FACULTAT
DE CIÈNCIES
I TECNOLOGIA
UVIC | UVIC·UCC

Treball de Fi de Grau Experimental

IMPLEMENTACIÓ D'UN MODEL DE MACHINE LEARNING PER LA PREDICCIÓ DEL RECEPTOR D'ESTROGEN EN EL CÀNCER DE MAMA

MARINA VILARDELL LLADÓ

Grau en Biotecnologia

Tutor/a: Maria Dolors Anton Solà i Lara Nonell Mazelon

Vic, Juny de 2022

Resum

Títol: Implementació d'un model de *Machine Learning* per la predicció del receptor d'estrogen en el càncer de mama.
Autora: Marina Vilardell Lladó
Co-Tutores: Maria Dolors Anton Solà i Lara Nonell Mazelon
Data: Juny de 2022
Paraules clau: *Machine Learning, XGBoost, càncer de mama, receptor d'estrogen (RE)*

El càncer de mama és la neoplàsia maligna més freqüent entre les dones, i ocupa la primera posició pel que fa a mortalitat respecte als altres tipus de càncers. Per reduir els índexs de mortalitat, és summament important diagnosticar el càncer amb el menor temps possible. Una de les tendències creixents en els últims anys per ajudar en la diagnosi de malalties, és la intel·ligència artificial.

En aquest treball, es desenvoluparà un algoritme supervisat d'aprenentatge automàtic amb XGBoost, amb el propòsit de predir l'estatus del receptor d'estrogen, implicat en la classificació molecular del càncer de mama. Per avaluar l'eficàcia del model, es comparen els resultats amb els generats a través del model de regressió logística clàssica. S'obté un valor F, un dels mètodes estadístics utilitzats per determinar com de precís el model és, del 97% en el model de *Machine Learning* amb XGBoost, i un valor del 54% pel model clàssic. Així doncs, es conclou que el model d'aprenentatge automàtic generat amb XGBoost obté uns millors resultats a l'hora de classificar el receptor d'estrogen, i que és una eina potencial per generar prediccions, en la qual s'ha de continuar investigant i optimitzant, per a què pugui arribar a ser utilitzada no només en el camp de la medicina de diagnosi, sinó que també en altres àrees mèdiques.

Summary

Title: Implementation of a Machine Learning model for predicting estrogen receptor status in breast cancer.
Author: Marina Vilardell Lladó
Supervisor: Maria Dolors Anton Solà i Lara Nonell Mazelon
Date: June 2022
Keywords: Machine Learning, XGBoost, breast cancer, estrogen receptor (ES)

Breast cancer (BC) is the most common malignant disease diagnosed in women. Amongst all the different types of cancer, it is considered one of the leading causes of death worldwide in females. Mortality rates are estimated to rise in the near future, hence the importance of an early diagnosis to reduce them. Artificial intelligence (AI), which is growing rapidly in the medical field, has emerged as a powerful tool for assisting diagnosis in healthcare.

This study aims to develop a machine learning supervised algorithm through the implementation of XGBoost for predicting estrogen receptor status, which is implied in breast cancer molecular identification. To evaluate the efficiency of this model, results are compared with the ones obtained through a classic logistic regression model. The F value, a statistical method to determine the accuracy of the model, is 97% for the machine learning algorithm, in contrast to the 54% obtained from the classical one. It is concluded that the XGBoost algorithm has the capacity to predict estrogen receptor status with a high correct rate, thus being a formidable tool to make predictions. Despite this fact, more research is needed to be focused on this field to optimize the algorithms, so that it can be used due to their benefits, not only in diagnostic medicine but also in other medical areas.

Índex de Continguts

Llista de Taules

Llista de Figures

1. Introducció

1.1 El càncer

El càncer és un conjunt de nombroses malalties caracteritzades per una divisió cel·lular anormal i descontrolada. Aquest creixement descontrolat arriba a formar una massa anomenada neoplàsia, i que a través dels vasos sanguinis i limfàtics, les cèl·lules es poden expandir a altres òrgans o teixits, procés anomenat metàstasis[1].

El càncer és originat per mutacions genètiques que produeixen una desregulació en l'expressió de determinats gens que controlen el funcionament del cicle cel·lular de les cèl·lules. El fet pel qual s'ocasionen les mutacions es deu a l'exposició de substàncies perjudicials del medi ambient i d'elements nutricionals, a errors no reparats durant la replicació del DNA, i a formes heretades. Per tal que les cèl·lules esdevinguin canceroses, les mutacions han d'afectar a tres tipus de gens [2].

- Oncogens: Gens implicats positivament en la regulació del cicle cel·lular. Un sol al·lel del gen ha de ser mutat per què aquest esdevingui responsable de la transformació de les cèl·lules canceroses.
- Gens supressors de tumors: Gens implicats en la regulació negativa del cicle cel·lular. Si es produeix una mutació en els dos al·lels del gen, aquest estarà inactiu, facilitant així el procés de divisió cel·lular[3].
- Gens de reparació del DNA: gens en la qual la seva funció es basa a reparar errors en el DNA. Si aquests gens estan mutats, no podran efectuar la seva funció correctament, pel qual no podran arreglar mutacions en altres gens, i més probabilitat que les cèl·lules esdevinguin tumorals.

1.1.1 Les xifres del càncer al món

Fa pocs anys, l'Organització Mundial de la Salut (OMS) anunciava que el càncer es trobava en la llista de les 10 principals causes de mortalitat en tot el món, lluny encara de la primera posició, que s'atribuïa a les malalties cardiovasculars [4]. Amb el pas del temps, el càncer ha anat i es preveu que vagi escalant posicions, fins a convertir-se una de les primeres causes de mortalitat mundial. Durant l'any 2020, es varen arribar a detectar més de 19 milions de casos a tot el món, i més de la meitat resultant en mort (9,96 milions) [5]. Els tumors diagnosticats més freqüentment varen ser el de mama, pulmó, còlon, pròstata i estómac, tots ells amb més d'un milió de casos (Figura 1). Les estadístiques recents de l'OMS apunten que l'any 2040 la incidència del càncer augmentarà més d'un 50%, amb la diagnosi de més de 30 milions de casos nous i encara que cada vegada es promoguin més campanyes de prevenció, la mortalitat es veurà molt augmentada, arribant fins a una xifra de 16,3 milions de defuncions.

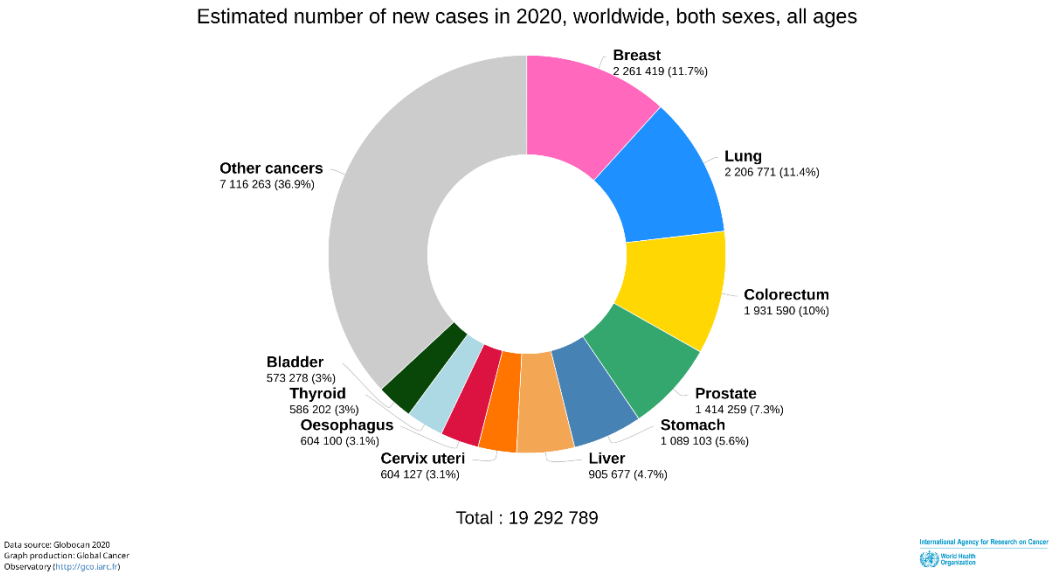


Figura 1: Estimació del nombre de nous casos de càncer a nivell mundial durant l'any 2020.

El nombre de casos diagnosticats ha experimentat un notable creixement en els últims anys, a causa de l'augment de la població i les tècniques de diagnòstic precoç, ja que en l'àmbit estatal durant l'any 2020 es van diagnosticar aproximadament 282 mil casos de càncer, dels quals 164 mil varen ser detectats en homes i 119 mil en dones [6]. Amb aquest nivell de creixement, les estadístiques de l'OMS en preveuen fins a 375 mil per l'any 2040. Encara que la incidència del càncer ha augmentat, en general, el risc de mortalitat s'ha vist disminuït amb el pas del temps gràcies a les innovadores teràpies i tractaments. Tot i això, cal continuar invertint recursos en investigació que donin lloc a noves estratègies que permetin reduir els índexs de mortalitat.

Separant per sexes, a Espanya, el tipus de càncer més freqüent en homes és el de pròstata (34.613 casos), seguit pel de còlon (24.610 casos) i pulmó (21.480). En dones, el càncer de mama encapçala la llista (34.088 casos), seguit de lluny pel de còlon (15.831 casos) i pulmó (7.708 casos)[6].

1.1.2 Càncer de mama

El càncer de mama és la neoplàsia maligna més freqüent entre les dones. Durant l'any 2020 es van arribar a diagnosticar més de 2,2 milions de casos arreu del món i s'estima un augment del 40% per l'any 2040, 3,19 milions de casos[5]. Aquest tipus de càncer ocupa també, la primera posició quant a mortalitat en dones respecte als altres tipus de càncers, arribant a una xifra aproximadament de 685 mil defuncions. S'estima, per l'any 2040, més d'un milió de morts. Aquesta situació també es veu i es veurà reflectida a Espanya, que durant l'any 2020 es varen diagnosticar 34.088 casos de càncer de mama, del qual se'n preveu un lleuger augment durant les dues pròximes dècades. Es varen registrar més de 6,5 mil defuncions i se n'estimen més de 8 mil per l'any 2040, segons les estadístiques de l'OMS. A la figura 2 es pot observar l'evolució de la incidència i mortalitat dels casos de càncer de mama a Espanya.

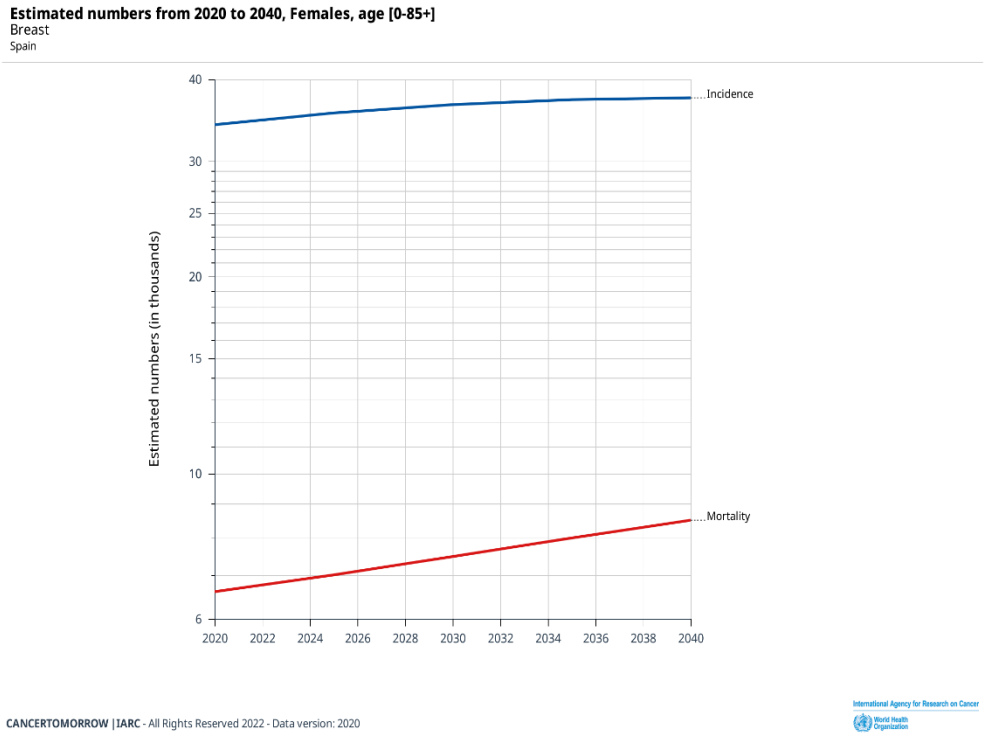


Figura 2: Evolució de la incidència i mortalitat del càncer de mama a Espanya, des de l'any 2020 i fins al 2040

1.1.3 Immunohistoquímica

Els progressos en la investigació del càncer de mama i la biologia molecular des de finals del segle XX han permès definir el càncer com una malaltia heterogènia, mostrant diferents tipologies de tumors, cadascuna amb unes característiques biològiques especifiques, una resposta al tractament i un pronòstic diferent [7], [8]. Dur a terme una anàlisi molecular a través del microscopi per identificar les característiques de cada tumor és un procés crític per

determinar el subtipus de càncer que pateix un pacient, i per obtenir informació sobre quin tipus de tractament pot ser el més adequat i eficaç.

Tradicionalment, el mètode més freqüent emprat per definir el subtipus molecular de càncer de mama, és la immunohistoquímica (IHC). Aquesta prova es realitza per determinar si les cèl·lules tumorals presenten uns marcadors cel·lulars concrets, els quals són utilitzats amb la finalitat de classificar el tumor i avaluar el pronòstic. En el càncer de mama, els principis de la immunohistoquímica es basen en, mitjançant anticossos marcats, mesurar l'expressió proteica dels receptors hormonals, estrogen (ER) i progesterona (PR), i de l'oncogen HER2, per establir una classificació tumoral que defineixi el pronòstic i les opcions terapèutiques més eficaces per a cada subtipus [9]. Per tant, la immunohistoquímica té un paper molt important en la detecció del càncer i conseqüentment s'ha de dur a terme de manera precisa i acurada, puix que si el pacient és diagnosticat amb un fals resultat, aquest pot ser sotmès a tractaments no adequats, posant en risc la seva supervivència.

1.1.4 Biologia del càncer de mama

El càncer es caracteritza per ser una malaltia formada per una gran xarxa de mecanismes complexos i vies de senyalització, fet pel qual és difícil de donar amb un tractament que resulti en la inhibició de la seva progressió, ja que si un fàrmac ataca a una proteïna diana per aturar una via de senyalització, n'hi poden haver d'altres d'actives que permeten continuar afectant el desenvolupament i creixement de les cèl·lules canceroses. Existeixen, per tant, diferents mecanismes de senyalització i diferents tipus de receptors implicats en l'estimulació del càncer de mama.

L'estrogen i la progesterona són hormones esteroidals que, conjuntament amb els seus receptors específics, tenen un paper molt important en el desenvolupament i la progressió del càncer de mama luminal A i B (vegeu apartat 1.1.5). La seva unió amb el receptor d'estrogen (ER) i el receptor de progesterona (PR), pertanyents a la família de receptors nuclears, donarà lloc a processos transcripcionals que activen el cicle cel·lular i la proliferació de cèl·lules malignes, regulant així el creixement del tumor [10]. A causa d'aquestes propietats, ambdues hormones es consideren factors de creixement, amb una alta afinitat per unir-se al seu receptor i activar-lo. Per consegüent, es desencadenen un seguit d'efectes en la via de senyalització PI3K/AKT, entre d'altres, que induirà l'activació de factors de transcripció que regulen l'expressió de gens relacionats amb el cicle cel·lular [11].

Per altra banda, les cèl·lules tumorals poden presentar el receptor 2 del factor de creixement epidèrmic humà, més conegut com a HER2. Si el receptor, amb activitat tirosina-cinasa, és activat gràcies a la unió amb el seu lligand, és transdueix el senyal mitjançant fosforilacions, estimulant la proteïna RAS i activant la via MAPK, que resulta en una sèrie de canvis en l'expressió de gens involucrats en la divisió cel·lular. A més a més, el senyal també es pot transduir per la via PI3K/AKT, inhibint l'apoptosi i facilitant la supervivència cel·lular [12],[13].

Si la cèl·lula no mostra indicis de tenir nivells d'expressió del receptor d'estrogen, progesterona o HER2, el desenvolupament del càncer, doncs, es donarà per altres vies.

1.1.5 Subtipus moleculars del càncer de mama

S'han descrit principalment 4 subtipus moleculars [7], resumits a la taula 1, i ordenats de menor a major agressivitat. Es considera que el subtipus luminal A té un millor pronòstic, i el triple negatiu és el que presenta un pitjor pronòstic, ja que les cèl·lules creixen i s'expandeixen més ràpidament que els altres subtipus moleculars de càncer de mama.

Taula 1: Subtipus moleculars del càncer de mama.

(*): El percentatge d'afectació pot variar depenent de la font bibliogràfica consultada. (**): Estratègies terapèutiques generals per a cada subtipus.

		LUMINAL A	LUMINAL B	HER2 ENRIQUIT	TRIPLE NEGATIU
Percentatge d'afectació*		50%	10-20%	10-15%	10-15%
Expressió dels receptors	Estrogen	+	+	-	-
	Progesterona	+	+/-	-	-
	HER2	-	+/-	+	-
Estratègies terapèutiques**		Tractament hormonal			
		QUIMIOTERÀPIA			
			Teràpia dirigida a HER2		
		Teràpies diana			

1.2 Importància d'una diagnosi precoç

Un factor clau que influeix en les possibilitats de curació i supervivència d'una malaltia, es basa a determinar l'etapa en la qual aquesta es diagnostica, ja que la causa principal de les morts per càncer de mama es deu a un diagnòstic tard, quan el càncer ja es troba en estadis més desenvolupats.

Detectar el càncer en les seves etapes primerenques és de gran importància perquè permet optar a un ventall més ampli de tractaments, que aquests siguin més exitosos, i que lluitin contra el càncer abans que faci metàstasis, envaeixi altres teixits i/o arribi en estats més avançats on sigui més difícil de curar. La prevenció i el diagnòstic precoç s'han convertit en àrees d'investigació molt importants degut als avantatges i beneficis que aporten per frenar malalties, augmentar les probabilitats de supervivència i millorar la qualitat de vida del pacient. D'aquesta manera, una de les possibles solucions per reduir la taxa de mortalitat del càncer de mama en aquest cas, és realitzar un diagnòstic precoç. Com més aviat es diagnostiqui i es caracteritzi el càncer de mama que pateix un pacient, més ràpid es pot començar a tractar-lo i més possibilitats de curació tindrà. Per contra, si no s'obté un diagnòstic que detecti el càncer en les seves fases inicials, és molt possible que el nombre de defuncions vagi més a l'alça durant els pròxims anys, ja sigui tant a Espanya com a escala mundial.

Una de les aplicacions de l'aprenentatge automàtic, disciplina científica de la intel·ligència artificial, se centra en el camp de la medicina predictiva. Concretament, en els últims anys ha estat utilitzat en diagnosticar el càncer i altres malalties de manera més precisa i ràpida que els mètodes tradicionals.

1.3 Aprenentatge automàtic

Degut als grans avenços dels últims anys en tecnologia, s'ha estimulat a les màquines a què duguin a terme tasques d'aprenentatge cada vegada més complexes, i facilitar-les a l'ésser humà, ja que requeririen molt de temps poder-les realitzar.

L'aprenentatge automàtic o *Machine Learning* (ML) en anglès, és una disciplina científica de la intel·ligència artificial (IA) que se centra, mitjançant algoritmes, en què les màquines tinguin la capacitat de deduir nou coneixement a partir del que han observat prèviament. Aquesta característica és essencial per fer les màquines no només intel·ligents, sinó que també autònomes.

Un algoritme d'intel·ligència artificial es defineix per tenir la capacitat d'identificar una sèrie de patrons analitzant grans quantitats de dades, amb la finalitat de generar models que puguin predir informació sobre fets que encara no s'han observat, i deduir els millors resultats per un problema donat. Els mètodes estadístics, sense dubte, són la base fonamental dels algoritmes d'aprenentatge automàtic, doncs a través d'ells, el model podrà realitzar futures prediccions [14].

Aquests algoritmes es basen i es pot divideixen en:

- Procés de decisió: A partir d'unes dades d'entrada, l'algoritme durà a terme un seguit de passos i decisions que el portaran cap a fer una predicció.
- Funció d'error: És molt important poder avaluar la predicció del model, doncs de poc servirà un model que faci prediccions errònies.
- Procés d'optimització: Després de conèixer l'error entre el model i el conjunt de dades d'entrenament, s'ajusten un seguit de paràmetres per intentar reduir aquest error.

Depenent de com sigui l'algoritme d'aprenentatge que la màquina tingui implementat per aprendre de les dades, l'aprenentatge es pot classificar en tres subcategories: Aprenentatge supervisat, semi supervisat, i no supervisat [15].

1.3.1 Aprenentatge supervisat

L'aprenentatge supervisat té com a objectiu predir un resultat ja conegut en el conjunt de dades d'entrenament. És a dir, l'algoritme s'ensenya a partir de dades que ja tenen la resposta correcta de la variable que es vol predir. A continuació se li faciliten noves dades a l'algoritme, però aquesta vegada sense els resultats correctes de la variable la qual es vol predir. L'algoritme llavors, utilitza el coneixement adquirit durant l'entrenament per realitzar noves prediccions.

Com més quantitat de dades s'ensenyin a la màquina, més podrà aprendre i generar un model més fort.

Aquest tipus d'algoritmes poden ser de classificació o regressió [16].

- Algoritme de classificació: El problema que ha de resoldre aquest algoritme es basa a classificar un conjunt de dades en diferents subgrups, i depenent de les seves característiques, s'ajustaran millor a un subgrup o altre. En aquest cas, la variable resposta ha de ser categòrica. Depenent dels subgrups que abordi la variable, pot ser dicotòmica si només hi ha dos subgrups, o multinominal, si hi ha més de dos subgrups.
- Algoritmes de regressió: En aquest cas, la variable resposta que es vol predir és de caràcter continu.

En aquest treball, s'ha optat per generar el model amb un algoritme de tipus supervisat, XGBoost (*Extreme Gradient Boosting*). És un dels algoritmes més coneguts en l'actualitat, ja que es caracteritza per utilitzar el principi de *boosting,* en el qual es genera un model robust a partir dels resultats de models anteriors més febles, i permet obtenir bons resultats comparats amb altres models més complexos. S'explica detalladament a l'apartat 3.3.

1.3.2 Aprenentatge no supervisat

L'aprenentatge no supervisat representa tot el contrari que el supervisat. En aquest cas no hi ha un resultat concret per predir. En els problemes d'aprenentatge no supervisat, l'algoritme és entrenat amb un conjunt de dades sense cap etiqueta, doncs no s'indica què representen les dades. La idea principal d'aquest aprenentatge, és que sigui el mateix algoritme qui identifiqui patrons ocults, semblances i agrupacions en les dades. Sens dubte, aquest tipus d'aprenentatge és una taxa més complexa i també és més difícil de valorar.

1.3.3 Aprenentatge semisupervisat

L'aprenentatge de tipus semisupervisat fa referència a un terme mitjà entre els dos escrits anteriorment. Els algoritmes d'aprenentatge semisupervisat fan servir conjunts de dades etiquetades, però amb poca proporció, és a dir s'utilitzen poques dades etiquetades durant l'entrenament del model. La classificació d'aquestes poques dades serveix per guiar la classificació d'un gran conjunt de dades sense etiquetar.

Aquest tipus d'aprenentatge és útil quan no es disposa de suficients dades etiquetades, o quan es requeriria molt de temps per etiquetar-les.

1.3.4 Aplicacions de l'aprenentatge automàtic

Els models d'aprenentatge automàtic només poden detectar patrons en les dades que prèviament han vist, per això és de gran importància entrenar les màquines amb la quantitat més gran de dades possibles, ja que d'aquesta manera el model serà més robust. Aquesta habilitat fa que la intel·ligència artificial es trobi present i estigui al darrere de diversos aspectes del dia a dia [14], com per exemple en les recomanacions de pel·lícules de plataformes digitals, en el correu electrònic, reconeixement facial, però sobretot, destaca la seva importància en l'àmbit de la medicina, ja que hi ha molts problemes mèdics de diferents àrees els quals se'n poden beneficiar.

Algunes de les seves aplicacions en aquest àmbit es basen a analitzar els historials clínics de pacients, obtenir informació sobre imatges mèdiques, agilitzar el desenvolupament d'un fàrmac, i a millorar la qualitat de vida de les persones a través de la medicina personalitzada i la diagnosi de precisió.

Avui dia, la medicina continua estant enfocada a utilitzar un mateix fàrmac i estratègies de prevenció per tractar a tots els pacients amb una determinada malaltia, fet que pot ocasionar efectes secundaris si el tractament no és ideal. Investigacions recents en el camp de la genòmica apunten que el perfil genètic de cada individu determina la reacció a un fàrmac en qüestió. Amb la intel·ligència artificial es pretén posar fi a aquest actual enfocament, anomenat medicina de talla única, i convertir-lo en medicina de precisió, el qual suposa una adaptació del tractament als gens, l'entorn i l'estil de vida d'un pacient [17],[18]. Així doncs, coneixent les característiques i el perfil genètic de cada persona, la intel·ligència artificial podrà realitzar prediccions per identificar quin tractament és el més ideal, precís i efectiu per a cada pacient, entre altres aplicacions.

Tanmateix, cal esmentar que l'ús de la intel·ligència artificial no substitueix al personal mèdic ni als experts, sinó que és una eina que aporta molts beneficis i ajuda a la feina dels professionals.

1.4 Marc referencial

L'àmbit sanitari és un dels camps on es preveu que la intel·ligència artificial tingui un major desenvolupament i rellevància en el futur, però ara com ara encara és una tecnologia en creixement i se'n fa molt poc ús en hospitals, ja que no està prou optimitzada i tampoc hi ha un model de mercat que la faci arribar i sigui accessible per a tots [19]. Sí que estan donant lloc, per això, diferents investigacions dirigides per hospitals i empreses que tracten d'utilitzar diferents models d'aprenentatge automàtic per predir malalties o determinar quins pacients es poden beneficiar de determinades teràpies, entre altres. El principal objectiu de totes les investigacions és ampliar coneixements en aquest àmbit i que la intel·ligència artificial es vagi obrint camí cap a la medicina i ús en hospitals.

En l'aplicació de l'aprenentatge automàtic concretament per al càncer de mama, consten diferents investigacions on han generat models partint de les mamografies. En un estudi de la universitat de València dins d'un projecte internacional, desenvolupen un model de *Machine Learning* on l'algoritme intenta aprendre a través d'imatges procedents de mamografies, amb el propòsit de diagnosticar precoçment el càncer de mama i millorar falsos positius, per evitar així, biòpsies innecessàries. El resultat de la fiabilitat d'aquest algoritme és del 93% si es combina amb la valoració d'un professional[20]. En una altra investigació, s'empren mètodes de classificació de *Machine Learning* com K-nearest neighbor i Naïve Bayes per demostrar una precisió pròxima al 100% a l'hora de classificar tumors en mamografies [21].

S'han dut a terme també, una gran quantitat d'estudis que tenen la finalitat de predir la supervivència i la prognosi de pacients de càncer de mama mitjançant diferents mètodes de *Machine Learning* [22]. En una investigació, científics van dissenyar un algoritme mitjançant el mètode de classificació XGBoost, entre d'altres, on els resultats concloïen amb èxit que l'aprenentatge automàtic permet la detecció del càncer i la seva prognosi a través de ressonàncies magnètiques [23]. Altres estudis es basen a comparar diferents models per determinar quin és el millor amb un nivell de precisió més alt, i que pugui ser útil per la predicció de la supervivència i ajudar a prendre decisions mèdiques [24].

Recentment, s'han desenvolupat nous algoritmes com el *Deep Learning* (o aprenentatge profund), que són capaços d'aprendre sense la necessitat humana prèvia, i treure les seves pròpies conclusions. Els algoritmes utilitzats es basen en xarxes neuronals artificials que intenten imitar al cervell humà, aprenent a partir de grans quantitats de dades, però són molt més complexes i requereixen molta més potència de càlcul [25].

Diferents estudis ja han desenvolupat models basats en *Deep Learning* per predir la supervivència dels pacients de càncer de mama, com ve a ser el cas d'una investigació realitzada per la Universitat de Malaia, on varen demostrar que un algoritme de *Deep Learning* donava lloc a resultats més precisos respecte a models de *Machine Learning* [26]. Altres equips de recerca i investigadors s'han centrat també, en generar algoritmes per determinar l'estatus del receptor d'estrogen i identificar els pacients que es poden beneficiar de teràpies hormonals. Encara que falta optimitzar els models, s'han obtingut resultats que demostren com són de precisos els beneficis que aporten per reduir el temps i costos de diagnòstic [27] [28].

2. Objectius

La raó principal d'aquest treball es basa en l'ús dels algoritmes d'aprenentatge automàtic per ajudar a obtenir un diagnòstic precoç del càncer de mama de manera més ràpida, amb el propòsit de contribuir en la reducció de la taxa de mortalitat per aquesta malaltia.

2.1 Objectius generals

El principal objectiu és desenvolupar, analitzar i implementar un model de *Machine Learning* que permeti realitzar prediccions relacionades amb el càncer de mama.

2.2 Objectius específics

- Dur a terme una prospecció de les dades clíniques per identificar una variable dicotòmica que pugui ser predita a través de les dades transcriptòmiques.
- Mitjançant una anàlisi d'expressió diferencial, dur a terme un filtratge de gens per determinar quins són rellevants i candidats al classificador del model.
- Realitzar un entrenament de l'algoritme XGBoost per determinar si aquest funciona de manera correcta i validar-lo.
- Dur a terme un estudi comparatiu de la precisió entre el model logístic clàssic de regressió i el model d'XGBOOST.
  Es parteix de la hipòtesi que el model generat amb XGBoost presentarà uns millors resultats que el model logístic clàssic.

3. Metodologia

3.1 Població sobre la qual s'ha fet l'estudi

El conjunt de dades de càncer de mama que s'utilitzaran per dur a terme aquest estudi, pertanyen al projecte *The Cancer Genome Atlas,* sent abreviat com a TCGA [29]. Aquest va sorgir l'any 2006 arran d'una col·laboració entre el *Human Genome Research Insitute* (NHGRI) i el *National Cancer Insitute* (NCI) per fer front a la demanda de la millora en el diagnòstic i al tractament del càncer. El TCGA és un projecte de caràcter públic que pretén estudiar les bases moleculars i alteracions genètiques de 33 tipus diferents de càncer amb el propòsit de crear un atles amb el perfil genòmic de cada càncer. El fet que totes les dades recollides pel projecte estiguin a l'abast de tot científic, investigador i equips de recerca ha donat lloc a grans avenços en coneixement i descobriments, i és per aquest motiu que el TCGA avui en dia es considera un projecte pioner i massiu que ha generat una gran quantitat de dades òmiques que han aportat nous coneixements per la millora del diagnòstic, l'optimització del tractament i la prevenció del càncer [30].

Per tal d'estudiar i analitzar íntegrament el perfil del càncer, dins del projecte TCGA hi queden recollides dades de diferents òmiques; com la genòmica, transcriptòmica, epigenòmica i proteòmica, a més de dades clíniques que aporten informació sobre els pacients. En aquest estudi, seran utilitzades les dades de transcriptòmica, que corresponen a dades d'expressió de gens, i que han estat obtingudes a través de tècniques basades en seqüenciació (*Next generation sequencing*). Per elaborar un perfil de transcriptòmica s'empra la seqüenciació d'ARN, una tècnica d'alta precisió que quantifica seqüències d'ARN, indicant quins són aquells gens que estan expressats en una cèl·lula i en un moment determinat.

L'enorme quantitat de dades que s'han generat en el projecte del TCGA, ha fet sorgit noves eines tecnològiques avançades de visualització i anàlisis. Així doncs, el conjunt de dades necessàries de TCGA que seran usades per a fer prediccions mitjançant diferents algoritmes de *Machine Learning* són extretes del cBioPortal.

3.1.1 cBioPortal

El cBioPortal, desenvolupat al *Memorial Sloan Kettering Cancer Center*, és un recurs web que conté conjunts de dades que s'han emprat en estudis relacionats amb el càncer de diferents projectes, i informació relacionada amb dades òmiques, ja siguin de genòmica, transcriptòmica, proteòmica i/o epigenòmica. Aquest recurs, permet visualitzar, analitzar, i descarregar un gran nombre de conjunts de dades de diferents tipus de càncers, entre els quals es troben les dades del càncer de mama del projecte TCGA [31].

3.2 Entorn

Per analitzar el conjunt de dades del TCGA i implementar el model de *Machine Learning* amb XGBoost, s'utilitzarà l'RStudio, una aplicació *software* que engloba programes informàtics, procediments i documentació dedicada a l'anàlisi estadística per poder tractar amb grans quantitats de dades [32]. El llenguatge de programació en què es basa és en R i un dels seus punts forts és la facilitat en la qual es poden dissenyar operacions matemàtiques i algoritmes, importants per la construcció del model amb *Machine Learning*, i gràfiques de qualitat. L'R és l'eina principal per al desenvolupament del treball, puix és l'entorn on es duran a terme totes les operacions.

3.3 XGBoost

Abans de començar a desenvolupar el model, és important seleccionar el tipus d'algoritme que es creu que funcionarà millor per resoldre el problema (supervisat; classificació o regressió, no supervisat, semisupervisat). El propòsit en aquest cas, és predir una variable dicotòmica, i, per tant, es construirà un model de classificació amb XGBoost.

XGBoost, sigles de *eXtreme Gradient Boosting* en anglès, és una llibreria que implementa algoritmes de *Machine Learning* basats en arbres *Gradient Boosting*. És un algoritme de tipus supervisat amb intenció de predir de manera més acurada, precisa, i ràpida possible una variable partint d'un model dèbil. XGBoost utilitza arbres de decisió per la generació d'un model cada vegada més complex, que pot ser utilitzat per problemes de classificació, regressió i de rang.

La idea principal del *Gradient Boosting* es basa a generar un model de classificació robust a partir de combinacions de models i arbres més febles, per millorar-ne la capacitat predictiva. És un procés seqüencial on es creen nous models a partir dels resultats erronis en les prediccions de models anteriors. Un cop generat un nou model, es compara amb els resultats de l'anterior, i se selecciona aquell que té menys error en les prediccions. És a dir que cada vegada es crea un nou model que millora les prediccions errònies de l'anterior. Aquest procés es repeteix fins que la

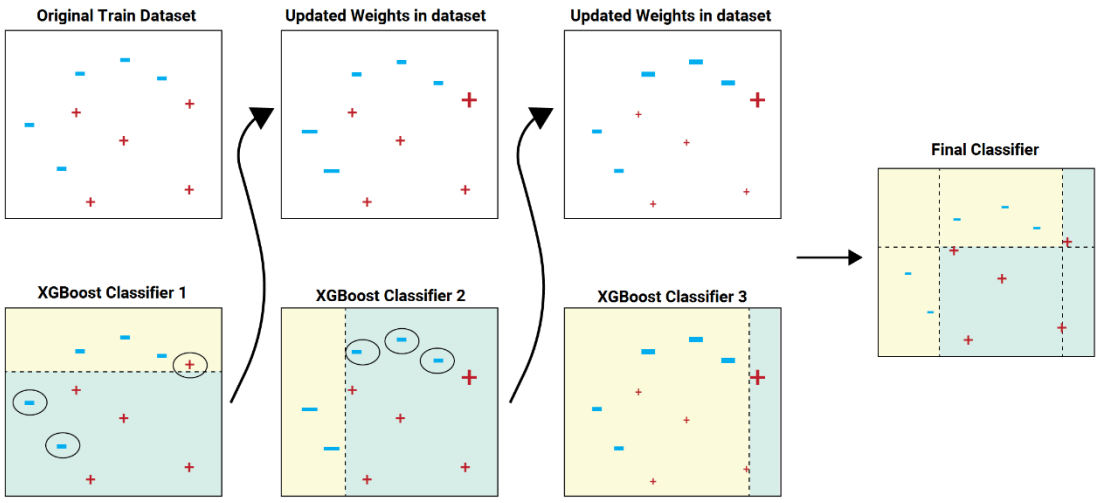


Figura 3: Representació esquemàtica del principi de *Boosting*.

diferència d'errors entre models consecutius són mínimes i/o insignificants. Finalment, els models es combinen en un classificador final.

A la figura 3, es mostra una representació esquemàtica simple del procés de *gradient boosting,* en el qual es basa l'algoritme. Es parteix de les dades d'entrenament per generar el primer classificador, on el fons groc significa la predicció d'un valor negatiu, i el verd, d'un valor positiu. Els valors predits erròniament, marcats amb un cercle, es tindran en compte a l'hora de generar el segon classificador, que prediu correctament dos dels valors que el model anterior no havia predit amb èxit. Aquest procés va continuant fins a minimitzar tots els errors. Finalment, es construeix un classificador final combinant tots els models, capaç de realitzar prediccions el màxim de precises [33].

Per treballar amb tot model d'aprenentatge automàtic cal separar el conjunt de dades en dues parts: dades d'entrenament (*training)* i dades de prova (*set).* Les dades d'entrenament són les usades per a entrenar el model, és a dir, l'algoritme aprèn d'aquestes dades i hi busca patrons. Les dades de prova són aquelles utilitzades per comprovar l'eficàcia del model creat amb les dades d'entrenament, doncs s'avalua si les prediccions generades són correctes o no. Generalment, un 70% del conjunt de dades corresponen a dades d'entrenament, i la resta, un 30% a les dades de prova.

Originalment, XGBoost es va crear en llenguatge C++, però degut als seus bons resultats i a la facilitació a l'hora d'implementar-lo, s'ha desenvolupat en molts altres llenguatges de programació per l'anàlisi de dades, com ve a ser Python, Java, R, entre d'altres. En aquest treball, XGBoost s'implementarà en el llenguatge R.

3.3.1 Paràmetres

XGBoost té una gran quantitat de paràmetres que es poden ajustar. A continuació es mostren els paràmetres principals que s'utilitzaran i que són classificats en tres categories; paràmetres generals, paràmetres específics o dependents del *booster,* i paràmetres d'aprenentatge.

3.3.1.1 Paràmetres generals

- Paràmetres que fan referència al tipus d'aprenentatge del model (*booster)*. Per a la construcció del model, es pot partir d'un algoritme basat en arbres de decisió *(gbtree)* o en funcions lineals *(gblinear).*

  Per resoldre problemes de classificació, s'utilitza *gbtree* i per problemes de regressió es pot utilitzar qualsevol d'ambdós.

  En aquest treball, doncs, *booster=gbtree*

- Dades d'entrenament. Necessari separar totes les dades en dos subconjunts. Indicar a XGBoost amb quin d'aquests conjunts de dades es generarà el model.

- *Nthread:* Correspon al nombre de fils computacionals que s'utilitzaran per entrenar el model. Es refereix als nuclis del processador de l'ordinador.

  En el model, es defineix *nthread=2*

3.3.1.1 Paràmetres dependents del booster

Paràmetres que controlen el comportament de l'algoritme escollit (*booster).* Si s'utilitza *gbtree* o *gblinear,* els paràmetres a ajustar són diferents.

En utilitzar *gbtree* com a *booster* en el nostre model, només es descriuran i s'especificaran els paràmetres associats a ell.

- *GBTREE:*

  - *Eta:* Taxa d'aprenentatge del model. Correspon a un valor entre 0 i 1.

    S'especifica aquest paràmetre com a 0,4.

  - *Nrounds:* Nombre d'iteracions que es realitzen abans d'obtenir el model final. És a dir, el nombre de vegades que es creen nous models que generin millors prediccions. Generalment, com més gran és el nombre d'iteracions, més bons resultats de predicció retornarà el model.

    A l'hora d'implementar el model, s'especifica *nrounds=125*.

  - *Max.depth:* Màxima profunditat o nombre de ramificacions en els arbres de decisió.

    S'especifica aquest paràmetre amb un valor de 100.

3.3.1.3 Paràmetres d'aprenentatge

Paràmetres ajustats per especificar l'objectiu d'aprenentatge.

- OBJECTIVE: Especifica quin tipus de classificació es vol realitzar. Les opcions disponibles per ajustar aquest paràmetre inclouen; regressió, classificació binària, multi classificació, i rang.

  Per construir un model de regressió logística per realitzar una classificació binària, el paràmetre a especificar és el següent: *objective: 'binary:logistic'.* Els resultats es mostren en forma de probabilitat.

3.4 Mètodes estadístics

3.4.1 Regressió logística

S'utilitzarà la regressió logística, una tècnica d'aprenentatge automàtic, provinent del camp de l'estadística clàssica. Tot i el seu nom, no és considerat un algoritme aplicable a problemes de regressió lineal, on la variable que es vol predir és de caràcter continu, sinó que és un mètode per problemes de classificació binària.

Aquesta tècnica consisteix a mesurar la relació entre la variable resposta i les variables independents, determinant així, la probabilitat que la variable resposta sigui certa o falsa.

3.4.2 Mesures avaluadores del model

Per avaluar les prediccions generades pel model, es tenen en compte les següents mesures: (vegeu annex A per les fórmules detallades de cada mesura)

- Exactitud: Proporció de prediccions (tant positives com negatives) que el model ha generat correctament.
- Sensibilitat: Proporció de prediccions positives que el model ha generat correctament, respecte a tots els casos positius.
- Especificitat: Proporció de prediccions negatives que el model ha generat correctament, respecte a tots els casos negatius.
- Taxa d'error: Proporció de prediccions errònies respecte a totes les prediccions generades.
- F *score:* Paràmetre similar a l'exactitud, però que té en compte la distribució de la variable resposta. S'utilitza quan el fet de predir falsos negatius i positius és crític i pot tenir greus efectes negatius.
  Com més alt és el valor, millor capacitat té el model per classificar les observacions.

3.5 Procediment

A continuació es mostra una representació esquemàtica del procés seguit pel desenvolupament del model predictiu i el compliment dels objectius presentats a l'apartat anterior.

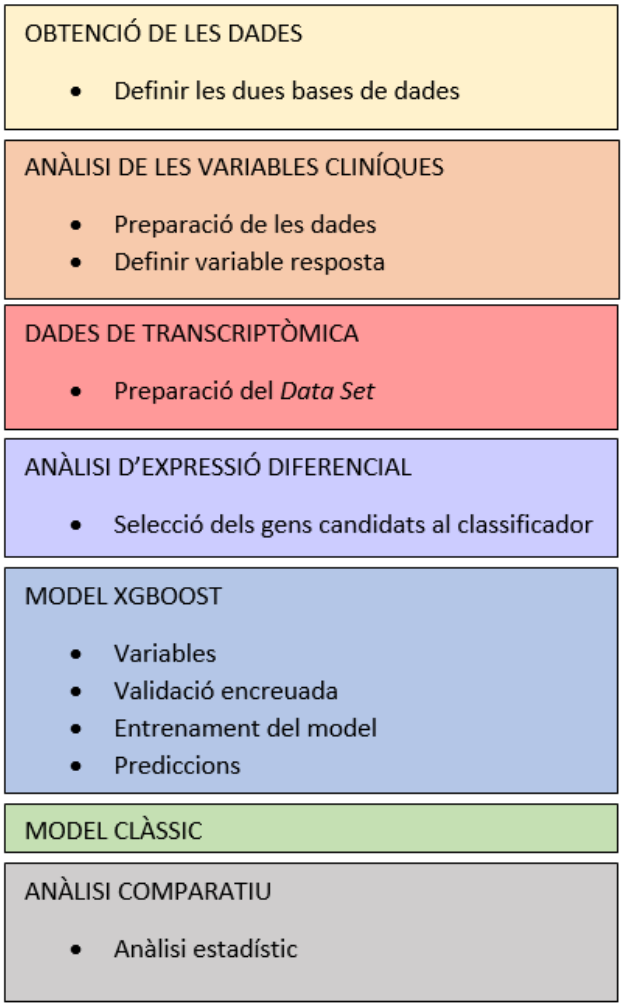


Figura 4: Esquema del procediment seguit.

3.5.1 Obtenció de les dades

Tal com s'ha descrit anteriorment a l'apartat 3.1, les dues bases de dades de les quals es parteix per generar el model predictiu, són procedents del projecte TCGA, es troben dins del cBioPortal, i s'ha accedit a elles mitjançant R Studio.

La primera base de dades conté informació sobre les variables clíniques dels pacients i la segona, conté les dades d'expressió gènica, ja en Z-scores, de tots els gens que es coneixen avui en dia i per cada pacient de l'estudi. Així doncs, el punt en comú que tenen les dues bases de dades i que servirà per unir-les, és l'identificador del pacient.

La base de dades de les variables clíniques, està composta per 1108 files i 108 columnes. Cada fila correspon al codi d'un determinat pacient, i cada columna fa referència a una variable clínica diferent.

La base de dades que conté la informació d'expressió gènica, està formada per 20.432 gens i 1102 pacients.

3.5.2 Anàlisi de les variables clíniques

A l'hora d'estudiar les variables clíniques, s'ha establert uns requisits que permetran determinar quines són les més adequades com a predictores del model i quines queden descartades. Primerament, la variable resposta ha de ser dicotòmica. Només pot prendre dos possibles valors que seran codificats com una variable binària, prenent com a valors 0 o 1, en funció de si una determinada característica és present o absent. Aquest fet determina que el model de *Machine Learning* a utilitzar és un model de classificació.

En segon lloc, és fonamental conèixer el nombre de dades disponibles per a cada variable que consta a la base de dades i verificar si a cada observació s'hi ha registrat un valor o no. Per tant, la variable resposta ha de contenir el menor nombre possible de valors absents o *missings*, per així evitar l'eliminació de pacients i la pèrdua d'informació. Com més gran sigui el nombre de mostres (pacients) més representativa serà, i més significatiu l'estudi.

3.5.3 Preparació de les dades de transcriptòmica

És necessari eliminar els gens que continguin valors absents per a tal d'entrar les dades a XGBoost. Es crearà una única base de dades que contingui les dades dels pacients comuns a la base de dades de les variables clíniques i la d'expressió. La dimensió de la nova base de dades és de 1037 x 18422. Per tant, hi consten 1037 pacients i 18422 variables d'expressió i clíniques. Cal ordenar per ordre alfabètic els pacients. Finalment, dels 1037 pacients, 239 seran negatius pel receptor d'estrogen i 799 positius.

3.5.4 Anàlisi d'expressió diferencial

Avui en dia, el nombre exacte de gens codificants per a proteïnes que conté el genoma humà encara és desconegut. Segons la base de dades *SeqRef*, dirigida pel *National Center for Biotecnhology Information* (NCBI), es considera un total de 19.558 gens codificants [34], i les estadístiques de l'última versió (39) de *Gencode*, base de dades de l'*European Molecular Biology Laboratoy* (EMBL), estima 19.982 gens codificants en el genoma humà [35]. Existeixen, doncs, diferències en el nombre de gens entre ambdues bases de dades degudes a la discrepància per la definició de gen, i a la diferent metodologia i eines bioinformàtiques que s'han emprat per a

la seva identificació, però de tota manera, els avenços científics i l'evolució de les bases de dades han permès aproximar-se al nombre real de gens codificants per a proteïnes que consten al genoma humà, ja que les primeres estimacions als anys 90 rondaven als 100.000 gens.

El principal objectiu d'aquest apartat és identificar els gens amb diferents nivells d'expressió d' RNA entre els pacients classificats com a receptor d'estrogen (ER) positiu i negatiu, amb el propòsit de dur a terme un filtratge de gens per determinar quins seran els candidats al classificador del model. Dur a terme una anàlisi d'expressió diferencial amb aproximadament els 20.000 gens que es coneixen avui en dia no tindria sentit biològicament, i, per tant, es pretén treballar amb només aquells gens dels quals es té la hipòtesi que puguin estar relacionats amb el càncer de mama, i en concret aquells gens que estan implicats en definir l'estatus del receptor d'estrogen.

S'utilitzarà un paquet d'R anomenat *Limma* per dur a terme l'anàlisi d'expressió diferencial. A l'hora d'obtenir els *top* gens expressats diferencialment, s'especificaran els següents paràmetres per seleccionar els gens candidats al classificador:

- P-valor ajustat inferior a 0.001
- Valor absolut de LogFC superior a 0.95

Els gens expressats diferencialment i que compleixin aquests dos requisits seran els candidats al classificador.

3.5.5 Generació del model XGBoost

3.5.5.1 Variables

XGBoost requereix matrius numèriques per funcionar correctament. Així doncs, caldrà convertir les columnes de dades de tipus caràcter a numèric, com venen a ser les columnes que aporten informació sobre l'expressió gènica i les variables independents. Per la variable resposta, que és dicotòmica, haurà de ser transformada a codi binari, establint valors 0 o 1.

3.5.5.2 Validació encreuada

Per la construcció del model, és necessari dividir les dades que s'entraran a XGBoost en dos subconjunts; un d'entrenament, on es generarà un model segons les característiques de les dades, i l'altre de prova, amb el qual es validarà el resultat del model generat. Aquesta separació de les dades es realitza aleatòriament, i generalment s'extreu un 70% de les dades per entrenar el model, i l'altre 30% restant són les dades de prova. A la figura 5, es pot observar un esquema del funcionament de l'algoritme.

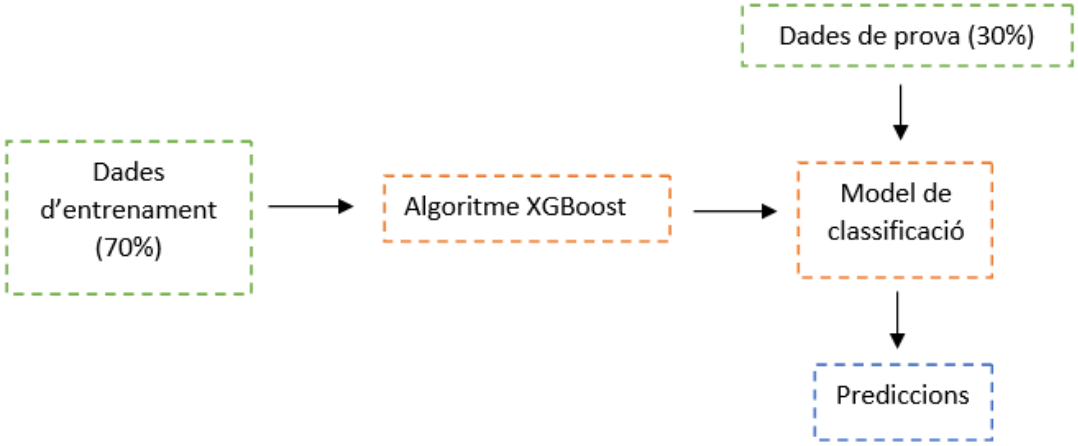


Figura 5: Esquema del funcionament de l'algoritme.

3.5.5.3 Entrenament del model i prediccions

S'utilitzaran els paràmetres esmentats a l'apartat 3.3.1 per ajustar el model, i a cada iteració obtenir-ne un de més robust, que millori els errors de classificació. Aquest procés es repeteix fins que no s'aconsegueixen millores entre dos models consecutius. Finalment, el model genera prediccions i retorna la probabilitat que un cas en particular pertanyi al valor 1 de la variable resposta. A continuació es mostren les funcions principals de XGBoost per a l'entrenament del model i per la generació de prediccions:

```
modelentrenament <- xgboost(data=traindata_matrix,
                            booster="gbtree",
                            nrounds=125,
                            max.depth=100,
                            eta=0.4,
                            nthread=2,
                            objective="binary:logistic")
```

```
p=predict(modelentrenament,newdata=testdata_matrix,ntreelimit=10)
```

3.5.6 Generació del model clàssic

A l'hora de generar el model clàssic de regressió logística, es segueixen passos similars als del model per XGBoost. Caldrà, també, separar el conjunt de dades, en dades d'entrenament i de prova.

Per ajustar el model de regressió logística (model lògit), s'utilitzarà la funció *glm(),* que permetrà generar prediccions de dues respostes, 0 o 1. A la sintaxi bàsica per obtenir el model logístic, cal indicar la funció de probabilitat (*Family).* En aquest cas, és una funció binomial.

A continuació es mostren les funcions principals per l'obtenció de prediccions amb el model logístic:

```
modelo.logit<-glm(A00ER ~ .,data=testrenament_mc,family =
"binomial"(link=logit),maxit=100 )

log.odds<- predict(modelo.logit, newdata = ttest_mc, ntreelimit=10)
```

3.5.7 Anàlisis comparatiu

Finalment, es durà a terme un estudi estadístic per comparar la precisió dels dos models (XGBoost i model clàssic). S'utilitzaran les mesures descrites a l'apartat 3.4.2 per avaluar-los.

4. Resultats

4.1 Prospecció de dades clíniques

Dins l'estudi del càncer de mama, existeix una base de dades que conté un conjunt de variables clíniques per cada pacient. Per dur a terme aquest projecte, no es farà ús de totes les diferents variables, sinó que només una serà la variable de major interès i la qual es vol predir emprant *Machine Learning.* El principal objectiu d'aquest apartat és, doncs, identificar una variable dicotòmica per classificar els pacients en dos subgrups i que aquesta pugui ser predita a través de dades de transcriptòmica.

A la taula 2, es mostren les variables clíniques que s'han determinat de major interès.

Taula 2: Variables clíniques de major interès.

(*):Nombre de dades classificades com a Indeterminades. (**): Nombre de dades classificades com a Equívoques. Les dades del qual el seu valor és indeterminat o equívoc, es consideren com a valors absents.

VARIABLE	VALORS ABSENTS	DESCRIPCIÓ
EDAT	5	Edat dels pacients de l'estudi.
ER_STATUS_BY_IHC	53 + 2*	Absència o presència del receptor d'estrogen.
IHC_HER2	185+12*+180**	Absència o presència de la proteïna HER2.
PR_STATUS_BY_IHC	54+4*	Absència o presència del receptor de progesterona.
OS_STATUS	4	Estat de supervivència dels pacients.
SEX	4	Sexe dels pacients.

La variable que classifica els pacients segons l'absència o presència del receptor d'estrogen (ER_STATUS_BY_IHC) serà la variable resposta o dependent, doncs és la que conté menys valors absents i és dicotòmica.

4.1.3 Distribució de les variables independents

4.1.3.1 Distribució del sexe

A la figura 6, s'observen clarament diferències en la proporció del sexe dels pacients de l'estudi de càncer de mama. 1092 observacions per pacients dones i només es consta de 12 pacients pel sexe masculí, un 1,08%. El fet que el nombre de mostres sigui tan baixa, tindrà influència en els resultats, doncs aquests no seran significatius. A més a més, durant la validació encreuada, quan les dades es reparteixen aleatòriament en els grups d'entrenament del model i prova, podria passar que algun grup no tingués cap dada del sexe masculí, de manera que es poden donar errors en els resultats que genera el model. Així doncs, és convenient eliminar les 12 observacions que fan referència al sexe masculí. A partir d'aquest punt, el treball se centra en el sexe femení.

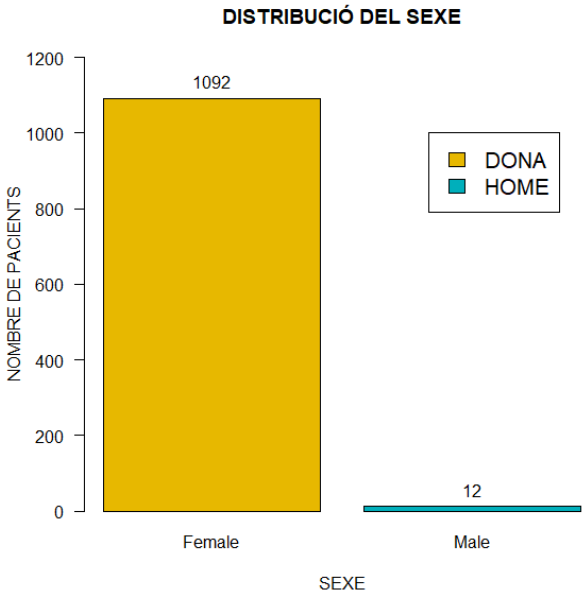


Figura 6: Distribució del sexe dels pacients.

4.1.3.2 Distribució de l'edat i de la variable resposta

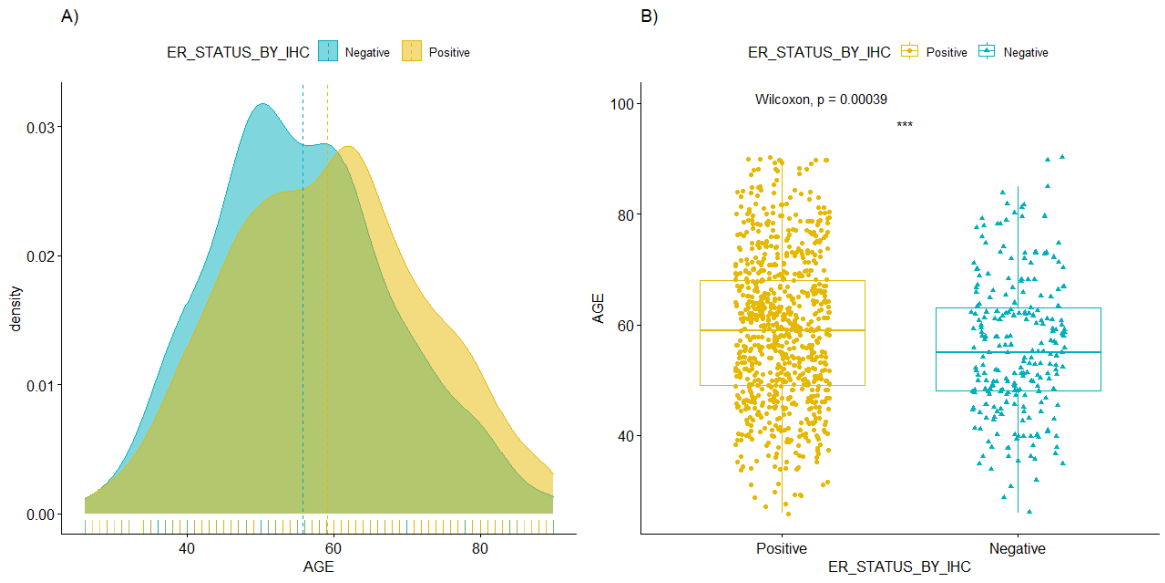


Figura 7: Distribució de l'edat i l'estatus del receptor d'estrogen.

A) Gràfic de densitat. Les línies blava i groga representen la mitjana d'edat de les pacients negatius i positius pel receptor d'estrogen, respectivament. B) Diagrama de caixes. Nivell de significació ***, p-value inferior a 0,001.

La distribució de l'edat dels pacients no és del tot exacte entre el grup de receptors d'estrogen negatius i positius, doncs la prevalença de presentar receptors d'estrogen (ER) és major en les dones de més edat que en les dones més joves (Figura 7A). Per contra les probabilitats de no

expressar aquest receptor son més altes en les pacients de menys edat, determinant un pitjor pronòstic. A la figura 7A i B, es poden observar les diferències en la mitjana d'edat de les pacients pels dos subgrups.

La mitjana d'edat de les pacients positives pel receptor d'estrogen és de 59,18 anys, i la mitjana per les pacients negatives és de 55,81 anys (vegeu apartat 2.4 de l'annex B). El p-valor obtingut després de realitzar el test d'igualtat de dues mitjanes (Figura 7B) indica que hi ha diferències significatives entre l'edat dels pacients positius per al receptor d'estrogen i els negatius.

4.1.4 Variables

A continuació es mostren les variables que el model ha de tenir en compte:

- Variables independents: Edat, en anys; i nivells d'expressió gènica.
- Variable resposta: Receptor d'estrogen (Negatiu, 0; Positiu, 1).

4.2 Distribució de la variable resposta

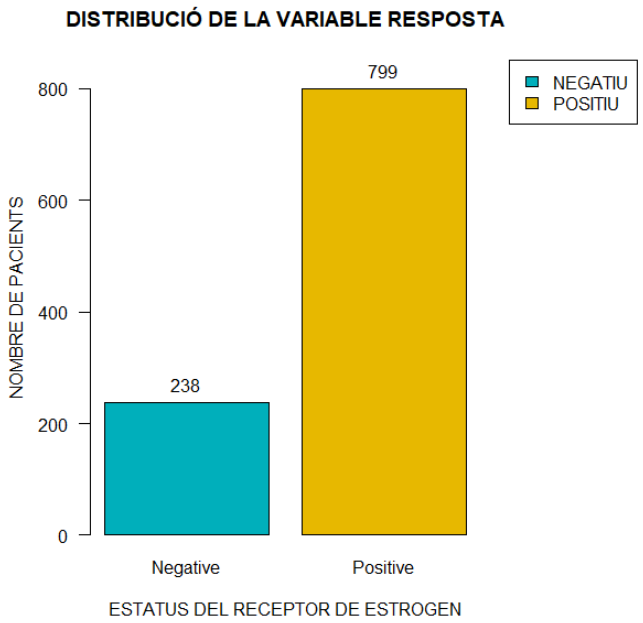


Figura 8: Distribució del receptor d'estrogen

Després d'haver eliminat les dades referents al sexe masculí, els valors absents i els pacients que no es troben a ambdues bases de dades (variables clíniques i expressió de gens), a la figura 8 es mostra la distribució del receptor d'estrogen, variable dependent la qual serà predita a través del model *Machine Learning.* Aquesta variable es divideix en 799 pacients positius i 238 pacients negatius.

4.3 Anàlisis d'expressió diferencial

La variable que es pretén predir és la del receptor d'estrogen, i, per tant, es formula la hipòtesi que el gen ESR1, que codifica per aquest receptor, serà un dels gens que mostra diferents nivells d'expressió entre els dos subgrups de pacients (Positius i Negatius).

Tal com s'ha esmentat a l'apartat 3.5.4, s'ha utilitzat la llibreria *limma* per dur a terme aquesta anàlisi i realitzar un filtratge de gens. Es parteix d'una base de dades que conté els nivells d'expressió de 20.433 gens, i després d'eliminar tots aquells que contenen valors absents, n'hi consten 18.421. Amb els paràmetres establerts (p-value menor a 0.001 i logFC superior a 0.95 o inferior a -0.95) es retornen 1.874 gens expressats diferencialment entre les mostres de pacients negatius i positius pel receptor d'estrogen. Tots aquests 1.874 gens seran entrats al model XGBoost per realitzar prediccions.

A la taula 3, es mostren els 15 primers gens els quals la seva expressió més es diferencia. El p-valor ajustat determina quins gens són significativament diferents entre les dues mostres i el logFC determina quins gens estan sobre expressats i quins sub expressats. És a dir, que un valor positiu per logFC significa que el gen en qüestió es troba sobre expressat en mostres negatives pel receptor d'estrogen, i si el valor és negatiu, el gen es troba poc expressat, el que determina que s'expressarà en mostres positives pel receptor d'estrogen.

Taula 3: *Top* 15 gens expressats diferencialment entre pacients amb receptor d'estrogen negatiu i positiu.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
ILF2	3.579442	1.3912745	27.37574	1.055474e-124	1.944288e-120	273.9581
B3GNT5	2.348810	0.4110344	26.81570	7.797989e-121	7.182338e-117	265.0907
DESI2	2.783041	1.0357054	26.17658	1.945094e-116	1.194353e-112	255.0119
FAM171A1	2.712459	0.5029082	25.79997	7.410620e-114	3.412776e-110	249.0958
FOXA1	-1.427812	-0.2507851	-25.78453	9.451169e-114	3.482000e-110	248.8537
UCK2	2.827927	0.9382346	25.72730	2.328139e-113	7.147774e-110	247.9562
ZBTB4	-1.533485	-0.7443442	-25.15384	1.899173e-109	4.997809e-106	238.9900
ANP32E	3.607203	1.1135186	25.04731	1.006710e-108	2.318075e-105	237.3296
HDGF	3.039430	1.2578967	24.96952	3.399327e-108	6.957668e-105	236.1182
PDE7A	2.319838	0.6527601	24.90749	8.964840e-108	1.651413e-104	235.1529
SFT2D2	2.530334	0.7496463	24.86749	1.674643e-107	2.804419e-104	234.5308
RABEP1	-1.258442	-0.4849750	-24.55962	2.037961e-105	3.128440e-102	229.7509
YEATS2	2.269913	0.5341045	24.48418	6.593240e-105	9.342622e-102	228.5821
SUV39H2	2.937077	0.7079760	24.22770	3.545651e-103	4.665316e-100	224.6152
GATA3	-1.313239	-0.2417125	-24.19011	6.351958e-103	7.800628e-100	224.0348

Dels 15 gens mostrats a la taula 3, el gens FOXA1, ZBTB4, RABEP1 i GATA3 es troben poc expressats en mostres negatives pel receptor d'estrogen. S'estima que aquests es trobaran sobre expressats en mostres positives pel receptor d'estrogen, i, per tant, són gens associats a un pronòstic favorable. Per contra, la resta de gens que estan sobre expressats en les mostres de pacients negatius per al receptor d'estrogen, estaran relacionats amb el càncer de mama HER2 enriquit o triple negatiu, dos subtipus on no hi ha expressió del receptor d'estrogen (vegeu taula 1) i, per tant, aquests gens estan associats a un pitjor pronòstic.

En quant a la hipòtesi formulada anteriorment, s'observa que el gen ESR1, que codifica pel receptor d'estrogen, es troba concretament a la posició 308 dels *top* gens expressats diferencialment.

Segons la taula 4, el gen ESR1, es considera diferencialment expressat entre les mostres negatives i positives. El valor negatiu del logFC indica que es trobarà poc expressat en els pacients negatius per receptor d'estrogen i s'estima que es trobi més expressat en pacients positius.

Taula 4: P-valor ajustat i logFC del gen ESR1.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
ESR1	-0,9545164	-0,123897	-17,61913	5,265509E-61	3,13903E-59	1279765

4.4 Regressió logística amb XGBoost

4.4.1 Variables importants

En realitzar l'entrenament del model, aquest no té en compte per igual totes les variables independents (gens i edat), sinó que hi poden haver gens que estan més implicats en la construcció del model, i que, per tant, són més importants. El guany és la mètrica que determinarà la contribució de cada variable en cada arbre del model. Un valor més alt, significa que el gen, en aquest cas, és més valuós per generar una predicció. A la figura 9, es representen els 10 primers gens que el model considera que tenen una major implicació. El primer de tots, és el ESR1, el gen que codifica pel receptor d'estrogen. Així doncs, és lògic que es consideri aquest gen important, ja que l'objectiu és predir l'estatus del receptor d'estrogen.

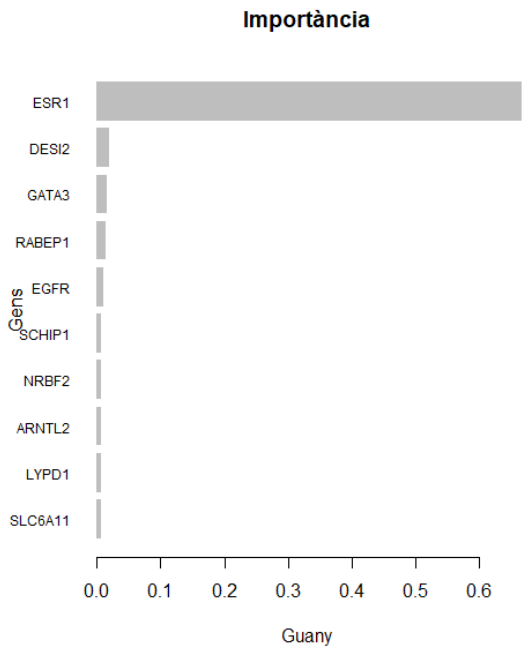


Figura 9: Representació gràfica dels 10 primers gens més importants pel model.

4.4.2 Prediccions

A partir del model d'entrenament, s'han generat noves prediccions amb les dades que s'havien destinat a provar el model. Tant les prediccions que ha generat el model, com el valor real del receptor d'estrogen, es poden comparar a la taula 5. El model ha predit amb èxit 68 casos de pacients negatius i 229 casos de pacients positius. Això i tot, s'han generat 5 casos de falsos positius i 9 de falsos negatius.

Taula 5: Matriu de confusió amb els resultats de predicció del model XGBoost.

L'estatus del receptor d'estrogen és negatiu si el valor es 0, i positiu si es 1.

Predicció	Valor actual	Freqüència
0	0	68
1	0	5
0	1	9
1	1	229

4.4.3 Avaluació del model

Per avaluar el model de classificació binària, es tindran en compte un seguit de conceptes estadístics (vegeu apartat 3.4.2) entre els quals destaca el valor F. Aquest model ha predit que 9 pacients seran negatius pel receptor d'estrogen, quan realment són positius, i 5 de positius quan actualment són negatius. En aquest context, existeix un greu problema a l'hora de predir falsos negatius i positius, ja que tindran un impacte en la salut del pacient. Per això, s'utilitza principalment el valor F per avaluar el model, puix que aquesta fórmula té en compte els falsos negatius i positius. A la taula 6, es mostren els resultats de l'avaluació del model generat amb XGBoost.

Taula 6: Paràmetres avaluadors del model XGBoost.

Exactitud	Taxa d'error	Especificitad	Sensiblitat	F Score
0.9549839	0.04501608	0.9315068	0.9621849	0.970339

4.5 Regressió logística amb el model clàssic

4.5.1 Prediccions

A la taula 7 es mostren els resultats de predicció generats amb el model clàssic de regressió logística. El model ha estat capaç de predir correctament 40 casos negatius i 87 casos positius de receptor d'estrogen. Per contra, ha predit 130 casos de falsos negatius i 14 casos de falsos positius.

Taula 7: Matriu de confusió amb els resultats de predicció del model clàssic.

L'estatus del receptor d'estrogen és negatiu si el valor es 0, i positiu si és 1.

Predicció	Valor actual	Freqüència
0	0	40
1	0	14
0	1	130
1	1	87

4.5.2 Avaluació del model

Per l'avaluació del model clàssic de regressió logística també s'han tingut en compte els mateixos paràmetres utilitzats per avaluar el model generat amb XGBoost. Per mesurar la precisió del model s'utilitza el valor F.

Taula 8: Paràmetres avaluadors del model clàssic de regressió logística.

Exactitud	Taxa d'error	Especificitat	Sensiblitat	F Score
0.4686347	0.5313653	0.4009217	0.7407407	0.5471698

5. Discussió de resultats

5.1 Prospecció de dades clíniques

En aquest estudi s'ha assignat el receptor d'estrogen com a variable resposta i la que serà predita a través del model de *Machine Learning.* Això i tot, clínicament, no és la variable de més interès. Respecte a les variables mostrades a la taula 2, la variable OS_STATUS, que determina si un pacient ha sobreviscut al càncer o no, és la que presenta un major interès clínic. Això és degut al fet que determinar l'efecte que tindrà un tractament en la supervivència d'un pacient és un factor molt important per reduir els nivells de mortalitat, ja que si es pot predir que un determinat tractament no serà efectiu pel pacient, aquest es pot evitar i decidir tractar el pacient amb altres opcions més exitoses.

En aquest cas, però, no hi ha constància de cap variable a la base de dades relacionada amb el tipus de tractament que ha rebut el pacient i, per tant, queda descartada la variable OS_STATUS com a variable resposta per aquest treball.

Sí que hi ha diferents estudis i s'han publicat articles on han emprat models d'aprenentatge automàtic per predir la supervivència d'un pacient i la seva resposta al tractament, per poder-se a anticipar al càncer i seleccionar el tractament més eficient, fet que es considera de gran importància per reduir les altes taxes de mortalitat, sobretot en països subdesenvolupats [36].

5.1.1 Beneficis de predir el receptor d'estrogen

L'aprenentatge automàtic és una disciplina amb una enorme capacitat de realitzar prediccions en molt menys temps que els especialistes mèdics, característica que el converteix en una eina potent per revolucionar la medicina en el futur degut als avantatges que presenta.

Un dels principals beneficis d'usar *Machine Learning* per la detecció del receptor d'estrogen és que s'evita realitzar una anàlisi immunohistoquímica al laboratori, fet que redueix dràsticament el temps de diagnosis, ja que s'obtenen els resultats pràcticament a l'instant. En cas contrari, rebre una diagnosi mèdica pot tardar setmanes o fins i tot uns mesos, si el sistema sanitari és lent [37]. El temps és un factor molt important en el càncer, doncs com més ràpid s'obtingui una diagnosi, més aviat es pot començar a tractar el pacient. El fet de predir el receptor d'estrogen ajudarà a determinar si el subtipus molecular del càncer és de tipus Luminal (sigui A o B) o no (vegeu apartat 1.1.5), avaluar el pronòstic i decidir si el pacient pot rebre un tractament hormonal dirigit al receptor d'estrogen (ER). És a dir, que si es prediu que un pacient és positiu per receptor d'estrogen, el seu subtipus molecular és luminal, fet que indica que el pacient té un pronòstic favorable.

A més a més, com que no es realitza la immunohistoquímica, no es requereix un patòleg i es redueixen els costos, fet rellevant en zones rurals de països en vies de desenvolupament o pobres, que tenen un sistema sanitari deficient [38].

5.1.2 Relació de l'edat i receptor d'estrogen

Tal com indiquen els resultats de la figura 7, existeixen diferències significatives entre la mitjana d'edat dels pacients positius per receptor d'estrogen i els negatius. Així doncs, l'edat es considera un factor important que influeix indirectament al pronòstic de les pacients, ja que és un condicionant de la presència o absència del receptor d'estrogen (ER), el qual sí que influeix directament en la supervivència. Les pacients de major edat són, per tant, més propenses a desenvolupar tumors de menor agressivitat, amb taxes més altes d'expressió de ER, i amb un pronòstic més favorable.

Aquests resultats concorden amb els obtinguts en diferents estudis on, clarament, posen en evidència que una diagnosi a una edat jove es relaciona amb una esperança menor de supervivència i una alta incidència de desenvolupar un subtipus de càncer de mama més agressiu, com el triple negatiu i el HER2, a més a més de detectar la malaltia en estats més avançats [39], [40].

5.2.1 Anàlisis d'expressió diferencial

Dins la llista de gens ordenats segons el seu grau de diferenciació hi trobem ESR1. El valor del seu logFC és de -0,95 aproximadament (vegeu taula 4), i a l'hora de realitzar el filtratge s'ha ajustat un |logFC| superior a 0,95 (vegeu apartat 3.5.4), en comptes de |logFC|>1, de manera que amb aquest paràmetre de filtratge, s'ha forçat a incloure el gen ESR1 dins dels gens filtrats, ja que es considera un gen important per la classificació del model.

A la taula 3, es mostra la llista dels 15 primers gens on la seva expressió més es diferencia entre mostres negatives i positives pel receptor d'estrogen, però es filtren més de 1.800 gens on els seus nivells d'expressió difereixen significativament entre les dues mostres. Així i tot, no tots aquests gens han estat estudiats en detall per científics i investigadors i, per tant, d'alguns d'ells no se'n té un coneixement sòlid que evidencií com estan relacionats amb el càncer de mama, en aquest cas. És per això que hi ha gens més coneguts que d'altres, i s'ha pogut establir diverses connexions entre alguns d'ells, com per exemple, avui en dia es coneix la relació entre els gens GATA3, FOXA1 i ESR1. Tots aquests 3 gens es troben expressats diferencialment, tal com es pot observar a les taules 3 i 4, respectivament. Els gens GATA3 i FOXA1 es troben entre els 15 primers gens més diferenciats entre mostres negatives i positives. En els pacients positius per receptor d'estrogen, aquests tendeixen a estar sobre expressats, determinant així, una relació positiva entre ells.

El gen FOXA1 pertany a la família de factors de transcripció que regula l'estructura de la cromatina i l'expressió gènica. La sobre expressió d'aquest en el càncer de mama, està implicat en el desenvolupament d'un subtipus luminal, sigui A o B, i, per tant, és un gen associat a un bon pronòstic [41]. El gen GATA3, pertany també a una família de factors de transcripció involucrats en regular l'expressió gènica d'altres gens, i en aquest cas regula l'expressió del gen que codifica pel receptor d'estrogen, ESR1. Com a resultat, ambdós gens codifiquen per factors de

transcripció associats amb la regulació del gen ESR1 [42]. La seva sobre expressió defineix un tumor de càncer de mama positiu pel receptor d'estrogen.

A més a més, les investigacions dutes a terme fins al dia d'avui han pogut establir aquests 3 gens dins la mateixa xarxa transcripcional i que existeix un grau elevat de dependència entre ells. Se suggereix també, que el gen GATA3 intervé en la unió de ESR1 i FOXA1 als elements ERE de la cromatina, i es detecten nivells elevats pel que fa a la relació d'aquests tres gens amb el grup positiu de pacients per receptor d'estrogen en el càncer de mama [43].

Per altra banda, els gens ILF2 i B3GNT5 són els que presenten nivells de diferenciació més significatius (Taula 3). El logFC d'ambdós és positiu, amb un valor de 3,57 i 2,34 respectivament. Aquests valors indiquen que els gens estan sobre expressats en els pacients negatius per receptor d'estrogen, i, per tant, estan associats a un pitjor pronòstic, ja que el fet de ser negatiu per ER, implica que el subtipus molecular de càncer de mama que un pacient pateix és HER2 enriquit o el triple negatiu, els quals són els més agressius. Aquests resultats són corroborats per diferents estudis en els quals van determinar que unes elevades taxes de transcripció del gen B3GNT5 promou tumors de mama més agressius. Els resultats d'aquests estudis van afirmar que aquest gen, B3GNT5, és molt probable que presenti alts nivells d'expressió en un fenotip triple negatiu [44]. Pel que fa al gen ILF2, diferents estudis demostren que una elevada expressió per aquest gen, està relacionada amb el fet que el tumor desenvolupi característiques més agressives, com per exemple un grau histològic més alt i mutacions al gen BRCA1, resultant en el subtipus triple negatiu o HER2 enriquit, que presenten un baix percentatge de supervivència del pacient [45].

5.3 Avaluació dels models de prediccions

Les prediccions que ha generat el model de XGBoost, mostren una clara millora de la precisió de l'algoritme, respecte a les prediccions que ha generat el model de regressió logística clàssica. El valor F és el més indicat per avaluar la precisió dels models. L'algoritme generat amb XGBoost té un valor F igual a 0,97. En canvi, el valor F del model clàssic és de 0,54 (Taula 6 i 8 respectivament). El fet que es forci a incloure el gen ESR1 dins dels gens filtrats, podria condicionar a que el model tingui un bon resultat (0,97).

La taxa d'error, que equival al percentatge de prediccions errònies que ha generat el model, és molt més baixa en el model de XGBoost (0,04) que no pas en el model clàssic (0,53), on és considerada molt elevada per un model de prediccions, ja que no es prediuen correctament ni la meitat dels casos. Aquests resultats van associats a la sensibilitat i l'especificitat del model clàssic, ja que tal com s'observa a la taula 8, els valors són molt baixos comparats amb els resultats del model generat amb XGBoost. Així doncs, sobta que el model clàssic obtingui, el que es consideren mals resultats (valor F 0,54). Tot i que s'esperava que els resultats no fossin igual de bons que els del model generat amb XGBoost, s'estimava que el model clàssic generés més prediccions correctes de les que ha generat i que el valor F fos més elevat. Aquest fet pot ser degut a que s'ha produït un sobre ajust del model i que, per tant, no es poden estimar tots els coeficients de cada variable, a més a més de no poder generar prediccions en alguns pacients.

La sensibilitat i especificitat del model XGBoost presenta uns resultats pròxims a 1 (taula 6), fet que indica que és capaç de generar un gran percentatge de prediccions correctes i un baix percentatge de generar-ne d'incorrectes. Així doncs, el model generat amb XGBoost presenta uns millors resultats que el model de regressió logística clàssica, i apunta al fet que l'ús dels algoritmes basats en intel·ligència artificial són una bona eina per generar prediccions reals, i que s'ha de seguir desenvolupant en un futur degut als avantatges que comporta.

Partint de les investigacions prèvies que indicaven que el *Machine Learning* és una tècnica més eficaç i precisa per generar prediccions, la hipòtesi formulada a l'inici del treball es pot afirmar, doncs analitzats els resultats, l'algoritme amb XGBoost és capaç de realitzar millors prediccions relacionades amb el càncer de mama, que el model clàssic.

6. Conclusió

El principal objectiu del treball, que consistia a desenvolupar, analitzar i implementar un algoritme basat en *Machine learning* per generar prediccions relacionades amb el càncer de mama, s'ha complert. Per avaluar-ne la precisió, s'ha comparat amb un model de regressió logística clàssica, i tot i que ambdós models tenien la mateixa finalitat, predir el receptor d'estrogen (ER), no han tingut els mateixos resultats.

El model generat a través de XGBoost té uns nivells de precisió molt elevats (97%), i, en canvi, la precisió del model de regressió logística clàssica és molt menor (54%). Per tant, es pot afirmar la hipòtesi formulada a l'inici del treball: l'algoritme d'aprenentatge automàtic generat amb XGBoost obté bons resultats per a la classificació del receptor d'estrogen en els pacients de càncer de mama.

Més específicament, també es conclou que el gen ESR1 presenta diferents nivells d'expressió entre mostres de pacients negatius i positius pel receptor d'estrogen, i que és summament important per les bones prediccions del model.

6.1 Limitacions i millores a realitzar en projectes futurs

Tot i els resultats obtinguts, hi poden haver diferents factors que hagin esbiaixat els resultats. Les dades utilitzades per la realització del treball han estat en Z-scores, una mesura d'estandardització on es mostra la desviació estàndard d'un valor respecte a la mitjana de les mostres. D'aquesta manera, es podrien haver seleccionat altres tècniques de normalització que representin millor les dades.

Un altre factor que ha pogut esbiaixar els resultats ha estat els paràmetres seleccionats de l'algoritme (vegeu apartat 3.3.1). En aquest cas, s'han realitzat diverses proves per donar amb els paràmetres que millor s'ajustaven i aconseguien més bons resultats, però es desconeix si altres ajusts donarien lloc a millors resultats.

6.1.1 El càncer de mama en homes

El càncer de mama és el càncer més freqüent entre el sexe femení, i a raó d'aquest fet, la societat el coneix per afectar a les dones, però no n'és una malaltia exclusiva, ja que en una minoria de casos també es pot desenvolupar en homes. Aproximadament l'1% de tots els càncers de mama són diagnosticats en homes, i durant l'any 2018 a Espanya varen ser detectats 328 casos en homes, respecte als 32.825 casos en dones [46].

Degut a les xifres de casos presentades per aquest tipus de càncer, el càncer de mama en homes és considerada una malaltia minoritària i, a la vegada, rara. Estudis apunten que existeixen diferències biològiques i epidemiològiques entre el càncer de mama d'ambdós sexes [47]. Els

càncers de mama desenvolupats en el sexe masculí son més propensos a ser del subtipus luminal, que expressen el receptor d'estrogen (RE) i el receptor de progesterona (RP), i menys propensos a expressar HER2. Concretament, un 92% dels càncers de mama en homes són positius pel receptor d'estrogen (RE), mentre que en les dones, el percentatge d'estrogen positiu recau al 78%, fet que indica que el càncer de mama en el sexe femení presenta una major diversitat en els subtipus moleculars [48] i, tot i que la taxa de mortalitat ha disminuït en els últims anys, encara és superior en homes si es compara amb les estadístiques del sexe femení [49]. Amb tot i això, el càncer de mama en homes no està completament estudiat ni entès, ja que en ser una malaltia minoritària en la qual la seva incidència és molt baixa, afectant només a 1 home per cada 99 dones, investigadors, equips de recerca, recursos en investigació i assajos clínics se centren principalment en el càncer de mama femení, fent que hi hagi una limitació pel que fa a estudis i conseqüentment en coneixement en el càncer de mama masculí [47]. És per això que en els pròxims anys s'haurien de destinar més recursos per entendre i donar visibilitat a aquest càncer, abordar noves estratègies terapèutiques, i aproximar-se a la igualtat de gènere pel que fa a assistència sanitària.

Aquest mateix treball és un exemple de les desigualtats que hi ha en el coneixement del càncer de mama femení i masculí, ja que s'ha desenvolupat envers les dones, doncs només s'han tingut en compte les dades referents a elles (vegeu apartat 4.1.3.1), degut a la falta d'informació dels homes. De cara al futur, seria convenient aplicar aquest model de *Machine Learning* als homes per veure si es donen resultats similars entre els dos sexes. Per això, caldria generar una base de dades amb el màxim nombre possible d'informació referent als pacients del sexe masculí.

6.2.2 Model pel receptor HER2

Algunes investigacions han revelat que el fet de determinar la presència de receptors HER2 a la superfície de les cèl·lules canceroses pot ser un procés dificultós i ambigu [50], a causa del criteri de cada especialista, i, per tant, alguns resultats de les anàlisis no permeten inclinar-se cap a un resultat positiu o negatiu. Així doncs, a vegades no s'obtenen els resultats reals que s'haurien d'obtenir, i aquesta imprecisió fa que alguns pacients no rebin el millor tractament i pronòstic possible. Si realment el càncer de mama presenta receptors de HER2, però s'ha determinat que és negatiu, es priva al pacient de rebre un tractament encarat a HER2 i que podria mostrar bons resultats. A més a més, el pacient pot quedar exposat als efectes secundaris dels medicaments sense beneficiar-se canvi, ja que el tractament no és adequat [51].

Una de les millores a realitzar en futures investigacions podria consistir a generar un model predictiu que ajudi als especialistes en aquest procés. De la mateixa manera que s'ha fet en aquest treball per predir el receptor d'estrogen, es podria generar un model que tingués la màxima capacitat possible per generar prediccions pel receptor HER2. A més a més, el model podria ser més complex i que fos capaç de predir no només HER2, sinó que també el receptor d'estrogen, amb l'objectiu de classificar el càncer de mama del pacient en un dels 4 subtipus moleculars, luminal A o B, HER2 enriquit, i triple negatiu.

7. Bibliografia

[1] "¿Qué es el Cáncer?" https://www.ivo.es/tipos-de-cancer/que-es-el-cancer/ (accessed May 29, 2022).

[2] "¿Qué es el cáncer? - NCI." https://www.cancer.gov/espanol/cancer/naturaleza/que-es (accessed May 29, 2022).

[3] K. Zhu *et al.*, "Oncogenes and tumor suppressor genes: Comparative genomics and network perspectives," *BMC Genomics*, vol. 16, no. 7, pp. 1–11, Jun. 2015, doi: 10.1186/1471-2164-16-S7-S8/FIGURES/6.

[4] "Enfermedades cardiovasculares." https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1 (accessed May 04, 2022).

[5] "Cancer Tomorrow." https://gco.iarc.fr/tomorrow/en/dataviz/isotype?types=1&single_unit=500000 (accessed May 04, 2022).

[6] "Cancer Today." https://gco.iarc.fr/today/online-analysis-multi-bars?v=2020&mode=cancer&mode_population=countries&population=900&populations=724&key=total&sex=2&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=10&group_cancer=1&include_nmsc=0&include_nmsc_other=1&type_multiple=%257B%2522inc%2522%253Atrue%252C%2522mort%2522%253Afalse%252C%2522prev%2522%253Afalse%257D&orientation=horizontal&type_sort=0&type_nb_items=%257B%2522top%2522%253Atrue%252C%2522bottom%2522%253Afalse%257D (accessed May 01, 2022).

[7] A. Lucena *et al.*, "Clasificación actual del cáncer de mama. Implicación en el tratamiento y pronóstico de la enfermedad. Especial Monográfico de Mama Molecular classification of breast cancer. Treatment and prognosis implications," doi: 10.37351/2021322.9.

[8] A. Roulot *et al.*, "Tumoral heterogeneity of breast cancer," *Ann. Biol. Clin. (Paris).*, vol. 74, no. 6, pp. 653–660, Nov. 2016, doi: 10.1684/ABC.2016.1192.

[9] M.-S. S. Andrés, G.-P. L. Tatiana, and O.-Z. S. Esperanza, "Clasificación inmunohistoquímica del cáncer de mama y su importancia en el diagnóstico, pronóstico y enfoque terapéutico," *MedUNAB*, vol. 18 (3), pp. 193–203, 2016.

[10] P. Yaşar, G. Ayaz, S. D. User, G. Güpür, and M. Muyan, "Molecular mechanism of estrogen–estrogen receptor signaling," *Reprod. Med. Biol.*, vol. 16, no. 1, p. 4, Jan. 2017, doi: 10.1002/RMB2.12006.

[11] R. Han *et al.*, "Estrogen promotes progression of hormone-dependent breast cancer through CCL2-CCR2 axis by upregulation of Twist via PI3K/AKT/NF-κB signaling," *Sci. Reports 2018 81*, vol. 8, no. 1, pp. 1–13, Jun. 2018, doi: 10.1038/s41598-018-27810-6.

[12] Y. Sukawa *et al.*, "HER2 Expression and PI3K-Akt Pathway Alterations in Gastric Cancer," *Digestion*, vol. 89, no. 1, pp. 12–17, Jan. 2014, doi: 10.1159/000356201.

[13] Z. Du and C. M. Lovly, "Mechanisms of receptor tyrosine kinase activation in cancer," *Mol. Cancer*, vol. 17, no. 1, Feb. 2018, doi: 10.1186/S12943-018-0782-4.

[14] "'Machine learning': ¿qué es y cómo funciona?" https://www.bbva.com/en/machine-learning-what-is-it-and-how-does-it-work/ (accessed May 31, 2022).

[15] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. Peter Campbell, "Introduction to Machine Learning, Neural Networks, and Deep Learning," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, 2020, doi: 10.1167/TVST.9.2.14.

[16] R. C. Deo, "Machine Learning in Medicine," *Circulation*, vol. 132, no. 20, p. 1920, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[17] "Medicina personalizada | NHGRI." https://www.genome.gov/es/genetics-glossary/Medicina-personalizada (accessed May 31, 2022).

[18] N. J. Schork, "ARTIFICIAL INTELLIGENCE AND PERSONALIZED MEDICINE," *Cancer Treat. Res.*, vol. 178, p. 265, 2019, doi: 10.1007/978-3-030-16391-4_11.

[19] "'Los hospitales no son conscientes aún de que en breve será mala praxis no utilizar inteligencia artificial' | @diariofarma." https://www.diariofarma.com/2022/04/20/los-hospitales-no-son-conscientes-de-que-en-breve-sera-mala-praxis-no-utilizar-inteligencia-artificial (accessed May 06, 2022).

[20] "La inteligencia artificial es el futuro de la salud | EL MUNDO." https://lab.elmundo.es/inteligencia-artificial/salud.html (accessed May 06, 2022).

[21] M. M. Alshammari, A. Almuhanna, and J. Alhiyafi, "Mammography Image-Based Diagnosis of Breast Cancer Using Machine Learning: A Pilot Study," *Sensors (Basel).*, vol. 22, no. 1, Jan. 2022, doi: 10.3390/S22010203.

[22] C. G. Yedjou, S. S. Tchounwou, R. A. Aló, R. Elhag, B. Mochona, and L. Latinwo, "Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification," *Int. J. Sci. Acad. Res.*, vol. 2, no. 1, p. 3081, 2021, Accessed: May 07, 2022. [Online]. Available: /pmc/articles/PMC8612371/.

[23] A. Tahmassebi *et al.*, "Impact of Machine Learning With Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients," *Invest. Radiol.*, vol. 54, no. 2, pp. 110–117, Feb. 2019, doi: 10.1097/RLI.0000000000000518.

[24] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," *Technol. Health Care*, vol. 24, no. 1, pp. 31–42, Jan. 2016, doi: 10.3233/THC-151071.

[25] "Deep Learning qué es y por qué es clave para la inteligencia artificial - Iberdrola." https://www.iberdrola.com/innovacion/deep-learning (accessed May 07, 2022).

[26] D. S. Kalafi EY, Nor NAM, Taib NA, Ganggayah MD, Town C, "Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data," *Folia Biol*, vol. 65(5–6), pp. 212-220., 2019.

[27] N. Naik *et al.*, "Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/S41467-020-19334-3.

[28] F. M. Alakwaa, K. Chaudhary, and L. X. Garmire, "Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data," *J. Proteome Res.*, vol. 17, no. 1, pp. 337–347, Jan. 2018, doi: 10.1021/ACS.JPROTEOME.7B00595/SUPPL_FILE/PR7B00595_SI_004.PDF.

[29] "The Cancer Genome Atlas Program - National Cancer Institute." https://www.cancer.gov/about-nci/organization/ccg/research/structural-

genomics/tcga (accessed Mar. 23, 2022).

[30] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,” *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015, doi: 10.5114/WO.2014.47136.

[31] “cBioPortal for Cancer Genomics.” https://www.cbioportal.org/ (accessed Nov. 03, 2021).

[32] “RStudio | Open source & professional software for data science teams - RStudio.” https://www.rstudio.com/ (accessed May 24, 2022).

[33] “Introduction to XGBoost in Python.” https://blog.quantinsti.com/xgboost-python/ (accessed May 24, 2022).

[34] “Homo sapiens Annotation Report.” https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/109.20211119/ (accessed Mar. 30, 2022).

[35] “GENCODE - Human Release 39 Statistics.” https://www.gencodegenes.org/human/stats_39.html (accessed Mar. 30, 2022).

[36] J. Li *et al.*, “Predicting breast cancer 5-year survival using machine learning: A systematic review,” *PLoS One*, vol. 16, no. 4, Apr. 2021, doi: 10.1371/JOURNAL.PONE.0250370.

[37] “Cáncer de mama: ahora se pueden detectar tumores de manera más rápida y eficaz | TN.” https://tn.com.ar/salud/noticias/2018/12/20/cancer-de-mama-ahora-se-pueden-detectar-tumores-de-manera-mas-rapida-y-eficaz/ (accessed Apr. 26, 2022).

[38] “La inteligencia artificial ya diagnostica enfermedades tan bien como los médicos.” https://www.lavanguardia.com/ciencia/cuerpo-humano/20180223/44950677766/inteligencia-artificial-machine-learning-diagnosticar-enfermedades-medicos-eficiencia.html (accessed Apr. 26, 2022).

[39] R. A. Freedman and A. H. Partridge, “Management of breast cancer in very young women,” *The Breast*, vol. 22, no. S2, pp. S176–S179, Aug. 2013, doi: 10.1016/J.BREAST.2013.07.034.

[40] C. K. Anders *et al.*, “Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression,” *J. Clin. Oncol.*, vol. 26, no. 20, pp. 3324–3330, 2008, doi: 10.1200/JCO.2007.14.2471.

[41] V. Theodorou, R. Stark, S. Menon, and J. S. Carroll, “GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility,” *Genome Res.*, vol. 23, no. 1, p. 12, Jan. 2013, doi: 10.1101/GR.139469.112.

[42] A. Albergaria *et al.*, “Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours,” *Breast Cancer Res.*, vol. 11, no. 3, Jun. 2009, doi: 10.1186/BCR2327.

[43] E. M. Martin, K. A. Orlando, K. Yokobori, and P. A. Wade, “The estrogen receptor/GATA3/FOXA1 transcriptional network: lessons learned from breast cancer,” *Curr. Opin. Struct. Biol.*, vol. 71, pp. 65–70, Dec. 2021, doi: 10.1016/J.SBI.2021.05.015.

[44] Z. Miao *et al.*, “Elevated transcription and glycosylation of B3GNT5 promotes breast cancer aggressiveness,” *J. Exp. Clin. Cancer Res. 2022 411*, vol. 41, no. 1, pp. 1–15, May 2022, doi: 10.1186/S13046-022-02375-5.

[45] X. Zhang *et al.*, “Interleukin enhancer-binding factor 2 promotes cell proliferation and DNA damage response in metastatic melanoma.,” *Clin. Transl. Med.*, vol. 11, no. 10, p. e608, Oct. 2021, doi: 10.1002/ctm2.608.

[46] “Cáncer de Mama en Hombres | Roche Pacientes.” https://rochepacientes.es/cancer/mama/en-hombres.html (accessed Mar. 26, 2022).

[47] A. Gucalp *et al.*, “Male breast cancer: a disease distinct from female breast cancer,” *Breast Cancer Res. Treat.*, vol. 173, no. 1, p. 37, Jan. 2019, doi: 10.1007/S10549-018-4921-9.

[48] K. J. Ruddy and E. P. Winer, “Male breast cancer: risk factors, biology, diagnosis, treatment, and survivorship,” *Ann. Oncol.*, vol. 24, pp. 1434–1443, 2013, doi: 10.1093/annonc/mdt025.

[49] S. Konduri, M. Singh, G. Bobustuc, R. Rovin, and A. Kassam, “Epidemiology of male breast cancer,” *Breast Off. J. Eur. Soc. Mastology*, vol. 54, p. 8, Dec. 2020, doi: 10.1016/J.BREAST.2020.08.010.

[50] E. E. Pala, Ü. Bayol, A. Özgüzer, Ü. Küçük, Ç. Y. Akdeniz, and Ö. Sezer, “Problems In Determining Her2 Status In Breast Carcinoma,” *J. Breast Heal.*, vol. 11, no. 1, p. 10, Jan. 2015, doi: 10.5152/TJBH.2014.2103.

[51] “Análisis de inmunohistoquímica.” https://www.breastcancer.org/es/pruebas-deteccion/analisis-inmunohistoquimica-ihq (accessed May 25, 2022).

Annex A

Formules avaluadrores del model:

- Exactitud: $\frac{Total\ positius+Total\ negatius\ predits\ correctament}{Total\ positius+Total\ negatius+Falsos\ positius+Falsos\ negatius}$
- Taxa d'error: $\frac{Falsos\ positius+Falsos\ negatius}{Total\ de\ casos}$
- Sensibilitat: $\frac{Positius\ predits\ correctament}{Total\ de\ positius}$
- Especificitat: $\frac{Negatius\ predits\ correctament}{Total\ de\ negatius}$
- F-score: $\frac{Positius\ predits\ correctament}{Positius\ predits\ correctament+\frac{1}{2}(falsos\ positius+falsos\ negatius)}$

Annex B

CODI: MANIPULACIÓ DE DADES I IMPLEMENTACIÓ D'XGBOOST

Marina Vilardell

3/6/2022

1. EXPORTACIÓ DE DADES

Llibreria que permetrà importar les dades.

```
library(cgdsr)
```

Creació de l'objecte

```
mycgds = CGDS("http://www.cbioportal.org/")
```

Amb la següent comanda es visualitza els estudis que hi ha dipositats al cBioportal. Només es mostren els 6 primers estudis.

```
cs <- getCancerStudies(mycgds)
head(cs)
##       cancer_study_id
## 1       paac_jhu_2014
## 2 mel_tsam_liang_2017
## 3     all_stjude_2015
## 4     all_stjude_2016
## 5       aml_ohsu_2018
## 6           laml_tcga
##                                                          name
## 1 Acinar Cell Carcinoma of the Pancreas (JHU, J Pathol 2014)
## 2                     Acral Melanoma (TGEN, Genome Res 2017)
## 3     Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2015)
## 4     Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2016)
## 5                 Acute Myeloid Leukemia (OHSU, Nature 2018)
## 6             Acute Myeloid Leukemia (TCGA, Firehose Legacy)
##
description
```

```
## 1
Whole exome sequencing of 23 surgically resected pancreatic
carcinomas with acinar differentiation and their matched normals.
## 2
Whole exome sequencing and transcriptome analysis of 34 Acral
Melanoma patients (33 with matched normals).
## 3
Comprehensive profiling of infant MLL-rearranged acute
lymphoblastic leukemia (MLL-R ALL)
## 4
Whole-genome and/or whole-exome sequencing of ERG-altered B-ALL
tumor/normal pairs.
## 5
Whole-exome sequencing of 672 acute myeloid leukemia samples (with
454 matched normals) from the Beat AML program.
## 6 TCGA Acute Myeloid Leukemia. Source data from <A
HREF="http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/
LAML/20160128/">GDAC Firehose</A>. Previously known as TCGA
Provisional.
```

```
colnames(cs)
```

```
## [1] "cancer_study_id" "name"            "description"
```

Per a cada estudi, hi consta un identificador, el nom complet i una descripció.

Se seleccionen tots aquells estudis, en els quals al seu identificador hi ha la paraula ‘tcga’. Se’n mostren els 20 primers.

```
estudi<-cs$cancer_study_id
estudi[grep("tcga",estudi)][1:20]
```

```
##  [1] "laml_tcga"                    "laml_tcga_pub"
##  [3] "laml_tcga_pan_can_atlas_2018" "acc_tcga"
##  [5] "acc_tcga_pan_can_atlas_2018"  "sarc_tcga_pub"
##  [7] "blca_msk_tcga_2020"           "blca_tcga_pub_2017"
##  [9] "blca_tcga"                    "blca_tcga_pub"
## [11] "blca_tcga_pan_can_atlas_2018" "lgg_tcga"
## [13] "lgg_tcga_pan_can_atlas_2018"  "brca_tcga_pub2015"
## [15] "brca_tcga"                    "brca_tcga_pub"
## [17] "brca_tcga_pan_can_atlas_2018"
"cesc_tcga_pan_can_atlas_2018"
## [19] "cesc_tcga"                    "chol_tcga"
```

L’estudi del qual obtindrem el Data-Set de càncer de mama és el ‘brca_tgca’.

```
dades_estudi='brca_tcga'
cs[cs$cancer_study_id=='brca_tcga',]
```

```
##    cancer_study_id
name
```

```
## 50       brca_tcga Breast Invasive Carcinoma (TCGA, Firehose
Legacy)
##
description
## 50 TCGA Breast Invasive Carcinoma. Source data from <A
HREF="http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/
BRCA/20160128/">GDAC Firehose</A>. Previously known as TCGA
Provisional.
```

Llista de casos:

```
llcasos<-getCaseLists(mycgds,dades_estudi)[,c(1:2)]
colnames(llcasos)
```

```
## [1] "case_list_id"   "case_list_name"
```

```
llcasos$case_list_id
```

```
##  [1] "brca_tcga_all"
##  [2] "brca_tcga_3way_complete"
##  [3] "brca_tcga_cna"
##  [4] "brca_tcga_methylation_all"
##  [5] "brca_tcga_methylation_hm27"
##  [6] "brca_tcga_methylation_hm450"
##  [7] "brca_tcga_mrna"
##  [8] "brca_tcga_rna_seq_v2_mrna"
##  [9] "brca_tcga_cnaseq"
## [10] "brca_tcga_sequenced"
## [11] "brca_tcga_phosphoprotein_quantification"
## [12] "brca_tcga_protein_quantification"
## [13] "brca_tcga_rppa"
```

Es parteix de totes les mostres disponibles d'aquest estudi, per tant, se seleccionen tots els casos, que corresponen a 'brca_tcga_all'.

Finalment, s'obté un data frame que conté informació clínica de tots els pacients:

```
brca<-getClinicalData(mycgds,'brca_tcga_all')
colnames(brca) #llista de totes les variables clíniques
```

```
##   [1] "AGE"
##   [2] "AJCC_METASTASIS_PATHOLOGIC_PM"
##   [3] "AJCC_NODES_PATHOLOGIC_PN"
##   [4] "AJCC_PATHOLOGIC_TUMOR_STAGE"
##   [5] "AJCC_STAGING_EDITION"
##   [6] "AJCC_TUMOR_PATHOLOGIC_PT"
##   [7] "BRACHYTHERAPY_TOTAL_DOSE_POINT_A"
##   [8] "CANCER_TYPE"
##   [9] "CANCER_TYPE_DETAILED"
##  [10] "CENT17_COPY_NUMBER"
##  [11] "DAYS_TO_COLLECTION"
```

```
##  [12] "DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS"
##  [13] "DFS_MONTHS"
##  [14] "DFS_STATUS"
##  [15] "DISEASE_CODE"
##  [16] "ER_POSITIVITY_SCALE_OTHER"
##  [17] "ER_POSITIVITY_SCALE_USED"
##  [18] "ER_STATUS_BY_IHC"
##  [19] "ER_STATUS_IHC_PERCENT_POSITIVE"
##  [20] "ETHNICITY"
##  [21] "FIRST_SURGICAL_PROCEDURE_OTHER"
##  [22] "FORM_COMPLETION_DATE"
##  [23] "FRACTION_GENOME_ALTERED"
##  [24] "HER2_AND_CENT17_CELLS_COUNT"
##  [25] "HER2_AND_CENT17_SCALE_OTHER"
##  [26] "HER2_CENT17_RATIO"
##  [27] "HER2_COPY_NUMBER"
##  [28] "HER2_FISH_METHOD"
##  [29] "HER2_FISH_STATUS"
##  [30] "HER2_IHC_PERCENT_POSITIVE"
##  [31] "HER2_IHC_SCORE"
##  [32] "HER2_POSITIVITY_METHOD_TEXT"
##  [33] "HER2_POSITIVITY_SCALE_OTHER"
##  [34] "HISTOLOGICAL_DIAGNOSIS"
##  [35] "HISTOLOGICAL_SUBTYPE"
##  [36] "HISTORY_NEOADJUVANT_TRTYN"
##  [37] "HISTORY_OTHER_MALIGNANCY"
##  [38] "ICD_10"
##  [39] "ICD_O_3_HISTOLOGY"
##  [40] "ICD_O_3_SITE"
##  [41] "IHC_HER2"
##  [42] "IHC_SCORE"
##  [43] "INFORMED_CONSENT_VERIFIED"
##  [44] "INITIAL_PATHOLOGIC_DX_YEAR"
##  [45] "IS_FFPE"
##  [46] "LYMPH_NODES_EXAMINED"
##  [47] "LYMPH_NODES_EXAMINED_HE_COUNT"
##  [48] "LYMPH_NODES_EXAMINED_IHC_COUNT"
##  [49] "LYMPH_NODE_EXAMINED_COUNT"
##  [50] "MARGIN_STATUS_REEXCISION"
##  [51] "MENOPAUSE_STATUS"
##  [52] "METASTATIC_SITE_OTHER"
##  [53] "METASTATIC_SITE_PATIENT"
##  [54] "METASTATIC_TUMOR_INDICATOR"
##  [55] "METHOD_OF_INITIAL_SAMPLE_PROCUREMENT"
##  [56] "METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_OTHER"
##  [57] "MICROMET_DETECTION_BY_IHC"
##  [58] "MUTATION_COUNT"
##  [59] "NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT"
##  [60] "NTE_CENT_17_HER2_RATIO"
##  [61] "NTE_ER_IHC_INTENSITY_SCORE"
```

```
##  [62] "NTE_ER_STATUS"
##  [63] "NTE_ER_STATUS_IHC_POSITIVE"
##  [64] "NTE_HER2_FISH_STATUS"
##  [65] "NTE_HER2_POSITIVITY_IHC_SCORE"
##  [66] "NTE_HER2_STATUS"
##  [67] "NTE_HER2_STATUS_IHC_POSITIVE"
##  [68] "NTE_PR_IHC_INTENSITY_SCORE"
##  [69] "NTE_PR_STATUS_BY_IHC"
##  [70] "NTE_PR_STATUS_IHC_POSITIVE"
##  [71] "OCT_EMBEDDED"
##  [72] "ONCOTREE_CODE"
##  [73] "OS_MONTHS"
##  [74] "OS_STATUS"
##  [75] "OTHER_PATIENT_ID"
##  [76] "OTHER_SAMPLE_ID"
##  [77] "PATHOLOGY_REPORT_FILE_NAME"
##  [78] "PATHOLOGY_REPORT_UUID"
##  [79] "PATH_MARGIN"
##  [80] "PHARMACEUTICAL_TX_ADJUVANT"
##  [81] "PRIMARY_SITE_PATIENT"
##  [82] "PROJECT_CODE"
##  [83] "PROSPECTIVE_COLLECTION"
##  [84] "PR_POSITIVITY_DEFINE_METHOD"
##  [85] "PR_POSITIVITY_IHC_INTENSITY_SCORE"
##  [86] "PR_POSITIVITY_SCALE_OTHER"
##  [87] "PR_POSITIVITY_SCALE_USED"
##  [88] "PR_STATUS_BY_IHC"
##  [89] "PR_STATUS_IHC_PERCENT_POSITIVE"
##  [90] "RACE"
##  [91] "RADIATION_TREATMENT_ADJUVANT"
##  [92] "RETROSPECTIVE_COLLECTION"
##  [93] "SAMPLE_COUNT"
##  [94] "SAMPLE_INITIAL_WEIGHT"
##  [95] "SAMPLE_TYPE"
##  [96] "SAMPLE_TYPE_ID"
##  [97] "SEX"
##  [98] "SITE_OF_TUMOR_TISSUE"
##  [99] "SOMATIC_STATUS"
## [100] "STAGING_SYSTEM"
## [101] "STAGING_SYSTEM_OTHER"
## [102] "SURGERY_FOR_POSITIVE_MARGINS"
## [103] "SURGERY_FOR_POSITIVE_MARGINS_OTHER"
## [104] "SURGICAL_PROCEDURE_FIRST"
## [105] "TISSUE_SOURCE_SITE"
## [106] "TMB_NONSYNONYMOUS"
## [107] "TUMOR_STATUS"
## [108] "VIAL_NUMBER"
```

2. ANÀLISI DE LES VARIABLES CLÍNIQUES

2.1 SEXE

```
table(brca$SEX)
```

```
##         Female   Male
##      4    1092     12
```

S'eliminen els pacients que no tenen definit el sexe.

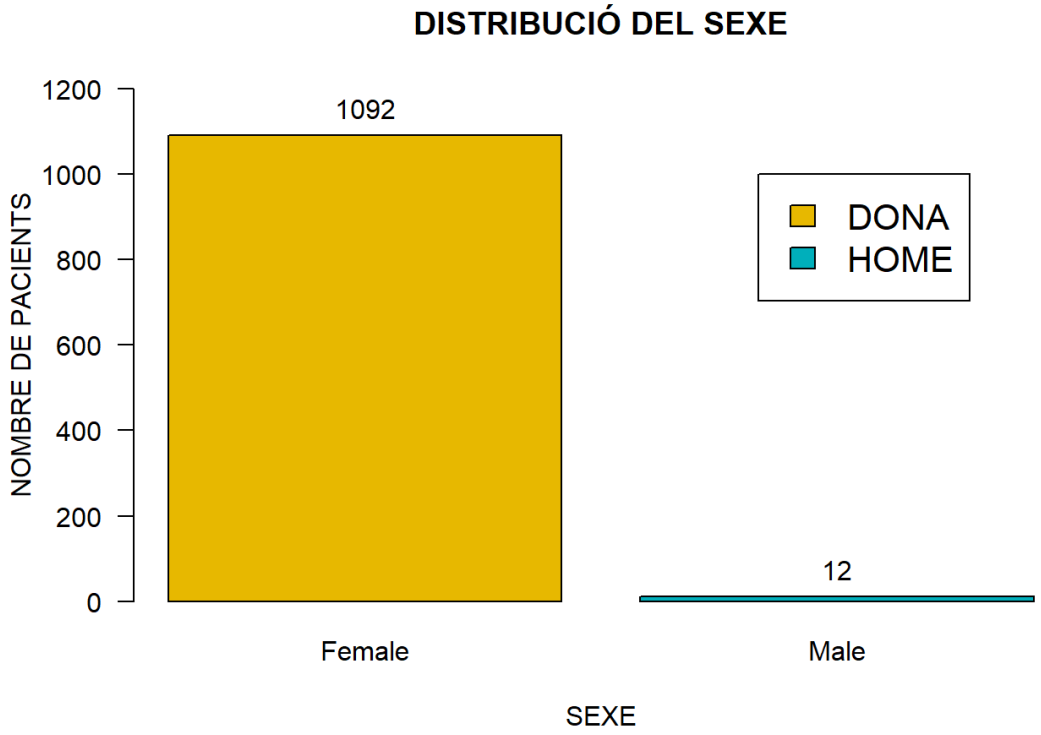
```
brca<-brca[brca$SEX!='',]
table(brca$SEX)
```

```
## Female   Male
##   1092     12
```

```
freq<-table(brca$SEX)
```

Gràfic de la distribució del sexe:

```
barplotsexe<-barplot(table(brca$SEX),
                     ylim = c(0,1200),
                     col=c('#E7B800','#00AFBB'),
                     xlab='SEXE',
                     ylab='NOMBRE DE PACIENTS',
                     main='DISTRIBUCIÓ DEL SEXE',
                     yaxt = "n")
axis(2, at = c(0,200,400,600,800,1000,1200),las=1)
text(barplotsexe,freq, labels=freq, pos=3)
legend(x = 1.7, y = 1000,
       legend = c('DONA','HOME'),
       fill=c('#E7B800','#00AFBB'),
       cex=1.3)
```



2.1.1 SEXE SENSE HOMES

S'eliminen els homes, doncs no es tenen en compte.

```
brca_dones<-brca[brca$SEX!='Male',]
table(brca_dones$SEX)
```

```
## Female
##   1092
```

A partir d'aquest punt, només es treballa amb dades corresponents al sexe femení.

2.2 EDAT

```
summary(brca_dones$AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   26.00   48.50   58.00   58.31   67.00   90.00       1
```

S'eliminen les pacients que no tenen definida l'edat (missings).

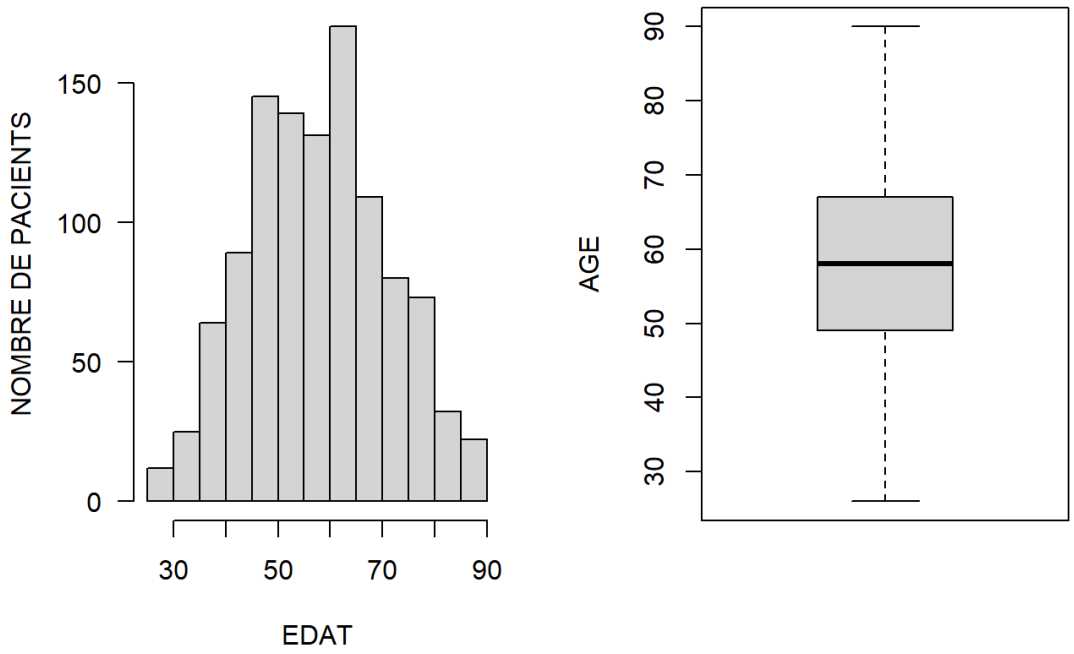
```
brca_dones_edat<-brca_dones[!is.na(brca_dones$AGE),]
summary(brca_dones_edat$AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   48.50   58.00   58.31   67.00   90.00
```

Gràfic de la distribució de l'edat.

```
par(mfrow=c(1,2))
```

```
hist(brca_dones_edat$AGE,
    xlab='EDAT',
    las=1,
    ylab='NOMBRE DE PACIENTS',
    main='',
    col = 'gray83')
boxplot(brca$AGE,
       ylab='AGE')
```



2.3 DISTRIBUCIÓ DE LA VARIABLE RESPOSTA, ER_STATUS_BY_IHC (RECEPTOR ESTROGEN)

```
table(brca_dones_edat$ER_STATUS_BY_IHC)
```

```
##                Indeterminate      Negative      Positive
##             49             2           239           801
```

S’eliminen les pacients que no tenen la variable resposta codificada com a negativa o positiva.

```
brca_dones_f<-
brca_dones_edat[brca_dones_edat$ER_STATUS_BY_IHC!='Indeterminate' &
brca_dones_edat$ER_STATUS_BY_IHC!='',]
table(brca_dones_f$ER_STATUS_BY_IHC)
```

```
## Negative Positive
##      239      801
```

2.4 DISTRIBUCIÓ DE L’EDAT I LA VARIABLE RESPOSTA

Llibreries necessàries:

```
library(ggplot2)
library(ggpubr)
```

Resum de la variable edat per als pacients negatius i positius del receptor d’estrogen.

```
tapply(brca_dones_f$AGE,brca_dones_f$ER_STATUS_BY_IHC,summary)
```

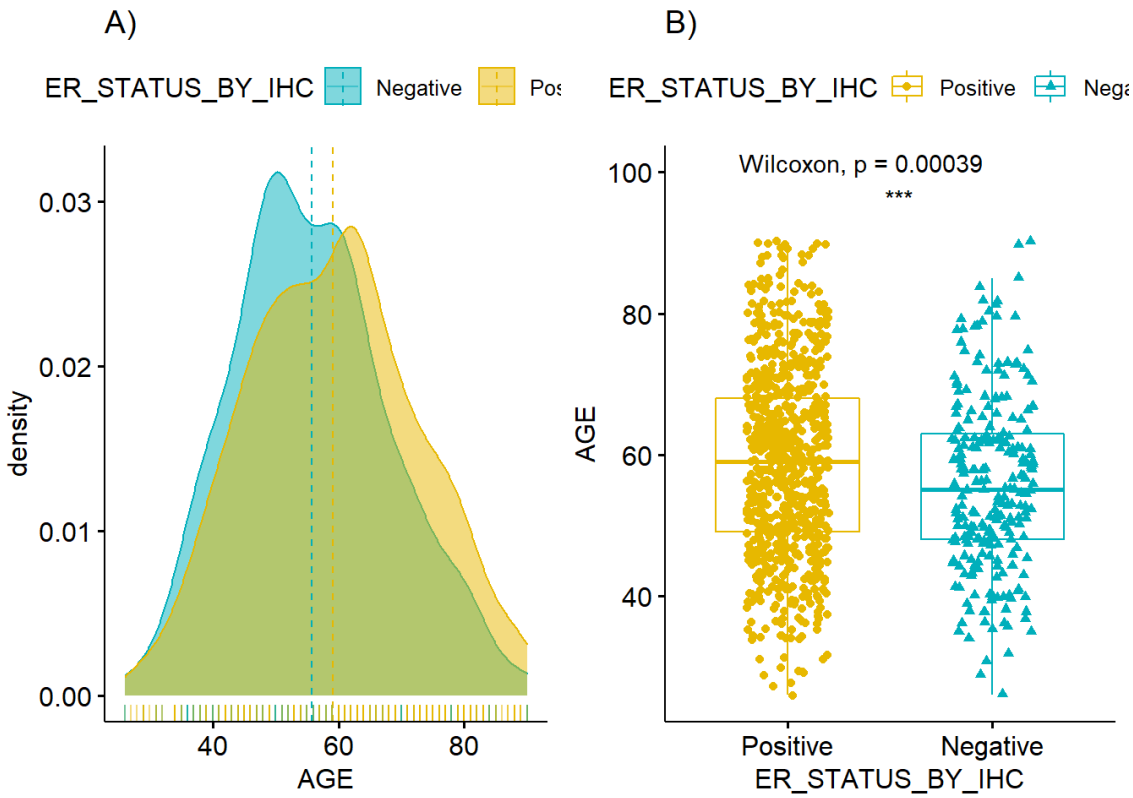
```
## $Negative
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   48.00   55.00   55.81   63.00   90.00
##
## $Positive
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   49.00   59.00   59.18   68.00   90.00
```

Mitjana d’edat de les pacients negatives: 55,81 anys.

Mitjana d’edat de les pacients positives: 59,18 anys.

```
a<-ggdensity(brca_dones_f, x='AGE',
          add = "mean", rug = TRUE,
          main='A)',
          color = "ER_STATUS_BY_IHC", fill = "ER_STATUS_BY_IHC",
          palette = c("#00AFBB", "#E7B800"))
b<-ggboxplot(brca_dones_f, x='ER_STATUS_BY_IHC', y='AGE',
          main='B)',
          color = "ER_STATUS_BY_IHC", palette =c("#E7B800", "#00AFBB"),
          add = "jitter", shape = "ER_STATUS_BY_IHC")
b<-b+stat_compare_means(label = "p.signif", label.y = 95, label.x=1.5)+
      stat_compare_means(label.y = 100)
age_er__plot<-ggarrange(a,b)

age_er__plot
```



2.4.1 PROVA DE NORMALITAT DE LES DADES:

H0: les dades de la variable EDAT segueixen una distribució normal.

H1: les dades de la variable EDAT no segueixen una distribució normal.

```
shapiro.test(brca_dones_f$AGE)
```

```
##  Shapiro-Wilk normality test
##
## data:  brca_dones_f$AGE
## W = 0.99193, p-value = 1.856e-05
```

p-value és més petit que 0.05, per tant, les dades no segueixen una distribució normal.

2.4.2 PROVA D'IGUALTAT DE MITJANES:

H0: les mitjanes d'edat de les pacients ER positives i negatives són iguals.

H1: les mitjanes d'edat de les pacients són diferents.

```
wilcox.test(brca_dones_f$AGE~brca_dones_f$ER_STATUS_BY_IHC)
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  brca_dones_f$AGE by brca_dones_f$ER_STATUS_BY_IHC
## W = 81282, p-value = 0.0003946
```

```
## alternative hypothesis: true location shift is not equal to 0
```

p-value és menor que 0.05, per tant, la hipòtesi nul·la és falsa, i s'accepta la H1. Hi ha diferències significatives entre la mitjana d'edat dels pacients positius i els negatius. Com a resultat, l'edat és un factor a tenir en compte.

3. PREPARACIÓ DE LES DADES DE TRABSCRIPTÒMICA

Carregar el fitxer el qual conté les dades de RNA, ja en Z-scores.

```
 data_RNA_Seq_v2_expression_median <-
read.delim("C:/Users/MARINA/Desktop/4T
BIOTEC/TFG/brca_tcga/data_RNA_Seq_v2_mRNA_median_Zscores.txt")
```

data_RNA_Seq_v2_expression_median[1:10,1:5] #Es mostren les 10 primeres files i les 5 primeres columnes de la base de dades.

```
##    Hugo_Symbol Entrez_Gene_Id TCGA.3C.AAAU.01 TCGA.3C.AALI.01
TCGA.3C.AALJ.01
## 1     UBE2Q2P2      100134869          0.9689          1.7786
0.2949
## 2      HMGB1P1          10357              NA              NA
NA
## 3    LOC155060         155060              NA              NA
NA
## 4     RNU12-2P          26823              NA              NA
NA
## 5         SSX9         280660         -0.0752          0.3153
-0.0752
## 6        EZHIP         340602         -0.0373          6.6560
-0.1184
## 7       EFCAB8         388795         -0.5721          2.0252
0.7840
## 8      SRP14P1         390284              NA              NA
NA
## 9    LOC391343         391343              NA              NA
NA
## 10     TRIM75P         391714              NA              NA
NA
```

Cal identificar quins pacients es troben a ambdues bases de dades (variables clíniques i RNA):

El nom de cada columna correspon a l'identificador del pacient.

```
n<-colnames(data_RNA_Seq_v2_expression_median)
```

El nom de cada fila correspon a l'identificador del pacient.

```
b<-rownames(brca_dones_f)
```

Intersectar els pacients.

```
length(intersect(b,n))
```

```
## [1] 1037
```

Hi ha 1037 pacients que es troben en ambdues bases de dades.

Determinar quins són els pacients es troben a la base de dades on hi ha informació RNA i que no es troben a la base de dades de les variables clíniques, prèviament manipulada.

```
d<-setdiff(n,b)
d
```

```
##  [1] "Hugo_Symbol"     "Entrez_Gene_Id"  "TCGA.A1.A0SM.01"
"TCGA.A7.A0CH.01"
##  [5] "TCGA.A8.A085.01" "TCGA.AC.A2FM.01" "TCGA.AC.A62V.01"
"TCGA.AO.A1KQ.01"
##  [9] "TCGA.AQ.A54O.01" "TCGA.AR.A1AV.01" "TCGA.B6.A0I2.01"
"TCGA.B6.A0I8.01"
## [13] "TCGA.B6.A0I9.01" "TCGA.BH.A0B4.01" "TCGA.BH.A0DD.01"
"TCGA.BH.A18R.01"
## [17] "TCGA.BH.A203.01" "TCGA.BH.A204.01" "TCGA.BH.A208.01"
"TCGA.C8.A12K.01"
## [21] "TCGA.C8.A12Y.01" "TCGA.C8.A275.01" "TCGA.C8.A8HR.01"
"TCGA.D8.A1XS.01"
## [25] "TCGA.E2.A14W.01" "TCGA.E9.A1N3.01" "TCGA.E9.A1QZ.01"
"TCGA.E9.A1R0.01"
## [29] "TCGA.E9.A1R3.01" "TCGA.E9.A1R4.01" "TCGA.E9.A1R5.01"
"TCGA.E9.A1R6.01"
## [33] "TCGA.E9.A1R7.01" "TCGA.E9.A1RA.01" "TCGA.E9.A1RB.01"
"TCGA.E9.A1RC.01"
## [37] "TCGA.E9.A1RD.01" "TCGA.E9.A1RE.01" "TCGA.E9.A1RF.01"
"TCGA.E9.A1RG.01"
## [41] "TCGA.E9.A1RH.01" "TCGA.E9.A1RI.01" "TCGA.E9.A226.01"
"TCGA.E9.A228.01"
## [45] "TCGA.E9.A229.01" "TCGA.E9.A243.01" "TCGA.E9.A244.01"
"TCGA.E9.A245.01"
## [49] "TCGA.E9.A247.01" "TCGA.E9.A248.01" "TCGA.E9.A249.01"
"TCGA.E9.A24A.01"
## [53] "TCGA.E9.A2JS.01" "TCGA.E9.A2JT.01" "TCGA.E9.A3HO.01"
"TCGA.E9.A3QA.01"
## [57] "TCGA.E9.A5UO.01" "TCGA.E9.A5UP.01" "TCGA.EW.A1PD.01"
"TCGA.EW.A6SA.01"
## [61] "TCGA.OL.A66H.01" "TCGA.PL.A8LV.01" "TCGA.PL.A8LX.01"
"TCGA.PL.A8LY.01"
```

```
## [65] "TCGA.PL.A8LZ.01"
```

```
eliminar<-d[2:65] #Seleccionar els pacients que caldrà eliminar a
la base de dades RNA. No s'elimina 'Hugo_Symbol'.
```

Caldrà eliminar tots aquests pacients de la base de dades RNA.

Llibreria que permet manipular data frames.

```
library(dplyr)
```

Eliminar els pacients de la base de dades RNA.

```
dats<-select(data_RNA_Seq_v2_expression_median, -eliminar)
```

Eliminar els gens duplicats.

```
elim<-which(duplicated(dats$Hugo_Symbol))
data3<-dats[-elim,]
```

Els noms dels gens passen a ser el nom de cada fila, en comptes de trobar-se en una columna.

```
rownames(data3)<-data3[,1]
data3<-data3[,-1]
```

Canvi de files per columnes.

```
data_RNA<-data.frame(t(data3))
```

data_RNA[1:5,1:5] #Es mostren les 5 primeres files (pacients) i les 5 primeres columnes (gens).

```
##                 UBE2Q2P2 HMGB1P1 LOC155060 RNU12.2P    SSX9
## TCGA.3C.AAAU.01   0.9689      NA         NA       NA -0.0752
## TCGA.3C.AALI.01   1.7786      NA         NA       NA  0.3153
## TCGA.3C.AALJ.01   0.2949      NA         NA       NA -0.0752
## TCGA.3C.AALK.01   0.6318      NA         NA       NA -0.0752
## TCGA.4H.AAAK.01   1.2416      NA         NA       NA -0.0752
```

Determinar els pacients que es troben a la base de dades on hi ha les variables clíniques i que no es troben a la base de dades RNA.

```
f<-setdiff(b,n)
```

f

```
## [1] "TCGA.AC.A5EI.01" "TCGA.AR.A0U1.01" "TCGA.A7.A0DC.01"
```

```
brca_transpose<-data.frame(t(brca_dones_f))
```

Eliminar els pacients de la base de dades de les variables clíniques.

```
brca_b<-select(brca_transpose,-f)
brca_dones_ff<-data.frame(t(brca_b))
```

Ordenar els pacients per ordre alfabètic segons el seu identificador.

```
brca_ordenat_patients<-
brca_dones_ff[order(rownames(brca_dones_ff)),]
```

brca_ordenat_patients[1:5,1:5]#Es mostren les 5 primeres files (pacients) i les 5 primeres columnes (variables).

```
##                 AGE AJCC_METASTASIS_PATHOLOGIC_PM
AJCC_NODES_PATHOLOGIC_PN
## TCGA.3C.AAAU.01  55                              MX
NX
## TCGA.3C.AALI.01  50                              M0
N1a
## TCGA.3C.AALJ.01  62                              M0
N1a
## TCGA.3C.AALK.01  52                              M0
N0 (i+)
## TCGA.4H.AAAK.01  50                              M0
N2a
##                 AJCC_PATHOLOGIC_TUMOR_STAGE AJCC_STAGING_EDITION
## TCGA.3C.AAAU.01                     Stage X                   6th
## TCGA.3C.AALI.01                   Stage IIB                   6th
## TCGA.3C.AALJ.01                   Stage IIB                   7th
## TCGA.3C.AALK.01                    Stage IA                   7th
## TCGA.4H.AAAK.01                  Stage IIIA                   7th
```

Els pacients d’ambdues bases de dades ja es troben en el mateix ordre.

Afegir la variable resposta a la base de dades RNA, on hi ha informació sobre l’expressió de gens per a cada pacient.

```
data_RNA$A0ER_<-brca_ordenat_patients$ER_STATUS_BY_IHC
```

Ordenar els gens.

```
data_RNA_gens_ordenats<-data_RNA[order(colnames(data_RNA))]
```

data_RNA_gens_ordenats[1:5,1:5] #Es mostren les 5 primeres files (gens) i les 5 primeres columnes.

```
##                   A0ER_    A1BG A1BG.AS1    A1CF     A2M
## TCGA.3C.AAAU.01 Positive  0.0103       NA -0.0589 -0.6589
## TCGA.3C.AALI.01 Positive  0.2155       NA -0.0589 -0.5208
## TCGA.3C.AALJ.01 Positive  1.1618       NA  0.1349 -0.4220
```

```
## TCGA.3C.AALK.01 Positive -0.0206       NA -0.0589 -0.2569
## TCGA.4H.AAAK.01 Positive  0.3759       NA  0.0320 -0.3640
```

Determinar si hi ha missings.

```
datatt<-data.frame(t(data_RNA_gens_ordenats))#canvi de files per
columns
sum(is.na(datatt))


## [1] 2086444
```

Eliminar els gens que tenen missings.

```
datafin<-na.omit(datatt)
```

sum(is.na(datafin))#Es comprova que s'hagin eliminat correctament

```
## [1] 0

data2<-data.frame(t(datafin))
```

data2[1:5,1:5] #Es mostren les 5 primeres files i columnes.

```
##                   A0ER_    A1BG    A1CF     A2M   A2ML1
## TCGA.3C.AAAU.01 Positive  0.0103 -0.0589 -0.6589 -0.2282
## TCGA.3C.AALI.01 Positive  0.2155 -0.0589 -0.5208 -0.2227
## TCGA.3C.AALJ.01 Positive  1.1618  0.1349 -0.4220 -0.2307
## TCGA.3C.AALK.01 Positive -0.0206 -0.0589 -0.2569 -0.2277
## TCGA.4H.AAAK.01 Positive  0.3759  0.0320 -0.3640 -0.2244
```

Ordenar els pacients segons el seu receptor d'estrogen. Primer els negatius i llavors els positius.

```
data4<-data2[order(data2$A0ER_),]
table(data4$A0ER_)
```

```
## Negative Positive
##      238      799
```

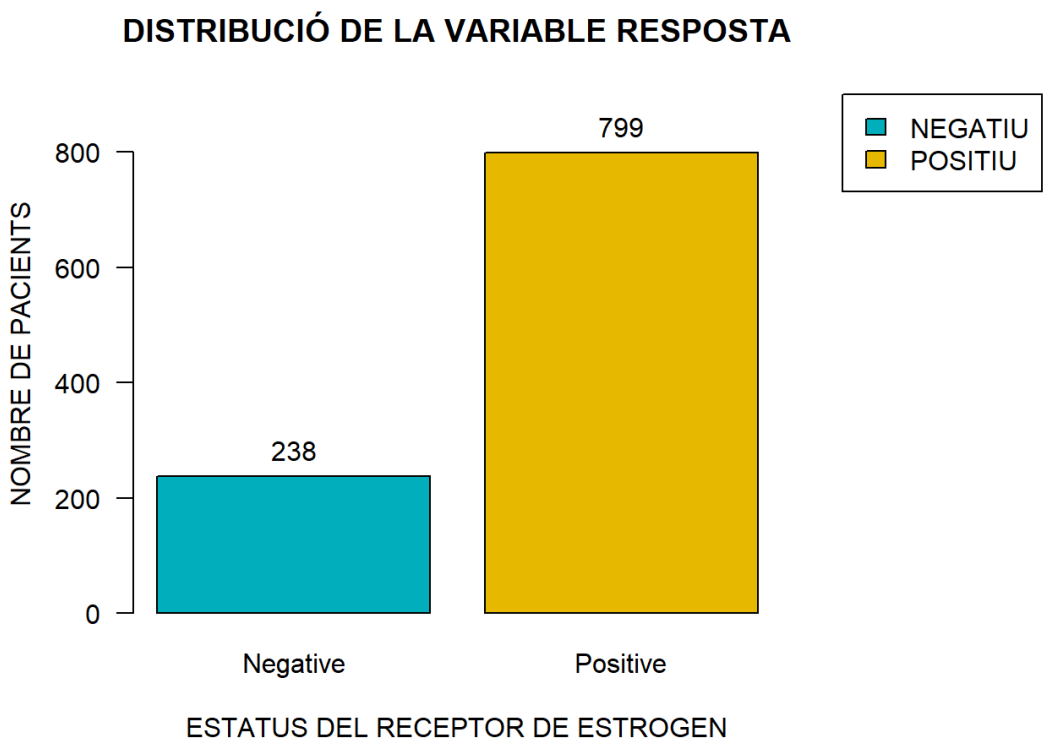
Finalment, després d'eliminar els pacients que no es troben en les dues bases de dades, i haver eliminat missings, es consta de 238 pacients negatius i 799 de positius.

DISTRIBUCIÓ FINAL DE LA VARIABLE RESPOSTA (RECEPTOR D'ESTROGEN):

```
par(mar=c(5,5,4,10))
```

```
frequer<-(table(data4$A0ER_))
barpl1<-barplot(table(data4$A0ER_),
     ylab='NOMBRE DE PACIENTS',
     main='DISTRIBUCIÓ DE LA VARIABLE RESPOSTA',
     xlab='ESTATUS DEL RECEPTOR DE ESTROGEN',
     col=c('#00AFBB','#E7B800'),
```

```
      las=1,ylim = c(0,900))
text(barpl1,frequer, labels=(table(data4$A0ER_)), pos=3)
legend(x='topright', inset = c(-0.4,0),
      legend=c('NEGATIU','POSITIU'),
      fill=c('#00AFBB','#E7B800'),
      xpd=TRUE)
```



Es crea un nou data frame amb només les dades Z-scores.

```
countsZ<-data4[,2:18422]
```

L'estructura de les dades és de tipus caràcter. Es necessiten com a numériques per a l'anàlisis d'expressió diferencial.

```
countsZ <- mutate_all(countsZ, function(x)
as.numeric(as.character(x)))

countsZ<-data.frame(t(countsZ))
```

countsZ[1:5,1:5]#Es mostren les 5 primeres files i columnes.

```
##         TCGA.A1.A0SH.01 TCGA.A1.A0SK.01 TCGA.A1.A0SO.01
TCGA.A1.A0SP.01
## A1BG           -0.0751         -0.5521          0.4991         -
0.8060
## A1CF           -0.0589         -0.0589         -0.0589         -
0.0589
```

```
## A2M            -0.0203         -0.7842         -0.7917         -
0.7018
## A2ML1          -0.2281         -0.1846         -0.1887
1.3511
## A4GALT         -0.0570         -0.9859         -1.1335
1.4192
##        TCGA.A2.A04P.01
## A1BG           -0.4395
## A1CF           -0.0589
## A2M            -0.3865
## A2ML1           2.8186
## A4GALT         -0.7416
```

```
dim(countsZ)
```

```
## [1] 18421  1037
```

Cada fila correspon als Z-scores d'un gen en concret.

4. ANÀLISIS D'EXPRESSIÓ DIFERENCIAL

Es requereix la llibreria Limma per dur a terme l'anàlisis d'expressió diferencial.

```
library(limma)
```

Creació d'un vector amb ER- i ER+, ordenats segons els pacients del data frame data4. Primer hi ha els pacients negatius i llavors els positius. Les mostres han de tenir el mateix ordre que el vector cond, molt important!

```
cond<-as.factor(c(rep("N",238),rep("P",799)))
```

```
design <-model.matrix(~0 + cond)
colnames(design) <- gsub("cond", "", colnames(design))
cont.matrix <- makeContrasts(con1=P-N, levels = design)
fit <- lmFit(countsZ, design)
fit <- eBayes(fit, trend=TRUE)
```

Resum.

```
summary(decideTests(fit, method = "separate"))
```

```
##            N     P
## Down    5430  4611
## NotSig  5535 10704
## Up      7456  3106
```

Llista ordenada dels gens diferenciats entre pacients negatius i positius.

```
top.Diff <- topTable(fit, n = Inf, coef = 1, adjust = "fdr")
```

S'extreuen els gens més significatius després de corregir per comparacions múltiples. S'ajusta el P-valor a 0.05, 0.01 i 0.001

```
res <- top.Diff[top.Diff$adj.P.Val<0.05,]
```

dim(res)

```
## [1] 12886     6
```

```
res2 <- top.Diff[top.Diff$adj.P.Val<0.01,]
```

dim(res2)

```
## [1] 11598     6
```

```
res3<- top.Diff[top.Diff$adj.P.Val<0.001 &
abs(top.Diff$logFC)>0.95,]
```

dim(res3)

```
## [1] 1874    6
```

```
res3[1:15,] #Es mostren els 15 primers gens
```

```
##              logFC    AveExpr         t       P.Value
adj.P.Val        B
## ILF2      3.579442  1.3912745  27.37574 1.055474e-124 1.944288e-
120 273.9581
## B3GNT5    2.348810  0.4110344  26.81570 7.797989e-121 7.182338e-
117 265.0907
## DESI2     2.783041  1.0357054  26.17658 1.945094e-116 1.194353e-
112 255.0119
## FAM171A1  2.712459  0.5029082  25.79997 7.410620e-114 3.412776e-
110 249.0958
## FOXA1    -1.427812 -0.2507851 -25.78453 9.451169e-114 3.482000e-
110 248.8537
## UCK2      2.827927  0.9382346  25.72730 2.328139e-113 7.147774e-
110 247.9562
## ZBTB4    -1.533485 -0.7443442 -25.15384 1.899173e-109 4.997809e-
106 238.9900
## ANP32E    3.607203  1.1135186  25.04731 1.006710e-108 2.318075e-
105 237.3296
## HDGF      3.039430  1.2578967  24.96952 3.399327e-108 6.957668e-
105 236.1182
## PDE7A     2.319838  0.6527601  24.90749 8.964840e-108 1.651413e-
104 235.1529
## SFT2D2    2.530334  0.7496463  24.86749 1.674643e-107 2.804419e-
104 234.5308
```

```
## RABEP1   -1.258442 -0.4849750 -24.55962 2.037961e-105 3.128440e-
102 229.7509
## YEATS2    2.269913  0.5341045  24.48418 6.593240e-105 9.342622e-
102 228.5821
## SUV39H2   2.937077  0.7079760  24.22770 3.545651e-103 4.665316e-
100 224.6152
## GATA3    -1.313239 -0.2417125 -24.19011 6.351958e-103 7.800628e-
100 224.0348
```

```
res3[308,]#LogFC i p-valor del gen ESR1, que codifica pel receptor
d'estrogen.
```

```
##           logFC   AveExpr         t      P.Value    adj.P.Val
B
```

```
## ESR1 -0.9545164 -0.123897 -17.61913 5.265509e-61 3.139027e-59
127.9765
```

Creació d'un nou data frame amb només la informació dels gens expressats diferencialment.

```
countsDEg<- countsZ[rownames(res3),]
```

countsDEg<-data.frame(t(countsDEg))

Comprovem que s'hagin seleccionat tots els gens expresats diferencialment.

```
dim(countsDEg)
```

```
## [1] 1037 1874
```

Ordenar els pacients.

```
counts_ord<-countsDEg[order(rownames(countsDEg)),]
```

S'afegeix la variable resposta (receptor d'estrogen) i l'edat:

```
counts_ord$A00ER<-brca_ordenat_patients$ER_STATUS_BY_IHC
```

counts_ord$A00AGE<-brca_ordenat_patients$AGE

Ordenar les columnes (gens) alfabèticament.

```
counts_fin<-counts_ord[order(colnames(counts_ord))]
```

counts_fin[1:5,1:5] #Es mostren les 5 primeres files(pacients) i columnes (Edat, receptor d'estrogen i gens).

```
##                A00AGE    A00ER   A2ML1   AAGAB    AARD
```

```
## TCGA.3C.AAAU.01     55 Positive -0.2282 -0.9332  4.6557
## TCGA.3C.AALI.01     50 Positive -0.2227  3.2441 -0.4702
## TCGA.3C.AALJ.01     62 Positive -0.2307 -0.8205 -0.3211
```

```
## TCGA.3C.AALK.01     52 Positive -0.2277 -0.2190 -0.2737
## TCGA.4H.AAAK.01     50 Positive -0.2244  0.0030 -0.4478
```

5. REGRESSIÓ LOGÍSTICA. MODEL AMB XGBOOST.

Es requereix les següents llibreries:

```
library('xgboost')
```

library(Matrix)

5.1 VARIABLES NUMÈRIQUES

```
prova_100_100<-counts_fin[1:1037,1:1876]
```

5.1.1 Variable resposta (codificada com a 0 o 1, en comptes de Negatiu o Positiu)

```
prova_100_100$A00ER<-as.numeric(factor(prova_100_100$A00ER))-1
```

5.1.2 Variable Edat

```
prova_100_100$A00AGE <-
as.numeric(as.character(prova_100_100$A00AGE))
```

5.2 SEPARAR ALEATÒRIAMENT LES DADES EN ENTRENAMENT (70%) I TEST (30%).

```
traindata<-sample_frac(prova_100_100,size=0.7)
```

testdata<-setdiff(prova_100_100,traindata)

5.2.1 Eliminem aquella variable que volem predir, en aquest cas ER (receptor d'estrogen), de les dades d'entrenament i test.

Dades entrenament:

```
sparse_matrix_train<-sparse.model.matrix(A00ER~.-1,data=traindata)
```

train_label<-traindata[,"A00ER"] #Guardem la columna que conté la variable la qual es vol predir

Dades test:

```
sparse_matrix_test<-sparse.model.matrix(A00ER~.-1,data=testdata)
```

test_label<-testdata[,'A00ER'] #Guardem la columna que conté la variable la qual es vol predir

5.3 CONVERTIR LES DADES A MATRIUS:

```
traindata_matrix<-xgb.DMatrix(data=as.matrix(sparse_matrix_train),
label=train_label)#Dades entrenament
```

testdata_matrix<-xgb.DMatrix(data=as.matrix(sparse_matrix_test), label=test_label)#Dades test

5.4 ENTRENAMENT DEL MODEL:

```
modelentrenament <- xgboost(data=traindata_matrix,
                booster="gbtree",
                nrounds=125,
                max.depth=100,
                eta=0.4,
                nthread=2,
                objective="binary:logistic")
```

```
## [1]  train-logloss:0.399701
```

```
## [2]  train-logloss:0.261285
## [3]  train-logloss:0.180099
## [4]  train-logloss:0.128220
## [5]  train-logloss:0.092981
## [6]  train-logloss:0.069617
## [7]  train-logloss:0.052987
## [8]  train-logloss:0.041331
## [9]  train-logloss:0.033456
## [10] train-logloss:0.027728
## [11] train-logloss:0.023525
## [12] train-logloss:0.020346
## [13] train-logloss:0.017812
## [14] train-logloss:0.015730
## [15] train-logloss:0.014330
## [16] train-logloss:0.013060
```

```
## [17] train-logloss:0.011911
## [18] train-logloss:0.010867
## [19] train-logloss:0.009960
## [20] train-logloss:0.009300
## [21] train-logloss:0.008748
## [22] train-logloss:0.008280
## [23] train-logloss:0.007873
## [24] train-logloss:0.007523
## [25] train-logloss:0.007165
## [26] train-logloss:0.006916
## [27] train-logloss:0.006701
## [28] train-logloss:0.006391
## [29] train-logloss:0.006214
## [30] train-logloss:0.006053
## [31] train-logloss:0.005906
## [32] train-logloss:0.005771
## [33] train-logloss:0.005605
## [34] train-logloss:0.005477
## [35] train-logloss:0.005326
## [36] train-logloss:0.005258
## [37] train-logloss:0.005122
## [38] train-logloss:0.005069
## [39] train-logloss:0.004935
## [40] train-logloss:0.004824
## [41] train-logloss:0.004773
## [42] train-logloss:0.004685
## [43] train-logloss:0.004592
## [44] train-logloss:0.004540
## [45] train-logloss:0.004489
## [46] train-logloss:0.004439
## [47] train-logloss:0.004393
## [48] train-logloss:0.004350
## [49] train-logloss:0.004306
## [50] train-logloss:0.004265
## [51] train-logloss:0.004225
## [52] train-logloss:0.004185
## [53] train-logloss:0.004145
## [54] train-logloss:0.004104
## [55] train-logloss:0.004067
## [56] train-logloss:0.004032
## [57] train-logloss:0.003994
## [58] train-logloss:0.003960
## [59] train-logloss:0.003928
## [60] train-logloss:0.003892
## [61] train-logloss:0.003860
## [62] train-logloss:0.003827
## [63] train-logloss:0.003794
## [64] train-logloss:0.003764
## [65] train-logloss:0.003732
## [66] train-logloss:0.003701
```

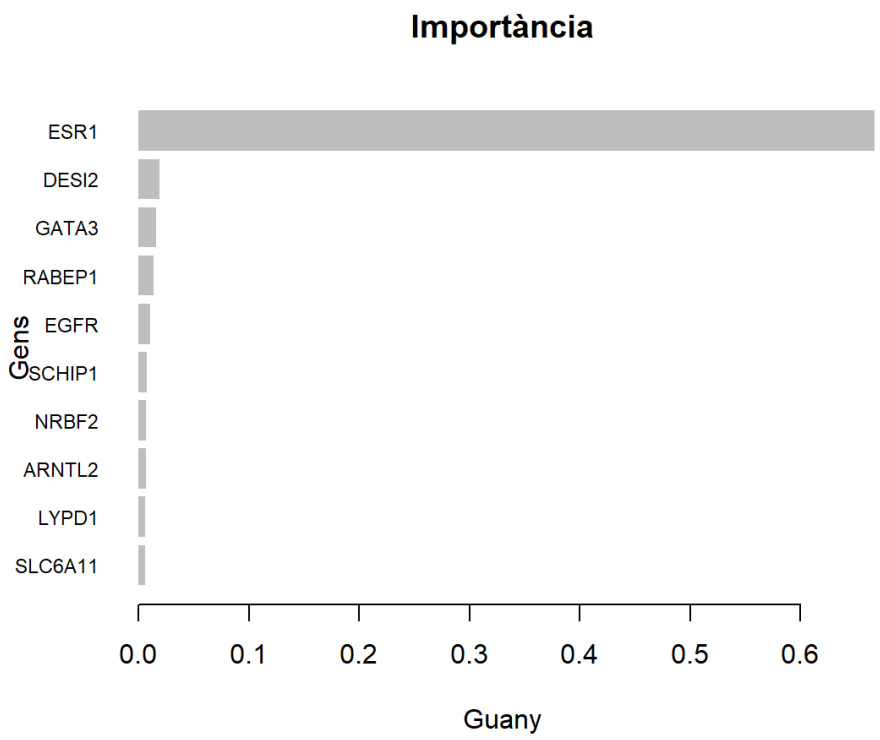
```
## [67] train-logloss:0.003668
## [68] train-logloss:0.003639
## [69] train-logloss:0.003611
## [70] train-logloss:0.003584
## [71] train-logloss:0.003556
## [72] train-logloss:0.003529
## [73] train-logloss:0.003502
## [74] train-logloss:0.003474
## [75] train-logloss:0.003449
## [76] train-logloss:0.003421
## [77] train-logloss:0.003394
## [78] train-logloss:0.003369
## [79] train-logloss:0.003343
## [80] train-logloss:0.003318
## [81] train-logloss:0.003294
## [82] train-logloss:0.003270
## [83] train-logloss:0.003246
## [84] train-logloss:0.003223
## [85] train-logloss:0.003200
## [86] train-logloss:0.003179
## [87] train-logloss:0.003156
## [88] train-logloss:0.003133
## [89] train-logloss:0.003111
## [90] train-logloss:0.003091
## [91] train-logloss:0.003071
## [92] train-logloss:0.003050
## [93] train-logloss:0.003030
## [94] train-logloss:0.003010
## [95] train-logloss:0.002992
## [96] train-logloss:0.002973
## [97] train-logloss:0.002956
## [98] train-logloss:0.002938
## [99] train-logloss:0.002920
## [100]    train-logloss:0.002901
## [101]    train-logloss:0.002885
## [102]    train-logloss:0.002869
## [103]    train-logloss:0.002853
## [104]    train-logloss:0.002839
## [105]    train-logloss:0.002826
## [106]    train-logloss:0.002813
## [107]    train-logloss:0.002813
## [108]    train-logloss:0.002813
## [109]    train-logloss:0.002813
## [110]    train-logloss:0.002813
## [111]    train-logloss:0.002813
## [112]    train-logloss:0.002813
## [113]    train-logloss:0.002813
## [114]    train-logloss:0.002813
## [115]    train-logloss:0.002813
## [116]    train-logloss:0.002813
```

```
## [117]    train-logloss:0.002813
## [118]    train-logloss:0.002813
## [119]    train-logloss:0.002813
## [120]    train-logloss:0.002813
## [121]    train-logloss:0.002813
## [122]    train-logloss:0.002813
## [123]    train-logloss:0.002813
## [124]    train-logloss:0.002813
## [125]    train-logloss:0.002813
```

5.4.1 Variables més importants per la construcció del model

```
importance=xgb.importance(feature_names =
colnames(sparse_matrix_train),model=modelentrenament)
```

```
importance<-importance[1:10,]
xgb.plot.importance(importance_matrix = importance,
                    main='Importància',
                    ylab='Gens',
                    xlab='Guany')
```



5.5 PREDICCIONS

```
nc=length(unique(train_label))
```

```
p=predict(modelentrenament,newdata = testdata_matrix,ntreelimit = 10)
```

```
## [20:02:08] WARNING: amalgamation/../src/c_api/c_api.cc:785:
`ntree_limit` is deprecated, use `iteration_range` instead.
```

```
head(p)
```

```
## [1] 0.9861253 0.9893348 0.9705517 0.9851966 0.1631522 0.9910133
```

Arrodonir a 1 o a 0.

```
pround<-round(p)
```

5.6. AVALUACIÓ DEL MODEL

5.6.1 Creació d'una matriu de confusió per comparar les prediccions amb els valors reals.

```
matriudeconfusio<-table(prediction=pround, actual=testdata$A00ER)
```

```
precision=(sum(diag(matriudeconfusio)))/sum(matriudeconfusio)

matriuconf_data<-data.frame(matriudeconfusio)
matriuconf_data
```

```
##   prediction actual Freq
## 1          0      0   68
## 2          1      0    5
## 3          0      1    9
## 4          1      1  229
```

```
tn<-matriuconf_data[1,3]#Negatius
```

```
fp<-matriuconf_data[2,3]#Falsos positius
fn<-matriuconf_data[3,3]#Falsos negatius
tp<-matriuconf_data[4,3]#Positius

total<-tn+tp+fn+fp#Total de casos predits
positiu<-tp+fn#Total de positius
neg<-tn+fp#Total de negatius
```

Exactitud del model

```
acuraccy<-(tp+tn)/total
```

Marge d’error

```
error_rate<-(fp+fn)/total
```

Sensibilitat del model

```
sensitivity<-tp/positiu
```

Especificitat del model

```
especificitat<-tn/neg
```

F-score

```
f_score<-tp/(tp+((fp+fn)/2))
```

data.frame(acuraccy,error_rate,especificitat,sensitivity,f_score)

```
##    acuraccy error_rate especificitat sensitivity  f_score
## 1 0.9549839 0.04501608     0.9315068   0.9621849 0.970339
```

6. REGRESSIÓ LOGÍSTICA. MODEL CLÀSSIC.

Cal assegurar que les variables siguin numèriques. En aquest cas, ja s'hi han transformat a l'apartat anterior per al model amb XGBoost.

6.1 DADES ENTRENAMENT (70%) I TEST (30%).

S'utilitza la mateixa separació de dades que en el model XGBoost.

```
testrenament_mc<-traindata

ttest_mc<-testdata
```

6.2 MODEL D’ENTRENAMENT.

```
modelo.logit<-glm(A00ER ~ .,data=testrenament_mc,family =
"binomial"(link=logit),maxit=100 )

c<-summary(modelo.logit)
```

```
d<-exp(coefficients(modelo.logit))
```

6.3 PREDICCIONS

```
log.odds<- predict(modelo.logit, newdata = ttest_mc, ntreelimit=10)

p_glm<-exp(log.odds)/(1+exp(log.odds))

head(p_glm)

proud_glm<-round(p_glm)#Arrodonir a 1
```

6.4 AVALUACIÓ DEL MODEL

Matriu de confusió:

```
matriudeconfusio_mc<-
table(prediction=proud_glm,actual=ttest_mc$A00ER)

precisio_mc<-
(sum(diag(matriudeconfusio_mc)))/sum(matriudeconfusio_mc)


matriuconfusion_data<-data.frame(matriudeconfusio_mc)

matriuconfusion_data
```

```
##   prediction actual Freq
## 1          0      0   40
## 2          1      0   14
## 3          0      1  130
## 4          1      1   87
```

```
tng<-matriuconfusion_data[1,3]#Negatius

fpg<-matriuconfusion_data[2,3]#Falsos positius

fng<-matriuconfusion_data[3,3]#Falsos negatius

tpg<-matriuconfusion_data[4,3]#Positius


totalg<-tng+fpg+fng+tpg #Total de casos predits

posg<-tpg+fng#Total de positius

negglm<-tng+fpg #Total de negatius
```

Exactitud del model

```
acuracyglm<-(tpg+tng)/totalg
```

Marge d'error

```
error_rateg<-(fpg+fng)/totalg
```

Sensibilitat del model

```
sensitivitatg<-tpg/posg
```

Especificitat del model

```
especificityg<-tng/negglm
```

F-score

```
f_scoreg<-tpg/(tpg+((fpg+fng)/2))
```

Conjunt dels diferents parametres per l'avaluació del model:

```
data.frame(acuracyglm,error_rateg,sensitivitatg,especificityg, f_scoreg)
```

```
##    acuraccy error_rate especificitat sensitivity  f_score
## 1 0.4686347 0.5313653    0.4009217  0.7407407   0.5471698
```