown*.
- Annex 3: Informe estadístic reproduïble en *PDF.*
- Annex 4: Codi R utilitzat per la generació de l'aplicació *Shiny*.