UNIVERSITAT DE BARCELONA

Treball final de grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Aprenentatge automàtic per predir risc cardiovascular amb dades clíniques

---

Autor: Iker Honorato López

Directora: Dra. Laura Igual
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 20 de juny de 2021

Resum

L'ateroesclerosi és un dels principals precursors de les patologies cardiovasculars, la primera causa de defunció al primer món. Un dels mètodes més utilitzats per diagnosticar-ho són les ecografies d'artèria caròtida, ja que són poc intrusives i econòmiques. No obstant això, produeixen imatges de baixa qualitat que fan la diagnosi de l'ateroesclerosi complicada.

Tanmateix, existeixen altres mesures de risc cardiovascular. Les taules de risc, que a partir de diversos factors d'estil de vida i mèdics, assignen la probabilitat d'un individu de patir un episodi cardiovascular. Aquest tipus de taules hereten la seva funcionalitat de l'estudi de Framingham, que va analitzar dades de població estatunidenca per crear la seva pròpia funció de risc, sent així el primer estudi a fer-ho.

No obstant això, adaptar les taules a tota la població no és precís, ja que existeixen diferents variables epidemiològiques que poden fer variar els valors de les taules, i fer estudis per ajustar-les és complicat. A més a més, tenen altres limitacions: edat màxima per aplicar-la, necessitat de no tenir cap valor desconegut, i també, que gran part dels futurs episodis cardiovasculars són classificats com a grups de risc mitjà i per tant no són medicats.

Aquest projecte té com a objectiu millorar l'actual funció de REGICOR, calculada en població catalana, mitjançant algoritmes de machine learning i una combinació de dades mèdiques i d'imatges d'ecografia de voluntaris.

Abstract

Atherosclerosis is one of the main precursors to cardiovascular pathologies, the first defunction cause on developed countries. One of its principal diagnosis methodologies is carotid ultrasound images due to their low cost and intrusivity. Nonetheless, these produce low quality representations, which makes the diagnosis of atherosclerotic plaques a laborious task.

In spite of that, other risk measurement methodologies exist. Risk tables which, taking into consideration diverse lifestyle and medical data, assign the probability of an individual to suffer a cardiovascular event. These types of tables inherit their functionality from the Framingham study, which analyzed data of United States population to create its risk function, thus being the first study to do so.

However, adapting these tables to all population is not precise, as there are different epidemiological factors that can affect the values of the tables, and conducting studies to adjust them is expensive. Moreover, other limitations exist, as it has been proved that most of the future cardiovascular events end up classified on mid-range risk groups, thus not being medicated, besides an age limit to apply the tables, and not accepting missing values.

This project sets out to improve the current REGICOR risk function, computed in catalan population, using machine learning prediction models and a combination of medical and ultrasound data of volunteers.

Índex

1 Introducció **1**
- 1.1 Contextualització . . . 1
- 1.2 Motivació i Objectius . . . 5
- 1.3 Estructura del document . . . 6

2 Treballs previs **7**
- 2.1 Estat de l'art . . . 7
- 2.2 L'estudi de Framingham . . . 7
- 2.3 Funció de risc de REGICOR . . . 8
- 2.4 Limitació de REGICOR . . . 9

3 Estudi de les dades **11**
- 3.1 Base de dades de REGICOR . . . 11
- 3.2 Exploració de les dades . . . 12
  - 3.2.1 Estudi general de les variables . . . 12
  - 3.2.2 Variable Sexe . . . 16
  - 3.2.3 Variable Diabetis . . . 17
  - 3.2.4 Variable Tabaquisme . . . 19
  - 3.2.5 Variable Hipertensió . . . 20
  - 3.2.6 Variable Estratificació . . . 21
  - 3.2.7 Estudi de defuncions . . . 22
  - 3.2.8 Estudi dels valors desconeguts . . . 23

4 Models de predicció **25**
- 4.1 K-Nearest Neighbors . . . 25
- 4.2 Models basats en arbres . . . 26
  - 4.2.1 Bagging . . . 27
  - 4.2.2 Random Forest . . . 28
  - 4.2.3 Boosting . . . 28
  - 4.2.4 Hybrid Random Forest Linear Model . . . 29
- 4.3 Support Vector Machines . . . 29
- 4.4 Regressió logística . . . 32
- 4.5 K-Means . . . 33
- 4.6 Deep Learning . . . 33
  - 4.6.1 Optimitzadors . . . 34
  - 4.6.2 Mètodes de Regularització . . . 35
  - 4.6.3 Estructura de la Xarxa Neuronal . . . 36

5 Tècniques de validació i processament de dades **37**
- 5.1 Metodologies de validació . . . 37
  - 5.1.1 Grid-search Cross-Validation . . . 37
  - 5.1.2 Mètriques . . . 38
- 5.2 Preprocessament de dades . . . 39
  - 5.2.1 Normalització de les dades . . . 39
  - 5.2.2 Codificació categòrica . . . 40

5.2.3 Outliers . . . 40
5.2.4 Dades Desconegudes . . . 41
5.3 Dades desequilibrades . . . 41
5.3.1 Recàlcul del pes de les classes . . . 42
5.3.2 Resampling . . . 42
5.4 Detalls d'implementació . . . 45

6 Experiments i resultats **47**
6.1 Plantejament . . . 47
6.1.1 Preprocessament de dades . . . 47
6.1.2 Estratègia de Cross-Validation . . . 48
6.2 Cerca no Supervisada . . . 48
6.3 Predicció d'episodi cardiovascular . . . 51
6.3.1 Dades dintre de les restriccions de la funció REGICOR . . . 51
6.3.2 Dades fora de les restriccions de la funció de REGICOR . . . 55
6.4 Interpretació dels resultats . . . 56

7 Conclusions i treball futur **58**
7.1 Conclusions . . . 58
7.2 Treball futur . . . 58

1 Introducció

1.1 Contextualització

Les malalties cardiovasculars (CVDs) són un grup de malalties que engloben afectacions al cor o al teixit sanguini i són la principal causa de defunció en el primer món, la majoria d'aquestes acaben en infarts o embòlies i una de les principals causes perquè es produeixen és l'ateroesclerosi.

L'ateroesclerosi és una malaltia inflamatòria dels vasos sanguinis que resulta de la interacció entre el flux sanguini, la sang i la paret arterial. Els processos involucrats en la formació de placa d'ateroma són:

- Inflamació
- Proliferació muscular
- Degeneració cel · lular
- Necrosi
- Calcificació

Apareix juntament amb altres afectacions de la paret de l'artèria que precedeixen complicacions cardiovasculars. Aquest procés pot començar en qualsevol punt de la vida i empitjora amb l'edat, de manera que la diagnosi prematura d'aquestes afectacions permet prendre mesures cautelars per evitar episodis crítics per la salut dels pacients.

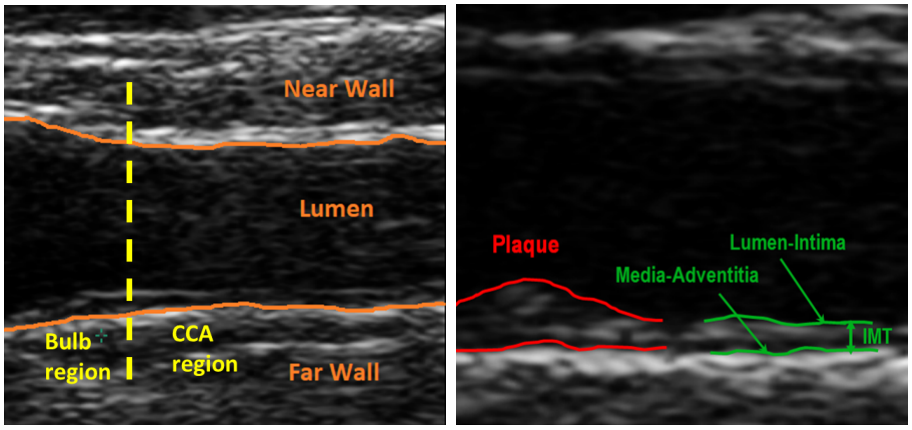
Hi ha diferents tipus de placa [31]:

- **Placa Vulnerable**: És aquella que és propensa a trencar-se, té una gran quantitat de component lípidic, i la capa de separació entre l'artèria i el contingut de la placa és molt fina.
- **Placa Estable** La placa creix i obstrueix de manera progressiva, les parets es calcifiquen i creen una barrera "dura"que evita que els components de la placa es trenquin.

En cas que la placa es trenqui, tot el seu contingut entra en contacte amb la circulació sanguínia i pot haver-hi trombosi que esdevindrà un infart agut de miocardi (IAM). El tractament de l'IAM es realitza mitjançant cateterisme, intentant destapar l'artèria aspirant els trombus resultants del trencament de la placa.

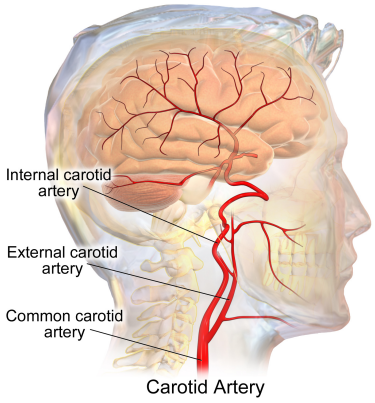
Actualment la Intima Media Thickness (IMT), **figura 1 dreta**, es considera un indicador de placa. La IMT és definida com la distància que hi ha entre la lumen-intima (LI) i Media-Adventitia (MA) i és normalment mesurada a la cara externa de la artèria caròtida (AC). Un dels mètodes més utilitzats per mesurar-la per la seva poca intrusivitat i preu econòmic és l'ecografia de l'AC, que permet tenir una imatge en blanc i negre de la zona. Però tot i poder visualitzar l'afectació, la qualitat de la imatge és baixa i conté molt soroll, com es pot veure a les imatges de la **figura 1**, cosa que dificulta el diagnòstic als experts.

Figura 1: Ecografia de la caròtida i IMT a la dreta, extretes de [10]



Cal tenir en compte també, que l'AC té diferents zones, **figura 2**, la part comuna de l'artèria (CCA), la cara interna (ICA) o el segment de bulb de l'AC i que la visualització és més fàcil a la part comuna, que no pas a les altres dues. No obstant això, es coneix que les plaques són més propenses a aparèixer a la bifurcació de l'artèria i a la part interna.

Figura 2: Segments de l'artèria caròtida, extreta de [10]



És important trobar una manera de millorar l'ecografia o intentar donar suport mitjançant tècniques d'aprenentatge automàtic per tal d'extreure informació d'aquestes i proporcionar eines de diagnòstic als experts clínics. Avui dia existeixen moltes metodologies per tractar aquestes dades, des de segmentació fins a caracterització o mesura del IMT de manera automàtica.

No obstant això, tot i que l'IMT es tracta d'un indicador clar de les CVDs, cal tenir en compte que hi ha molts altres factors que afecten en l'aparició d'aquestes condicions mèdiques [13]. Aquests factors es poden dividir en dos grups, els factors de risc **taula 1** i els factors protectors **taula 2**.

Taula 1: Factors de risc cardiovascular

	Modificables	**No modificables**
Ambientals	Tabac Infeccions Socioeconòmics Dietes greixoses	Edat Sexe masculí Genètics
Metabòlics	Displèmies Diabetis Hipertensió Homocisteïna	

Taula 2: Factors protectors de risc cardiovascular

	Modificables	**No modificables**
Ambientals	Dieta Exercici Físic Alcohol	Sexe Femení Genètics
Metabòlics	Colesterol HDL	

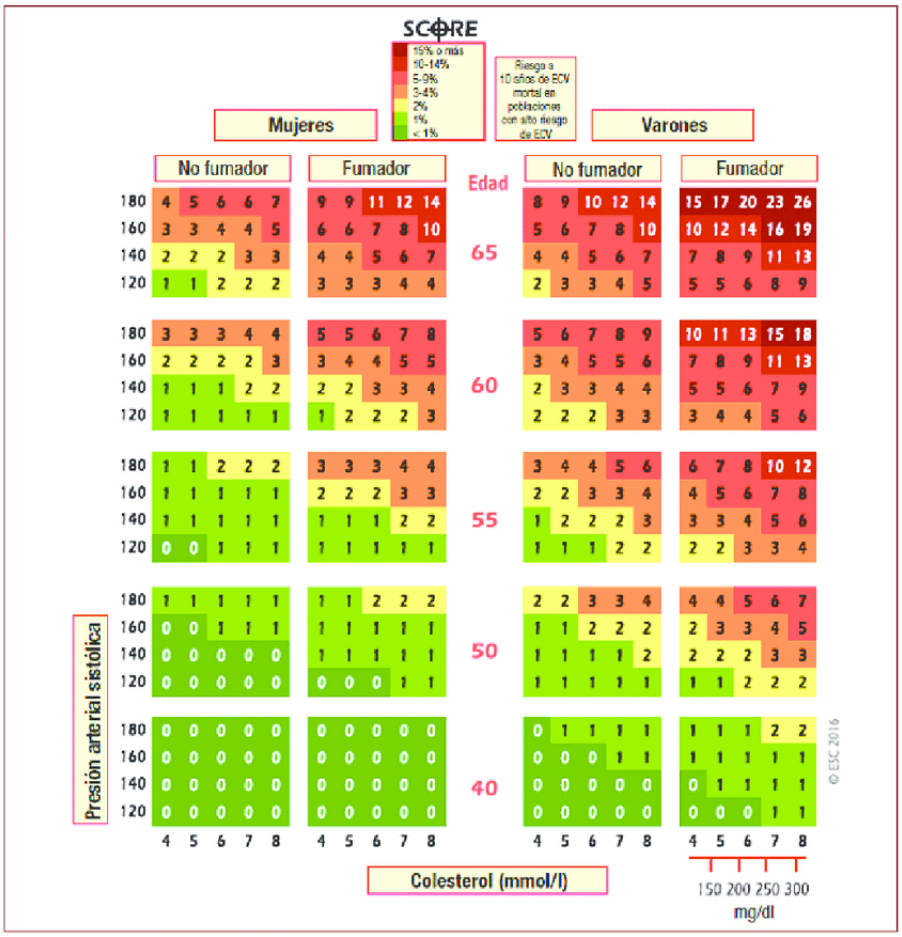
La displèmia és un conjunt d'alteracions del perfil lipídic que estan associades a un major risc de patir malaltia coronària:

- LDL $>$ 140 mg/dL
- HDL $<$ 40 mg/dL
- Hipertrigliceridèmia
- Lipoprotïna B elevada
- lipoproteïna A disminuïda

A mesura que s'acumulen aquests factors de risc, la mortalitat augmenta i per tant també la probabilitat de patir ateroesclerosis [38]. En l'àmbit mèdic, per fer el càlcul de risc, existeixen unes taules anomenades SCORE (Systematic Coronary Risk Estimation). Un exemple s'il · lustra a la **figura 3**, on es calcula el risc de patir un esdeveniment letal al cap de 10 anys. Aquestes taules es basen en funcions o

regressions de COX, com s'explicarà a la secció 2.3. En general les taules SCORE tenen en compte el sexe, els valors de la pressió, tabaquisme i els nivells de colesterol.

Figura 3: Taula SCORE, extreta i traduïda de [30]



A partir d'aquestes taules es considera que un pacient és de risc molt alt si compleix qualsevol dels següents punts:

- Han patit o pateixen alguna malaltia cardiovascular.
- Tenen diabetis que hagi produït danys als òrgans.
- Tenen un valor SCORE $> 10\%$.

Es considera risc alt si presenten tots els següents:

- Pacients amb 1 factor de risc cardiovascular molt elevat.
- Diabètics.
- Tenen un valor SCORE entre 5-10%.

Risc moderat si tenen valor SCORE entre 1-5% i baix si SCORE $< 1\%$.

1.2 Motivació i Objectius

L'objectiu d'aquest treball és trobar una millora a la metodologia de les taules SCORE, per tal que es pugui adaptar millor la predicció de risc cardiovascular mitjançant eines de machine learning o deep learning segons calgui.

Aquest TFG forma part d'un projecte de recerca que busca fusionar dades mèdiques de pacients amb imatges de la seva AC per intentar millorar la predicció de risc cardiovascular. Per aconseguir això, s'estudiarà de manera exhaustiva la base de dades i s'avaluarà el rendiment de diversos models de predicció clàssics i trobats a la literatura per veure quin s'ajusta millor al problema.

Així doncs, aquest treball ha estat dividit en tres grans blocs, el primer d'aquests, una revisió dels treballs previs, per tal de familiaritzar-se amb els mètodes que s'utilitzen per diagnosticar els episodis cardiovasculars i quins són els biomarcadors que es tenen més en compte. Així es podrà aconseguir:

- Ser capaç d'entendre les complicacions d'un cas clínic.
- Introduir-se en un tema mèdic des de 0.
- Aprendre a utilitzar les eines de cerca científica.

Seguidament, es farà un estudi de la base de dades de REGICOR, començant per les variables que utilitza la seva funció per avaluar el risc, a més a més d'altres que es mencionin a la literatura. Serà important aconseguir visualitzacions de la base de dades que permetin la fàcil comprensió d'aquesta i la relació que hi pugui haver entre biomarcadors i episodis. Així doncs els objectius d'aquesta part seran:

- Explorar la base de dades REGICOR.
- Extreure relacions entre variables.
- Fer visualitzacions útils per entendre la base de dades.
- Trobar possibles problemes en aquesta.
- Agregar noves variables a la funció.

Per últim, es desenvoluparan un conjunt de proves amb les quals poder avaluar l'eficàcia de diferents models de machine learning i escollir el millor d'aquests per realitzar la nova predicció. A més a més, també serà necessari trobar mètriques de validació pel problema plantejat. Es formulen els següents objectius:

- Aprofundir coneixement en els diferents algoritmes d'aprenentatge automàtic.
- Definir una prova estàndard per tots els models.
- Escollir quina és la mètrica més adient per mesurar el rendiment de cada mètode.
- Trobar un mètode de validació correcte.
- Aconseguir un mètode que funcioni igual o millor que la funció de REGICOR.

1.3 Estructura del document

La memòria ha estat dividida en 7 capítols:

- **Capítol 1, Introducció**, en el qual es presenta la contextualització del problema i s'exposa la motivació del treball i els diferents objectius que es plantegen.
- **Capítol 2, Treballs previs**, en el qual s'analitzen diferents articles de la literatura i es presenta l'estudi de Framingham, precursor de la funció de risc de REGICOR, que és la que pren el treball com a objectiu a millorar.
- **Capítol 3, Estudi de les dades**, on es fa un estudi de les diferents variables que s'utilitzaran per fer la predicció d'episodi cardiovascular i es plantegen les relacions que existeixen entre variables.
- **Capítol 4, Models de predicció**, en el qual es realitza un estudi bibliogràfic dels diferents models que s'han utilitzat per les prediccions, amb l'objectiu d'entendre el seu funcionament i aplicar-los de manera apropiada.
- **Capítol 5, Tècniques de validació i processament de dades**, on s'exposen les diferents tècniques que s'han utilitzat per tractar les dades i el desequilibri d'aquestes.
- **Capítol 6, Experiments i resultats**, en el qual s'expliquen de manera detallada els experiments que s'han realitzat i s'avaluen els resultats extrets d'aquests.
- **Capítol 7, Conclusions i treball futur**, on s'extreuen les conclusions del treball i es proposen millores de les metodologies utilitzades que podrien afavorir els resultats.

2 Treballs previs

2.1 Estat de l'art

Existeixen diferents treballs a la literatura que utilitzen dades clíniques per a avaluar risc cardiovascular o d'altres patologies.

A [33], s'utilitza una base de dades amb 1069 participants dels quals van utilitzar, edat, sexe, pressió diastòlica i sistòlica, colesterol total, LDL, HDL, diabetis i tabaquisme, a més a més de diferents indicadors de medicació i variables sobre la calcificació de l'artèria. També, mitjançant mètodes de feature selection basats en models d'arbres van afegir les 15 variables amb millor puntuació en aquest test. Construeixen els models a partir de XGBoost posat que funciona correctament en estratificacions de risc cardíac i utilitzen també estratègies de validació com el cross-validation. Conclouen que les metodologies de ML tenen millor eficàcia a llarg termini que les metodologies clàssiques de detecció de risc i que les dades provinents del sèrum han estat beneficioses.

A [27], també s'utilitza un algoritme de boosting, el LogitBoost, i variables de calcificació provinents d'una base de dades que compta amb 66.636 participants asimptomàtics dels quals el 67% són homes. Les seves mitges d'edat són de 54 $\pm$ 11 anys i l'edat mínima eren els 18. Es troben també resultats positius a l'hora de predir episodis en comparació amb els mètodes clàssics.

D'altra banda a [19], es pot trobar un recull bibliogràfic dels diferents algoritmes que s'han utilitzat a diversos estudis, i els seus resultats en l'aplicar els models en una base de dades amb 12 variables, essent aquestes edat, sexe, mal al pit, pressió sanguínia, colesterol, diabetis i altres resultats de proves d'esforç. Utilitza diferents mètriques per a comparar els mètodes (accuracy, precisió, sensitivitat, especificitat i f1_score). Els mètodes que compara són naïve bayes, models lineals, regressió, decision trees, random forests, Support Vector Machines (SVM) i HRFLM (Hybrid Random Forest Linear Models). Els resultats demostren que HRFLM, té la millor eficàcia de tots els models.

També, a l'estudi [18], centrat en predir casos d'episodis cardiovasculars en pacients que pateixen de Cardiomiopatia Hipertròfica mitjançant algoritmes de ML, han utilitzat random forest per cobrir els valors null i han codificat les variables categòriques com dummy variables. Han creat els models amb 20 variables provinents de criteri mèdic i de mètodes de feature selection basats en arbres. Novament troben que de tots els models que han provat, tots tenen una major precisió que els mètodes clàssics.

2.2 L'estudi de Framingham

L'estudi de Framingham (FHS)[4], és el precursor de la majoria de funcions de risc cardiovascular actuals. Ja en 1900, les malalties cardiovasculars suposaven una de les principals causes de mortalitat, però no es tenia coneixement de per què aquestes patologies es manifestaven.

Això va motivar al govern dels Estats Units a incentivar diversos estudis per investigar les CVD i millorar la seva prevenció. Donada la seva situació i població variada, la població de Framingham va ser escollida per dur a terme aquests estudis epidemiològics, que van començar el 1948 i que encara tenen continuïtat.

D'aquest estudi van néixer molts altres que van permetre definir el que són ara els estàndards de factors de risc de les malalties cardiovascular, i també la funció de risc de Framingham, que s'utilitza arreu del món. No obstant això, aquesta funció ha estat aproximada segons l'epidemiologia americana, i a causa de les variables d'entorn que hi pot haver entre continents, pot no ajustar-se correctament a totes les poblacions [20].

2.3 Funció de risc de REGICOR

Les taules SCORE són l'estàndard europeu, i estan calculades amb valors de diferents poblacions europees, per ajustar-se millor a aquestes. Des del Registre Gironí del Cor (REGICOR), es va fer un estudi [23] per tal de trobar els valors de les taules de la zona de Catalunya.

La funció de REGICOR pren 8 factors de risc, basant-se en l'estudi de Framingham, per fer la seva valoració sobre el risc. Les variables utilitzades són:

1. **Sexe**: Divisió entre homes i dones.
2. **Edat**: Edat del participant en el moment del seguiment, el seu valor mínim és 35 i el màxim 74.
3. **Colesterol total**: Valor del colesterol del participant en el moment de l'estudi.
4. **Colesterol HDL**: Valor del colesterol HDL en el moment de l'estudi.
5. **Pressió sistòlica**: Valor de la pressió sistòlica en el moment de l'estudi.
6. **Pressió diastòlica**: Valor de la pressió diastòlica en el moment de l'estudi.
7. **Diabetis**: Si el participant és diabètic o no.
8. **Fumador**: Diferencia si el participant ha fumat alguna vegada, és fumador, és exfumador o si no hi ha dades per determinar-ho.

Tant la funció de Framigham com la de REGICOR, són regressions de COX, que és un mètode d'anàlisis de supervivència, orientat a predir un episodi cardiovascular en un període de 10 anys. La funció està definida com:

$$prob(episodi|x_i) = 1 - S^{exp(\beta * x_i) - (\beta * \overline{x})}$$

on S és la probabilitat d'episodi, $\beta * x_i$ és el sumatori dels productes del coeficients de COX(β) de cada factor de risc pel valor del factor de risc de l'individu x_i de manera que:

$$\beta * x_i = \sum_{n=1}^{n=N} \beta_n * x_{in}$$

on N és el número de factors de risc, $\beta * x_{in}$ és el sumatori dels productes dels coeficients de COX (β) per valor mig del factor de risc a la població de manera que:

$$\beta * \overline{x} = \sum_{n=1}^{n=N} \beta_n * \overline{x}$$

La diferència entre la funció de REGICOR i Framingham són els valors de les S i les mitges de les x. Mentre a la primera aquestes variables corresponen als valors espanyols, a la segona corresponen a població nord-americana. D'altra banda, els valors de les β i les variables de factor de risc són iguals a les dues funcions. Per REGICOR, aquestes serien les funcions per homes i dones:

$$prob(episodi|home_i) = 1 - 0.951^{exp(\beta * x_i) - (3.489)}$$
$$prob(episodi|dona_i) = 1 - 0.978^{exp(\beta * x_i) - (10.279)}$$

A partir del valor que doni la fórmula, s'estratifiquen els grups de risc segons la taula 3 i als que tenen el risc més alt se'ls medica per intentar reduir la possibilitat d'episodi cardiovascular.

Taula 3: Estratificació dels grups de risc segons el valor resultant de REGICOR

Valor de la funció REGICOR	**Grup de risc**
> 0.05	Baix
≥ 0.5 o < 0.10	Mitjà baix
≥ 0.10 o < 0.15	Mitjà alt
≥ 0.15	Alt

2.4 Limitació de REGICOR

En comparació amb el FHS, que sobreestima el risc cardiovascular, la funció de REGICOR està validada per població espanyola. Així doncs és més precisa a l'hora de predir episodis cardiovasculars al territori. No obstant això, aquest tipus de taules tenen una llarga llista de limitacions.

La primera, el límit d'edat, ja que passats els 74 anys els valors de les taules varien de manera substancial. Per tant, tot i tenir dades sobre malalties com les cerebrovasculars, no es pot aplicar la funció, ja que la població que tendeix a patir-les és d'una edat avançada, i queden fora rang.

També, encara que l'estudi s'hagi fet dintre el mateix país, hi ha comunitats autònomes que tenen altres realitats epidemiològiques i per les quals la funció de REGICOR necessitaria petites variacions.

A més a més, si en algun participant alguna de les proves necessàries per extreure les dades de les variables no s'ha pogut fer, o ha donat un valor erroni, la funció no es podrà aplicar.

Per últim, l'ideal en aquestes funcions és que el màxim d'episodis es concentrin als grups de risc més alts, però més del 80% dels casos es troben concentrats als grups de risc moderat. Això pot ser degut a molts factors, però podria millorar afegint més bioindicadors al càlcul de risc.

3 Estudi de les dades

3.1 Base de dades de REGICOR

La base de dades utilitzada a aquest treball ha estat proporcionada pel Registre Gironí del Cor (REGICOR), i conté aproximadament 286 variables respecte a diferents biomarcadors d'interès clínic de 5074 subjectes juntament amb aproximadament 4751 imatges de la CCA.

Els participants van ser escollits seguint un mostreig aleatori betàpic, on primer es van seleccionar poblacions de Girona i després, d'aquestes, es van seleccionar mateix nombre d'homes i dones dividits en grups de 10 anys. Es va recollir informació sociodemogràfica, sobre factors de risc cardiovascular, activitat física, dieta o medicació. A la **figura 4** es detalla la cronologia dels seguiments, en els quals es va fer una enquesta telefònica preguntant per diversos aspectes relacionats amb les CVD i si havien patit algun episodi des del primer estudi, en cas que no es respongués, es consultava amb el registre de defunció dels hospitals.

Figura 4: Cronologia de l'estudi de REGICOR

En el primer seguiment, es va fer un pas més per ampliar la base de dades, afegint imatges de l'AC. Com aquest treball busca com a objectiu final una funció capaç d'extreure informació d'imatge i dades de pacients, el basal serà el seguiment 1, posat que és quan s'afegeixen aquestes. Així doncs, per tal d'entrenar els models de predicció es disposaran de 4453 individus, que són els que van respondre al seguiment 2.

Com els algoritmes que s'han estudiat no admeten imatges com a entrada, la base de dades conté 24 característiques de les imatges, anomenades fenotips, que s'afegiran com a variables numèriques. Per cada individu hi ha 4 imatges, que corresponen a:

1. CCA esquerra
2. CCA dreta

3. Bulb esquerre
4. Bulb dret

Aquestes s'han segmentat i s'han extret els següents fenotips:

1. Mitjana de l'IMT.
2. Valor màxim de l'IMT.
3. Valor mínim de l'IMT.
4. Variabilitat de l'IMT.
5. Àrea total de placa en mil · límitres quadrats.
6. Mediana de l'escala de grisos a la regió de la placa.

3.2 Exploració de les dades

Per fer l'estudi de la base de dades, s'han considerat les 8 variables que utilitza la funció de risc de REGICOR per estimar el risc, a més a més de 3 biomarcadors més que s'extreuen de [23, 13], que són hipertensió, índex de massa corporal (IMC) i triglicèrids.

Primer s'ha fet un estudi general de les mitges dels diferents grups de risc, per sexe o per voluntaris que hagin presentat episodi cardiovascular (taula 4, figura 5, figura 7, figura 7).

A continuació es realitza un estudi de les variables categòriques respecte la possibilitat de tenir episodi, i a més a més, la seva relació amb les altres variables numèriques mitjançant visualitzacions generals (secció 3.2.2, secció 3.2.3, secció 3.2.3, secció 3.2.4, secció 3.2.5, secció 3.2.6).

Després s'avaluen les causes de mort dintre de la base de dades (secció 3.2.7) i també s'exploren les dades que van ser deixades fora del càlcul de risc per no complir amb les restriccions de REGICOR, ja que aquestes dades també s'utilitzaran per entrenar els diferents models de ML seleccionats (secció 3.2.8).

3.2.1 Estudi general de les variables

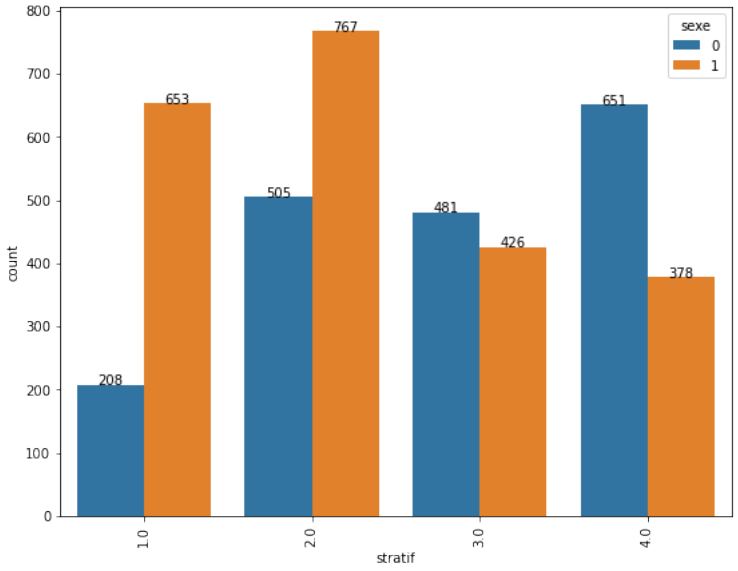
A la **taula 4** es presenten les mitges i desviació mitjana de la població general, homes i dones, de les variables de la funció de risc de REGICOR i les 3 noves afegides. Aquí es pot observar com els homes tenen una tendència a l'alta en colesterol, pressió sanguínia o en triglicèrids. També cal remarcar que el grup mitjà de risc és el 2.5, és a dir, a mig camí entre el risc mitjà baix i el risc mitjà alt.

Taula 4: Mitges de les variables per sexe

Nom variable	General	Homes	Dones
Edat	56.63 ± 9.43	56.58 ± 9.38	56.68 ± 9.47
Colesterol total	206.18 ± 35.74	203.34 ± 35.22	208.54 ± 36.01
Colesterol HDL	52.86 ± 11.84	48.10 ± 10.16	56.83 ± 11.68
Pressió Sistòlica	127.09 ± 19.00	132.00 ± 17.65	123.01 ± 19.12
Pressió Diastòlica	77.68 ± 10.03	80.46 ± 9.58	75.36 ± 9.80
Diabetis	1.88 ± 0.32	1.83 ± 0.36	1.91 ± 0.27
Tabaquisme	0.79 ± 1.02	1.12 ± 1.04	0.56 ± 0.94
IMC	27.08 ± 4.60	27.48 ± 3.80	26.74 ± 5.13
Hipertensió	1.56 ± 0.50	1.49 ± 0.50	1.63 ± 0.48
Triglicèrids	97.95 ± 56.76	107.68 ± 56.76	89.92 ± 48.17
Episodi	0.011 ± 0.10	0.018 ± 0.13	0.006 ± 0.07
Grup de risc	2.52 ± 1.08	2.85 ± 1.02	2.23 ± 1.05

A partir d'aquí es pot mirar la divisió entre homes i dones segons grup de risc a la **figura 5**. Les dones es troben en gran majoria als grups d'estratificació més baixos, en el cas dels homes passa el contrari.

Figura 5: Distribució dels participants segons l'estratificació i sexe



La **figura 6**, una gràfica de coordenades paral · leles, permet visualitzar la relació entre mostres i variables. Cada columna representa una de les variables, el valor màxim a dalt i el mínim a baix. Les línies són les mitges de cada grup de risc per cada variable. Prenent de referència l'última columna, que són els valors d'estratificació, es pot apreciar el següent:

1. **Sexe**: Es pot observar com la línia del grup de risc més baix, és el que té la mitja més alta a la variable sexe, això indica que la quantitat de dones és més elevada que la d'homes, i tal com es veu a les línies següents, a mesura que puja el risc aquesta mitja baixa. De manera que es pot assegurar que ser dona i el risc són inversament proporcionals.

2. **Edat**: Edat i risc són completament proporcionals, com més alta l'estratificació més alta la mitja d'edat.

3. **Colesterol**: A risc alt, el colesterol total és el més elevat, mentre que el HDL és el més baix, de manera que la funció de REGICOR puntua de manera negativa el colesterol LDL, el mateix passa als altres grups.

4. **Pressió**: La pressió torna a ser un clar exemple de proporcionalitat respecte a l'estratificació, a nivell més alt, més risc

5. **Diabetis**: La diabetis és una dada que té un comportament no normatiu, els dos grups on té la mitja més alta són els extrems de risc, mentre que on és més baixa és als valors del centre. En aquest cas, la base de dades defineix un voluntari amb diabetis com un 1 i un 2 a qui no en tingui. No obstant això, cal veure que la mitjana màxima és 1.91 mentre que la mínima és 1.85, cosa que fa pensar que els participants amb diabetis estan repartits en el conjunt de l'estratificació.

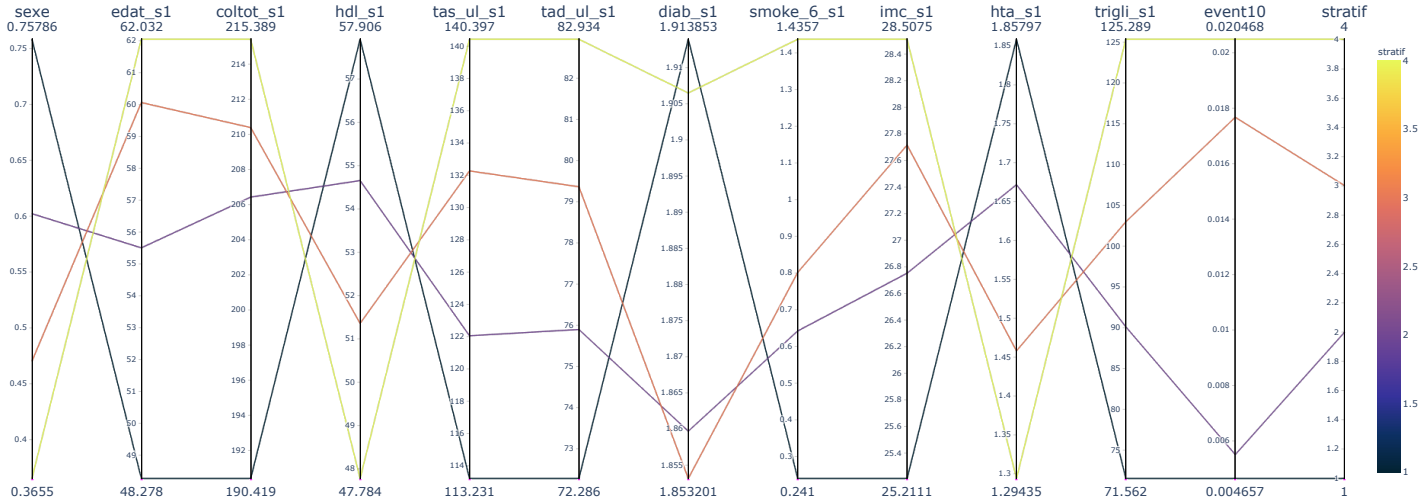
6. **Fumador**: En aquest cas, es valora com a 0 no haver fumat mai, 1 haver fumat, 2 ser exfumador i 9 el no tenir dades suficients. Per tant, sense saber com estan repartits els valors no es pot treure una conclusió ferma, a priori sembla que la mitja més alta, és a dir, haver fumat o ser fumador, corresponen als riscs més alts mentre que la mitja més baixa atribuïda a no fumar es troba als riscs més baixos.

7. **IMC** El IMC també es comporta de manera proporcional al risc de patir episodi.

8. **Hipertensió** Tot i que als grups d'estratificació del mig no segueix una distribució proporcional al risc, en els extrems sí que passa, la majoria dels individus amb hipertensió es troben al grup de risc més alt.

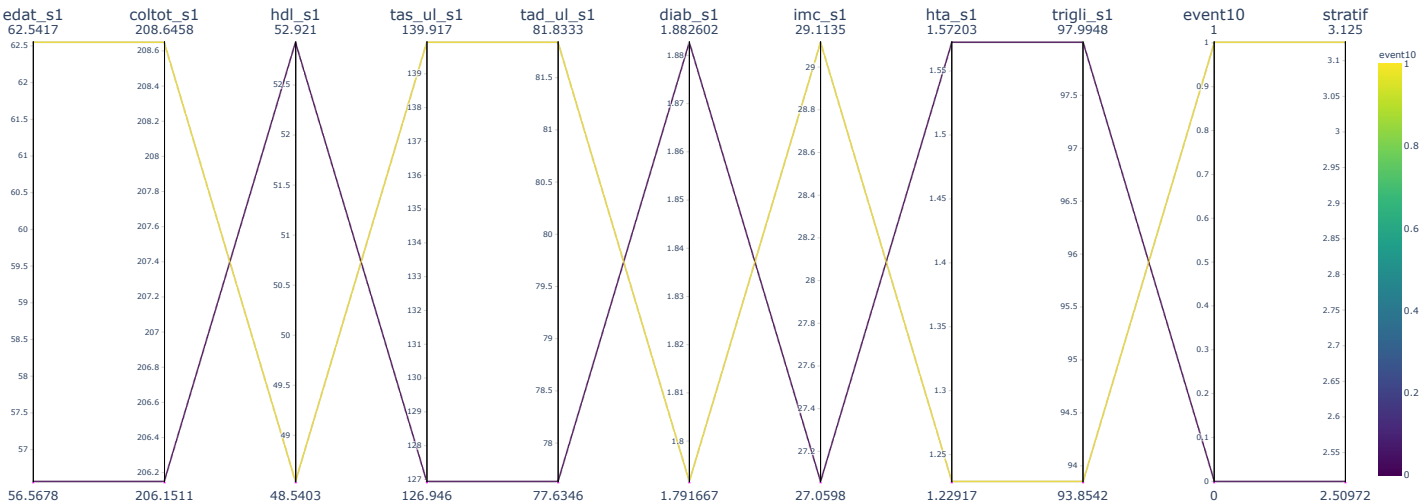
9. **Triglicèrids** Un altre cas de proporcionalitat respecte del risc cardiovascular.

Figura 6: Mitges dels diferents grups de risc



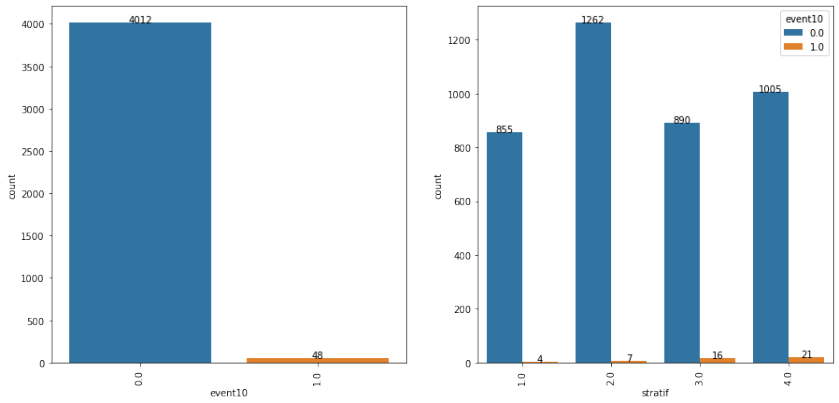
Aquestes gràfiques de coordenades paral · leles es basen en mitges, i els límits de les columnes prenen els valors màxims i mínims d'aquests. En el cas **figura 7**, on s'aprecien les mitges de mostres amb episodi i sense, trobem que la comparació és binària, per tant els extrems sempre seran d'una de les dues classes, atès que només hi ha 2 valors a comparar. Per tant, no es poden extreure conclusions a escala numèrica, ja que pot haver-hi poca diferencia entre el valor mínim i màxim de les columnes. No obstant això, sí que es pot observar que hi ha una clara diferencia entre biomarcadors, els episodis tenen un valor elevat en tots aquells que són considerats de risc i un valor baix en tots aquells que són beneficiosos.

Figura 7: Mitges segons event



A la **figura ??** es pot observar com els episodis són clarament una classe minoritària i que per tant la base de dades és desequilibrada. A més a més, com només hi ha 48 casos d'episodi i 4060 individus, a la **figura 7** la mitja dels que no en tenen es pot entendre com la mitja del total.

Figura 8: Distribució del total d'episodis



3.2.2 Variable Sexe

Es comença doncs amb l'estudi de les variables categòriques. És interessant seguir l'exemple de les funcions de Framingham i REGICOR, que tenen valors diferents per homes i per dones, i intentar buscar altres relacions entre les dades. A la **taula 5** es pot observar com es distribueixen els episodis cardiovasculars segons aquesta divisió en els diferents grups de risc. Com es pot apreciar, a mesura que s'avança en risc, el nombre de dones disminueix i el nombre d'homes augmenta. En canvi, els episodis relacionats amb dones es troben més repartits entre les diferents classes en comparació amb els dels homes, on s'aprecia un augment dels episodis acompanyat del risc. Per tant, es pot concloure que hi ha un biaix a la funció, és a dir, s'ajusta millor als homes.

A la **figura 9**, es pot examinar com es relaciona la variable sexe amb les altres. Com hi ha un nombre de dones més elevat que d'homes la funció de densitat sempre serà una mica més alta en aquelles variables amb poca variabilitat. Es poden apreciar petites diferències en el colesterol HDL, que els homes tenen en general més alt, i també, que aquests tenen un IMC menys variat. També és destacable que en edat homes i dones tenen distribucions molt similars.

Taula 5: Repartició d'episodis en els diferents grups de risc i incidència acumulada

	Baix			Moderat Baix		
	Dona	**Home**	**Total**	**Dona**	**Home**	**Total**
No episodi	651	206	857	764	501	1265
Episodi	2	2	4	3	4	7

	Moderat Alt			Alt		
	Dona	**Home**	**Total**	**Dona**	**Home**	**Total**
No episodi	423	468	891	372	636	1008
Episodi	3	13	16	6	15	21

Figura 9: Funcions de densitat de cada variable i comparació amb els altres biomarcadors segons sexe

3.2.3 Variable Diabetis

Altres variables que la funció de REGICOR pren com a importants per predir el risc, són la diabetis i la relació amb el tabac. A la **taula 6**, es pot veure la relació entre diabetis i els diferents grups de risc, posat que a la **figura 6** no es podia apreciar correctament. Tal com es deia, partint de la poca diferència entre mitges, els voluntaris amb diabetis estan bastant repartits entre grups de risc. Això que fa

pensar que no és una variable determinant en el risc cardiovascular, però en ser un factor de risc clàssic a la literatura, a aquelles persones que pateixen diabetis se les medica per prevenir episodis cardiovasculars [35].

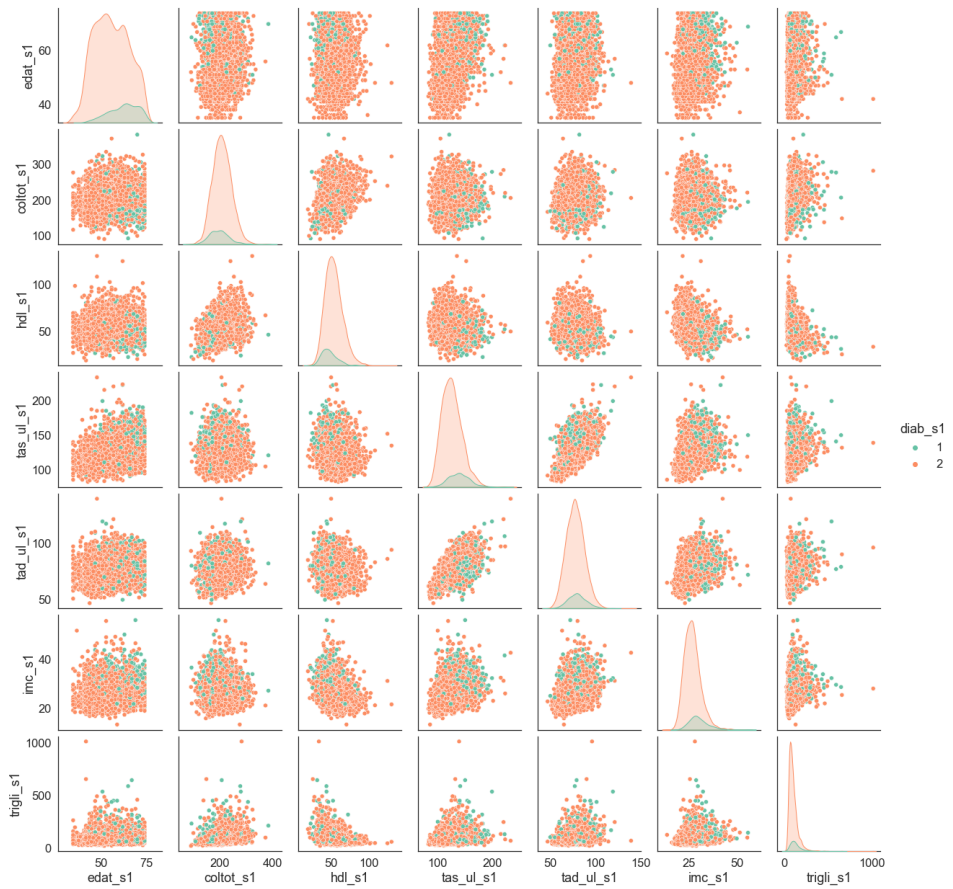
A la **figura 10** es pot apreciar com en el conjunt de persones diabètiques l'edat és major i el colesterol HDL és menor.

Taula 6: Repartició d'episodis segons diabetis i grup de risc

	Baix			Moderat Baix		
	Diabètic	**No diabètic**	**Total**	**Diabètic**	**No diabètic**	**Total**
No episodi	74	787	857	178	1094	1265
Episodi	2	2	4	2	5	7

	Moderat Alt			Alt		
	Diabètic	**No diabètic**	**Total**	**Diabètic**	**No diabètic**	**Total**
No episodi	133	774	891	97	932	1008
Episodi	4	12	16	2	19	21

Figura 10: Funcions de densitat de cada variable i comparació amb els altres biomarcadors segons diabetis



3.2.4 Variable Tabaquisme

D'altra banda a la **taula 7**, es pot contemplar la distribució de la variable tabaquisme. La majoria dels no fumadors, és a dir, els marcats com a 0 a la base de dades, es troben als grups de risc més baix, per tant és lògic que aquests tendeixin a tenir una mitja més baixa a la **figura 6**.

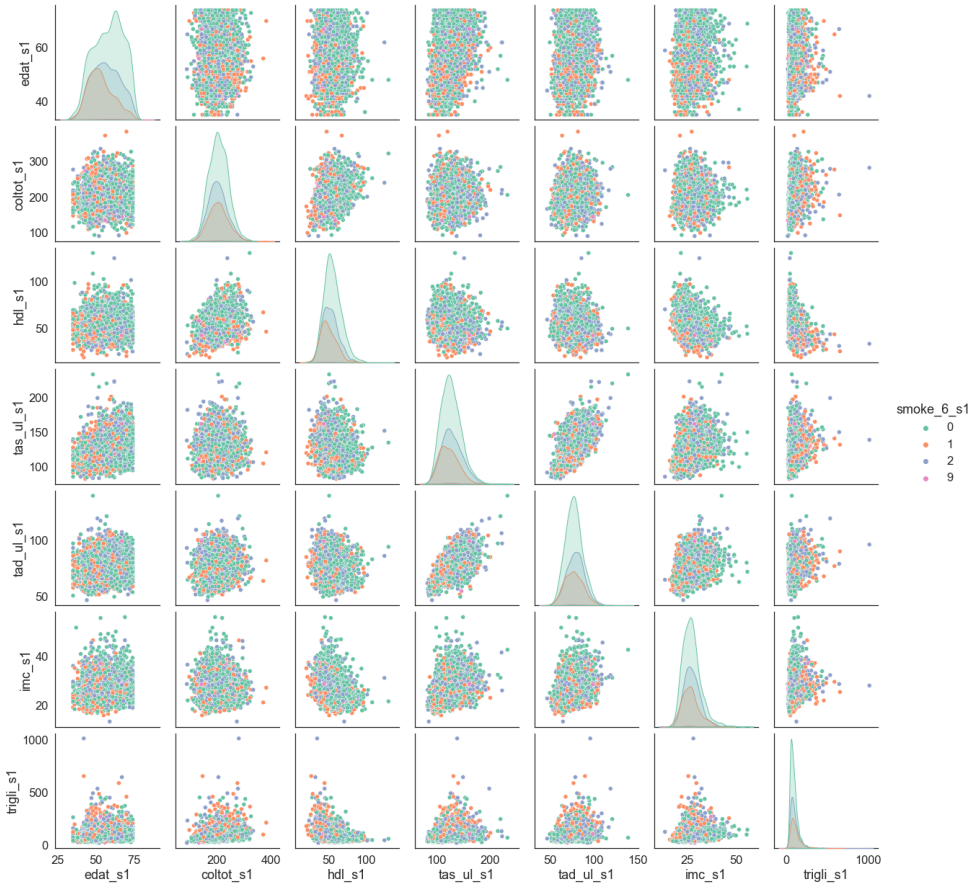
D'altra banda, una altra dada destacable, és que el grup d'exfumadors té més mostres als grups d'estratificació alt, concretament un 70% d'aquest grup, i a més a més, generen més episodis que els fumadors.

A la **figura 11** s'aprecia com es relaciona el tabac amb les altres variables. Cal destacar que en general el col · lectiu de fumadors és més jove que el d'exfumadors, cosa que podria explicar el biaix de risc que hi ha entre les dues classes, atès que en la resta de variables, són similars. Existeix una certa tendència a què els fumadors tinguin valors més baixos, però segurament és a causa de la diferència d'edat.

Taula 7: Repartició d'episodis segons variable fumador i grup de risc

			Baix		
	No fumador	**Ex-fumador**	**Fumador**	**Dades insuficients**	**Total**
No episodi	695	41	125	0	857
Episodi	3	0	1	0	4
			Moderat Baix		
	No fumador	**Ex-fumador**	**Fumador**	**Dades insuficients**	**Total**
No episodi	729	264	288	0	1265
Episodi	3	1	3	0	7
			Moderat Alt		
	No fumador	**Ex-fumador**	**Fumador**	**Dades insuficients**	**Total**
No episodi	450	270	187	0	891
Episodi	9	4	3	0	16
			Alt		
	No fumador	**Ex-fumador**	**Fumador**	**Dades insuficients**	**Total**
No episodi	266	560	182	21	1008
Episodi	4	15	2	0	21

Figura 11: Funcions de densitat de cada variable i comparació amb els altres biomarcadors segons tabaquisme



3.2.5 Variable Hipertensió

La variable hipertensió és la que sembla que pot aportar més informació. A la **taula 8**, es pot apreciar que clarament hi ha molts més individus a risc alt que baix, i que són els que generen més episodis. De fet, de 48 totals, 37 els tenen voluntaris amb hipertensió.

Una de les limitacions de la funció de REGICOR és que els grups d'estratificació del mig acumulen més episodis cardiovasculars que al grup de risc més alt. En aquesta taula, es pot veure com també passa amb els hipertensos, que tenen aproximadament el mateix nombre de participants en els grup grups de risc mitjà, i també són els que generen més episodis.

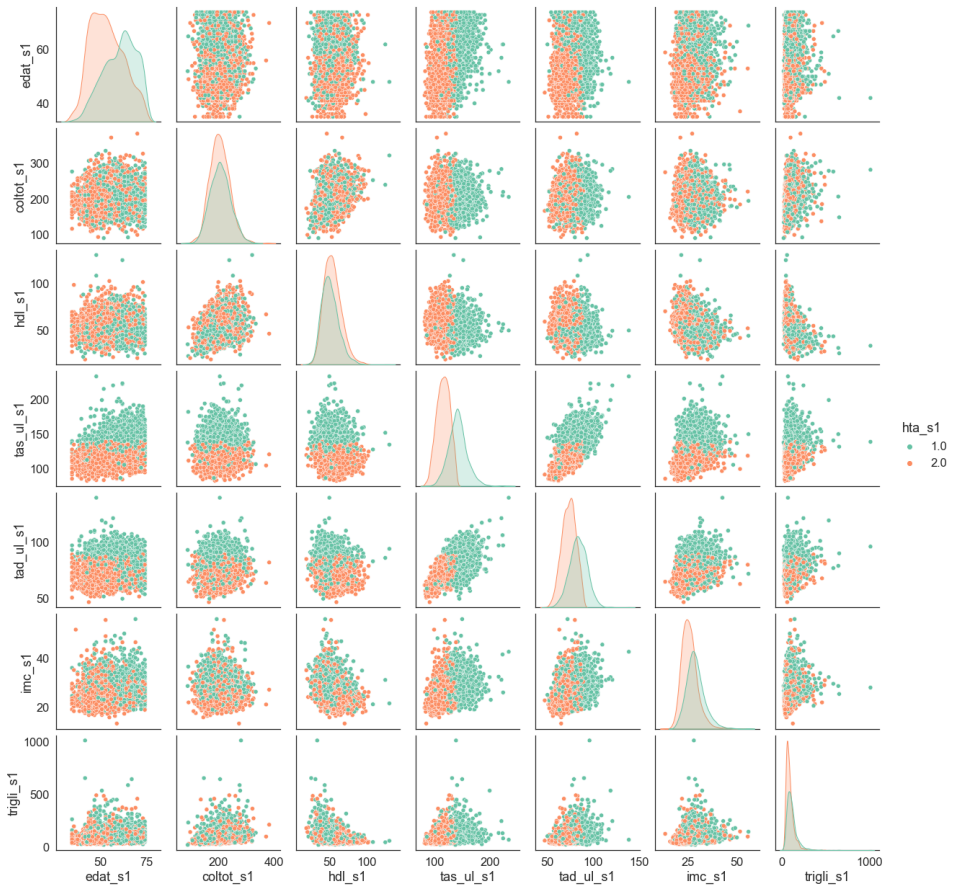
Un camí de millora de la funció de risc, podria ser intentar tenir més en compte aquesta variable donat els arguments anteriors. A la **figura 12** s'observa que en general hi ha una tendència de les persones amb hipertensió a tenir una edat més avançada igual que una pressió sistòlica i diastòlica més elevada. També és remarcable que en general tenen un IMC més elevat.

Taula 8: Repartició d'episodis segons variable hipertensió i grup de risc

	Baix			Moderat Baix		
	No hipertens	**Hipertens**	**Total**	**No hipertens**	**Hipertens**	**Total**
No episodi	737	122	857	852	417	1265
Episodi	1	3	4	2	5	7

	Moderat Alt			Alt		
	No hipertens	**Hipertens**	**Total**	**No hipertens**	**Hipertens**	**Total**
No episodi	491	415	891	302	724	1008
Episodi	5	11	16	3	18	21

Figura 12: Funcions de densitat de cada variable i comparació amb els altres biomarcadors segons hipertensió

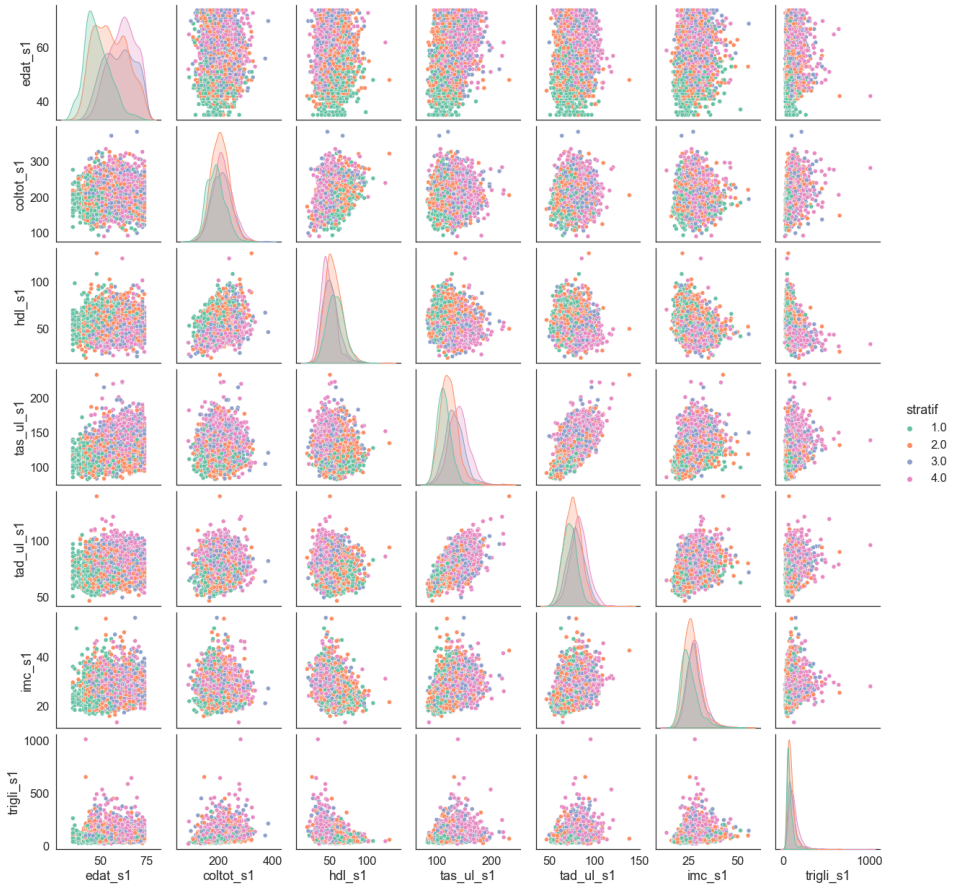


3.2.6 Variable Estratificació

A la **figura 13**, es poden observar les diferents funcions de densitat de cada grup d'estratificació per corroborar les conclusions extretes de la **figura 6**. S'observa que la diagonal és en general molt similar per tots els grups, excepte per l'edat, on es noten oscil · lacions clares. El grup de risc baix té la gent més jove, el de risc mitjà-baix persones de tota la distribució d'edats, el mitjà-alt tendeix a tenir una edat més elevada i el de risc alt els d'edat més avançada.

La resta de les variables, són bastant similars tot i que en general s'aprecia que pels factors de risc negatius, el grup més baix tendeix a valors menors, i per positius, majors.

Figura 13: Funcions de densitat de cada variable i comparació amb els altres biomarcadors segons estratificació

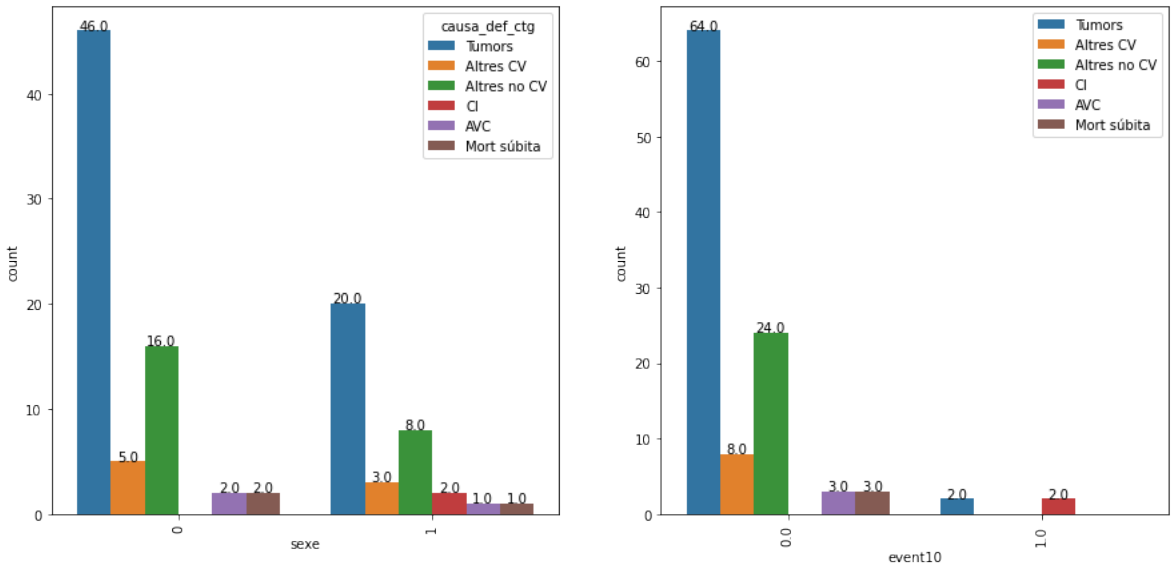


3.2.7 Estudi de defuncions

Es realitza un estudi de les defuncions dels participants que entren dintre els barems de la funció de REGICOR per descobrir si les malalties cardiovasculars són la principal causa de mort.

A la **figura 14** es pot observar com la primera causa de mort són els tumors, seguides d'altres morts no CV. Això pot tenir a veure amb el fet que és un estudi controlat que precisament busca intentar predir el risc cardiovascular, per tant, no s'hauria de prendre com un resultat genèric. Això s'aprecia també a la **figura 14 dreta**, ja que aquells que han patit un episodi cardiovascular tenen menys defuncions que el grup "sa".

Figura 14: Causes de mort per sexe a la dreta i dividit per participants amb episodi i sense a l'esquerra

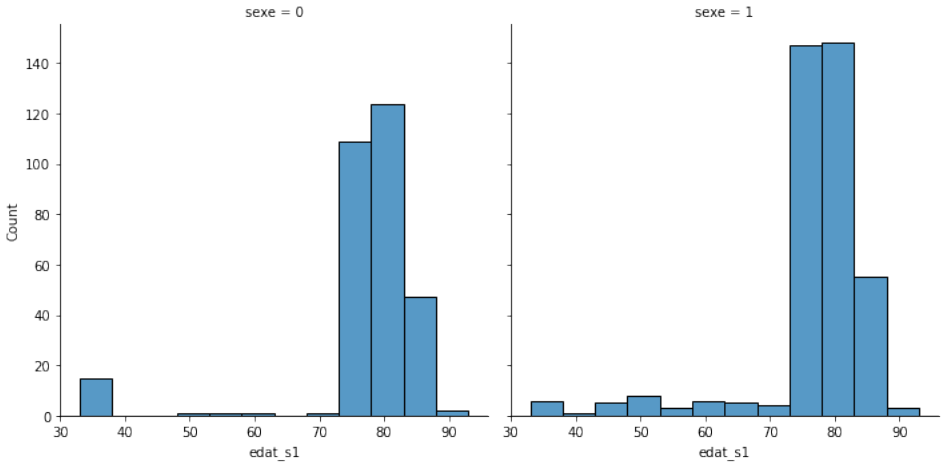


3.2.8 Estudi dels valors desconeguts

A continuació es fa un estudi d'aquells exemples que han estat descartats per ser aplicats a la funció de REGICOR pel fet que no compleixen les condicions necessàries, o perquè tenen dades nul · les entre les 8 variables de risc.

Com es pot observar a la **figura 15** el gruix de persones es troba a les edats superiors a 70 anys i probablement, superiors a 74 anys, de manera que no es poden acollir a la funció de REGICOR. Tots aquells que es troben per sota d'aquesta edat i siguin a risc desconegut segurament es podran tornar a afegir a un model fent una imputació de valors nuls. A la **figura 15** es comprova que no hi hagi un desequilibri a la variable sexe per tal d'afegir-los al conjunt total de manera segura.

Figura 15: Distribució de sexe i edat al grup de persones amb risc desconegut



D'altra banda a la **figura 16** s'aprecia que només hi ha episodis en aquells individus amb edats superiors a 74. Com la idea és entrenar un model que tingui menys limitacions que la funció de REGICOR, aquests episodis es poden utilitzar per fer una predicció més acurada. A la **figura 17** es pot observar que hi ha 57 episodis més.

Figura 16: Distribució d'episodis i edat al grup de persones amb risc desconegut

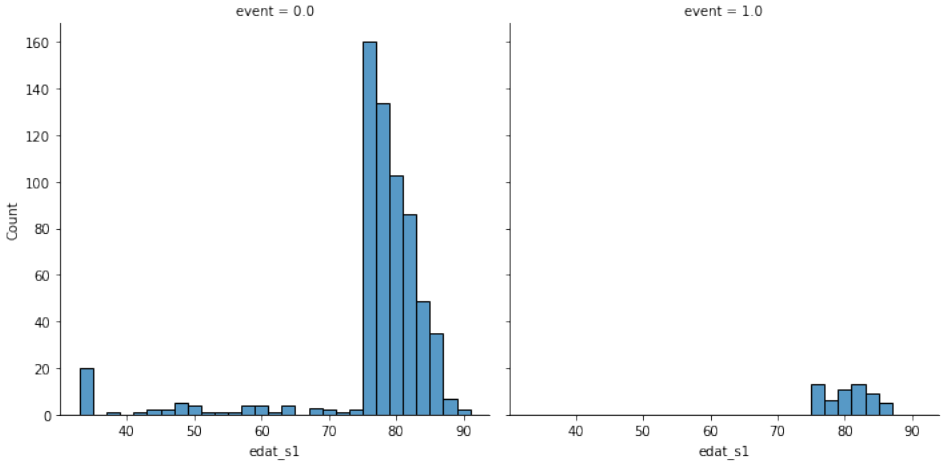
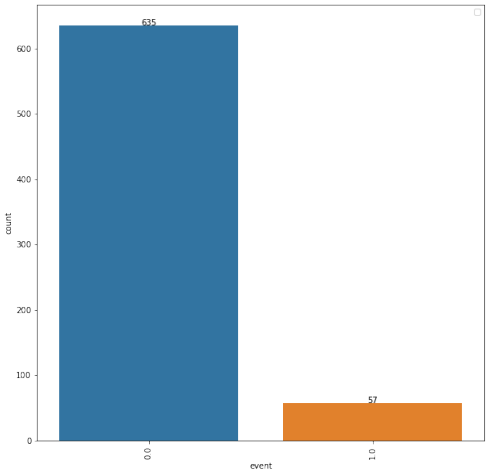


Figura 17: Distribució d'episodis i edat al grup de persones amb risc desconegut



4 Models de predicció

En aquesta secció s'explicaran els diferents models amb els quals s'ha treballat. Però abans és necessari definir una sèrie de conceptes:

Aprenentatge Automàtic L'aprenentatge automàtic o machine learning, és el camp d'estudi que permet als ordinadors aprendre sense haver de ser explícitament programats.

Aquest tipus d'algoritmes tenen diferents maneres de ser entrenats:

Aprenentatge supervisat És una tècnica que intenta obtenir una funció a partir d'uns vectors d'entrada amb els seus respectius vectors de sortida. Quan aquest últim és una variable quantitativa, s'està fent una regressió, si el valor és categòric, una classificació.

Aprenentatge no supervisat És una tècnica que extreu propietats i característiques del conjunt de vectors d'entrada sense la necessitat de tenir-ne un de sortida, s'utilitza generalment per fer una anàlisi exploratòria de les dades o extreure patrons interns d'aquestes.

Aprenentatge per reforç El programa aprèn a partir d'una sèrie de premis o càstigs, si el resultat és bo, se li donarà una recompensa positiva, al contrari, una negativa, de manera que a la següent iteració l'algoritme funcionarà de manera diferent en base aquestes.

4.1 K-Nearest Neighbors

La idea darrere aquest algoritme és estimar la classificació d'una instància no visitada segons la classificació d'altres mostres a prop d'ella. El nombre d'instàncies en què es fixa l'algoritme és d'on prové la K, que defineix el nombre d'elements amb què s'ha de comparar el no classificat.

El pseudocodi bàsic d'aquest algoritme consistiria en el següent:

1. Troba les k instàncies que es troben més a prop de l'element no classificat.
2. Agafa la classe més recurrent segons aquesta distància.

Hi ha moltes maneres de computar la distància entre dos punts A i B, però el càlcul sempre ha de complir els següents requisits:

1. La distància des de A fins a ell mateix sempre ha de ser 0.
2. La distància entre A i B és la mateixa que entre B i A.
3. S'ha de complir la desigualtat triangular, que diu que la distància més curta entre 2 punts és sempre una línia recta.

Un dels mètodes més utilitzats per calcular la distància és l'Euclidiana:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

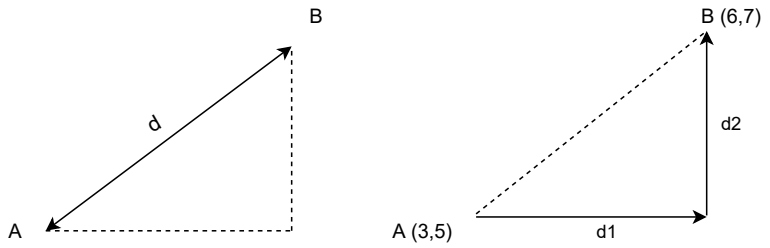
Que aplicada a sistemes n-coordenades es transforma en:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)}$$

I en casos en els quals sigui impossible el càlcul d'una línia recta es calcula la distància de Manhattan, prenent la **figura 18**, la distància de Manhattan entre els punts A i B es calcularia:

$$d_{manhattan} = (A_2 - A_1) + (B_2 - B_1) = (6 - 3) + (5 - 3) = 5$$

Figura 18: Distància Euclidiana i de Manhattan



4.2 Models basats en arbres

Els models basats en arbres es poden utilitzar tant per regressió com per classificació i funcionen dividint l'espai de predicció en regions, primer comença amb tot el conjunt i el va partint d'acord amb una de les variables, de manera que cada segment forma un node de l'arbre. Aquest procés es repeteix fins a arribar a les últimes divisions, que són les fulles.

En el cas de la classificació, per decidir com fer cada decisió aquests models utilitzen l'índex de Gini, que està definit com:

$$G = \sum_{k=1} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

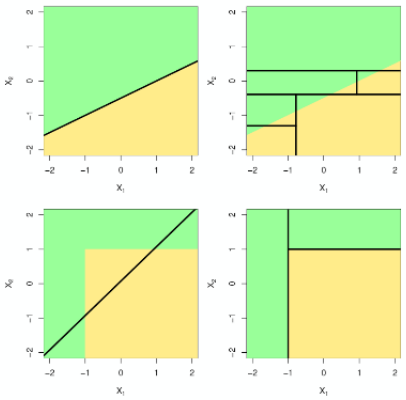
Que és la mesura de la variància a les K classes. Si $\hat{p_{mk}}$ pren un valor proper a 0, l'índex de Gini retorna un valor baix. Es diu que es un indicador de puresa del node, ja que un valor petit vol dir que el node té observacions predominants sobre una classe.

O l'alternativa a aquesta mesura, la cross-entropy, definida com:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Que dona un valor similar a l'índex de Gini.

Figura 19: Models lineals vs Arbres, extreta de [32]



No obstant això, els arbres de decisió sols no són uns grans classificadors, a la **figura 19** es pot observar com funcionen correctament per dades no lineals però no tan bé per dades lineals.

Existeixen altres mètodes com el Bagging, Random Forest o Boosting que fusionen múltiples arbres, obtenint així una gran millora en la predicció. Són els anomenats algoritmes d'ensemble o la seva traducció al català, d'acoblament.

4.2.1 Bagging

El Bagging o Bootstrap Aggregation és un algorisme de propòsit general que s'utilitza per reduir la variància d'un mètode d'aprenentatge estadístic.

Donat un grup d'n observacions $Z_1 \dots Z_n$, cada una amb una variància σ^2, la variància de la mitja $\hat{Z}$ és definida per $\frac{\sigma^2}{n}$.

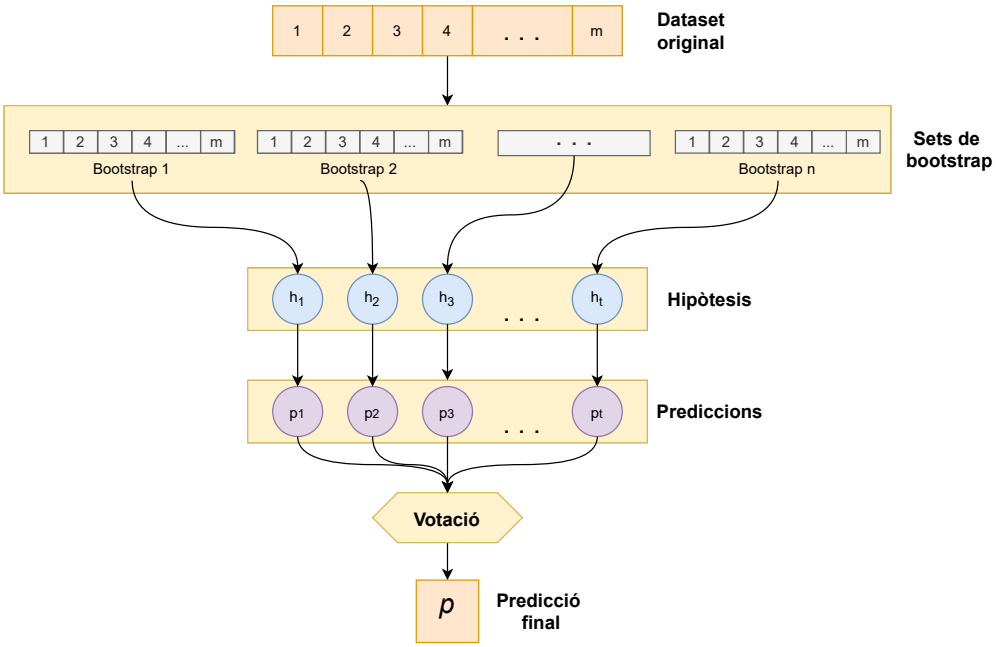
En cas de tenir n conjunts de dades, es pot reduir la variància dels resultats fent $\frac{\sigma^2}{n}$. Però normalment no es té més d'un grup de dades, per tant, s'utilitza el mètode de bootstrap al dataset. S'agafen mostres de les dades d'entrenament utilitzant un mostreig aleatori amb repetició, generant així B conjunts de dades mitjançant.

Després s'entrenarà el model utilitzant un set b_i per obtenir $\hat{f}^{*b}(x)$, la predicció al punt x. I per últim es fa la mitja de tots els resultats, és a dir:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

La **figura 20** il · lustra aquest tipus d'algoritmes.

Figura 20: Bagging, adaptada de [32]



4.2.2 Random Forest

Els Random Forest són una millora dels algorismes de Bagging, fent una petita modificació que aconsegueix reduir la correlació entre arbres i per tant la variància en fer la mitja dels resultats dels arbres.

Igual que al Bagging, primer es construeixen una sèrie d'arbres a partir de sets generats utilitzant bootstrap, però cada cop que es crea una branca als arbres, es té en compte un subconjunt aleatori m del conjunt total de predictors p, però del qual només s'utilitza un predictor. Per cada divisió s'escull una nova remesa de predictors m i normalment s'agafen aproximadament $\sqrt{p}$ predictors.

4.2.3 Boosting

Els algorismes de Bagging i Random Forests, generen els arbres paral · lelament, és a dir, no tenen comunicació entre ells. En canvi, els models de Boosting, generen els arbres seqüencialment, de manera que es creen tenint informació dels anteriors. També, al contrari que els altres dos mètodes, el Boosting pot fer overfit si el nombre d'arbres és molt elevat, tot i que de manera lenta.

L'algoritme és senzill, a cada iteració es procedeix a crear l'arbre tenint en compte els residuals de l'arbre anterior, de manera que el nou arbre millori el rendiment en aquells llocs on l'anterior s'ha equivocat. A més a més, controlant la ràtio d'aprenentatge, es pot aconseguir un aprenentatge lent, permetent crear arbres de diferents formes per atacar els residuals amb més precisió.

4.2.4 Hybrid Random Forest Linear Model

Els models Hybrid Random Forest Linear Model o HRFLM no entren exclusivament dintre els models que funcionen amb arbres, malgrat això, s'afegeixen a aquest apartat posat que segueixen una filosofia semblant. Els models híbrids fusionen diferents algoritmes clàssics de ML i creen un nou algoritme d'ensemble basat en els resultats d'aquests. Es pot escollir entre agafar simplement la classe que ha estat escollida majoritàriament o agafar la mitja de les probabilitats i d'aquí fer la classificació.

4.3 Support Vector Machines

Els support vector machines (SVM) busquen trobar un pla que separi les classes a l'espai de dades. Es necessari definir una sèrie de conceptes abans:

Hiperplà Un hiperplà en dimensions p és un subespai afí pla de dimensió $p-1$, de manera que a $p=2$ dimensions l'hiperplà és una recta. La seva equació general és:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

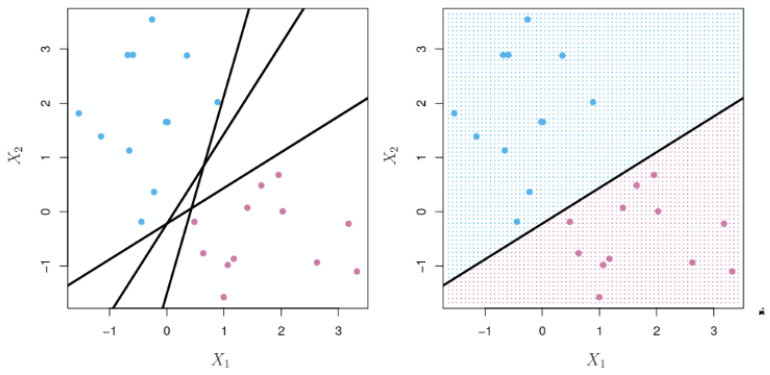
Si $\beta_0 = 0$ l'hiperplà passa per l'origen, i igual que a un pla normal, el vector $\beta = \beta_1, \beta_2, \ldots, \beta_p$ és el vector normal i és perpendicular a l'hiperplà. En cas de voler utilitzar aquest pla per dividir dues classes:

si

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

substituint un punt X a l'espai aleshores obtenim $f(X) > 0$ o $f(X) < 0$ segons de la zona on es trobin del pla.

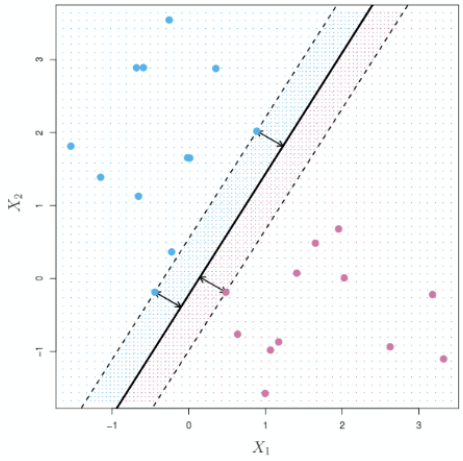
Figura 21: Possibles plans i divisió de les dades, valors en negatiu pintats en vermell i positius en blau, extretes de [32]



Com es pot observar a la **figura 21**, poden existir diferents plans que tallin les dades de manera perfecta, per tant, cal definir un criteri per seleccionar quin és el

millor pla. Un d'aquests mètodes és el *maximal margin classifier*, que escull el pla que té la distància més gran entre els punts més pròxims a aquest tal com es veu a **figura 22**

Figura 22: Pla que equidista el màxim de les dues classes, extreta de [32]

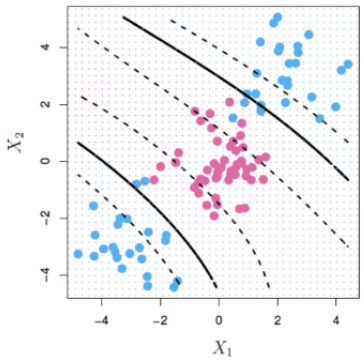


Això no obstant, aquestes dades són fàcilment separables, però la realitat és diferent, hi ha molts datasets en els quals les dades són sorolloses, no són separables o simplement no tenen límits lineals.

1. **Dades sorolloses**: El SVM maximitza un soft margin en comptes d'un maximal.
2. **Dades no separables**: En aquest cas, sempre hi haurà un punt que no satisfarà els límits, per tant, es deixa que alguns punts es classifiquin malament i s'afegeix un terme de cost al càlcul de la distància per penalitzar-ho, és el paràmetre C. Un valor alt vol dir una penalització alta.
3. **Espai de dades no lineal**: En aquesta situació, és necessari definir un altre tipus de metodologia per dividir les classes.

   Feature Expansion Es transformen les dades de manera que si es té un espai definit per X_1, X_2 es fan transformacions del tipus $X_1, X_2, X_1^2, X_1X_2 \ldots$, així es passa d'una dimensionalitat ρ a una $M > \rho$.

Figura 23: Separacions no lineals als SVM, extreta de [32]



Una vegada fet aquest pas, s'entrena el model en aquest nou espai, cosa que resulta en límits no lineals a l'espai de dades original, ja que les noves fronteres estarien definides com:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_3 X_1 X_2 = 0$$

Seguint els exemples anteriors. A la **figura 23** hi ha un exemple de com quedarien les fronteres.

Kernels No obstant això permeti solucionar els problemes, les solucions polinòmiques poden arribar a ser molt costoses, per tant, s'introdueix una altra manera per calcular solucions no lineals, els kernels.

Un SVM lineal es pot representar com:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i (x, x_i)$$

on n són el número de paràmetres. Per estimar els paràmetres $\alpha_1, \ldots, \alpha_n$ i β_0 es necessiten els $\binom{n}{k}$ dels productes interns (x, x_i) entre tots els parells d'observacions.

Però resulta que la majoria dels paràmetres poden donar 0, per tant, la formula passa a ser:

$$f(x) = \beta_0 + \sum_{i \in S}^{n} \hat{\alpha}_i (x, x_i)$$

On S és el vector d'índex en el qual $\hat{\alpha}_i > 0$, també conegut com el vector de suport. Són aquells punts que es troben més a prop de l'hiperplà divisor i els que acabaran definint el comportament d'aquest de manera més decisiva.

Per tal de decidir quin pla és el que separa al màxim les dades, és necessari maximitzar els marges amb la condició que no hi hagi punts entre ells. Es pot deduir que la distància entre els marges és:

$$dist(H_1, H_2) = \frac{2}{||w||}$$

on w és un vector de pes i $H_1 i H_2$ són els plans equidistants del classificador. Per tant per maximitzar la distància es necessari minimitzar $||w||$, cosa que es pot fer amb la forma primal del classificador que és:

$$f(x) = w^T x + b$$

Però només és útil en espais de dades lineals. Per tant, per ampliar-ho a espais no lineals, es pot utilitzar el concepte de dualitat a optimització, si abans s'estava minimitzant una variable per obtenir la millor variable, ara es buscarà maximitzar-ne una per arribar al mateix resultat. La formula dual del problema és:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Que pot ser reescrita com:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i^T x_j)$$

De manera que l'optimització no canvia i es poden introduir noves dimensionalitats a partir d'un producte escalar.

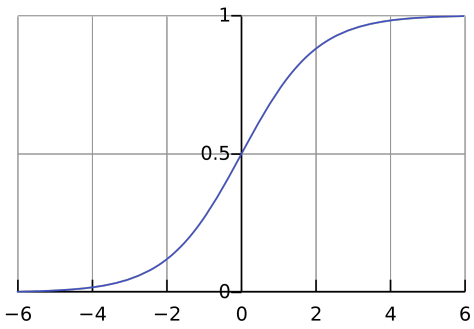
4.4 Regressió logística

Aquest model pren el seu nom de la funció que utilitza com a base, la funció logística també coneguda com a funció sigmoidal, i permet modelar les probabilitats de pertànyer a una classe a través de funcions lineals en x definida per:

$$Logistic(\overrightarrow{\omega} * \overrightarrow{x}) = \frac{1}{1 + e^{-\overrightarrow{\omega} * \overrightarrow{x}}}$$

El procés d'ajustar els pesos per fer la predicció és l'anomenada regressió, i es basa en un procés anomenat *semblança màxima* que funciona ajustant els pesos de manera que els resultants sempre siguin els més probables. A la **figura 24** es pot observar la corba Logística, que s'utilitzaria per classificar, si el valor és molt proper al 0 serà una classe, i si no, serà l'altra.

Figura 24: Exemple de corba Logística, [32]



4.5 K-Means

El K-means és un algoritme de clustering no supervisat, que genera grups de dades no superposats, és a dir, que cada dada només forma part d'un conjunt. És un algoritme molt senzill que només necessita el nombre de clústers K per començar a funcionar.

Donat un conjunt de dades no classificades, s'assigna un clúster aleatori de valor entre 1 i K a cada mostra, a manera d'estat inicial de l'algoritme.

A continuació s'itera sobre els clústers fins que deixin d'haver-hi noves assignacions:

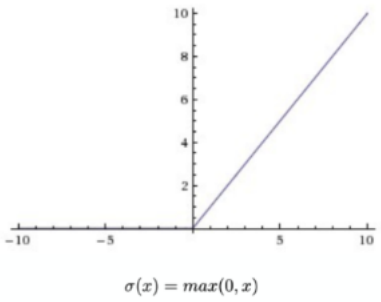
1. Per cada un dels K clústers es computa el seu centroide, que és el vector de característiques mig dels elements que hi ha a un clúster.
2. A cada individu se li reassigna el clúster segon el centroide més pròxim a ell.

4.6 Deep Learning

Els algoritmes de Deep Learnig formen part de la família dels algoritmes de Machine Learning, però les estructures dels algorismes reben el nom de Xarxes Neuronals.

Les xarxes neuronals estan formades per unitats anomenades neurones per la seva semblança de funcionament amb les del cervell. Aquestes calculen una funció segons els inputs i la propaguen cap a les següents neurones, l'aprenentatge apareix quan es canvien els pesos que connecten les neurones i es minimitza l'error comès a les prediccions. L'ajustament de molts parells de neurones permet refinar la funció al màxim fins al moment de ser capaç de reconèixer dades que no havia vist abans. La informació entre neurones s'envia mitjançant funcions d'activació, que interpreten els valors anteriors i extreuen un valor per propagar. Algunes de les més utilitzades són les funcions ReLU, **figura 25** o la sigmoide, que s'ha mencionat també a la **secció 4.4**.

Figura 25: Funció ReLU, extreta de [32]



Donat un vector de característiques, la xarxa neuronal ajusta els pesos de cada un dels components per decidir quins són els més importants i quins prescindibles. En aquest treball s'utilitzen xarxes neuronals senzilles de dues o tres capes de neurones, amb diferents paràmetres que cal explicar.

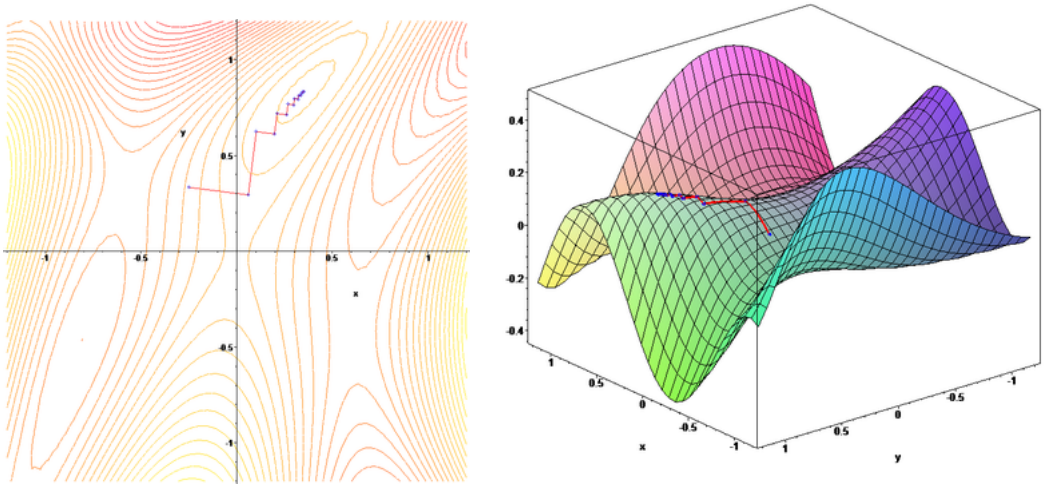
4.6.1 Optimitzadors

Els optimitzadors són algoritmes que defineixen com han de canviar els pesos de les neurones per aconseguir el millor valor de loss possible. El loss és la semblança de l'actual predicció a la realitat.

El descens de gradient és el més bàsic de tots, i es basa a trobar el valor mínim de la primera derivada del loss mitjançant variacions dels pesos. Per cada paràmetre calcula el gradient mínim i fa un pas en aquella direcció, el paràmetre que regula la distància del pas és l'anomenada ràtio d'aprenentatge.

Depenent del valor d'aquesta, l'algoritme es pot passar del mínim local si és molt gran, o pot simplement trigar a trobar-lo si és molt baix. A la **figura 26** es presenta un exemple il · lustratiu. A la dreta hi ha la representació en 3D de la funció mentre que a l'esquerra la 2D. A la primera, els cercles més extensos són els mínims, i com es pot observar n'hi ha de diferents i de més o menys profunds. L'objectiu és arribar sempre al mínim màxim i no quedar-se en un local. En vermell es marca el camí que fa el descens de gradient per arribar al seu, i s'hi poden apreciar els diferents passos que fa.

Figura 26: Descens de gradient, *primera* i *segona* extretes de Wikimedia Commons.



No obstant el descens de gradient és un algoritme senzill, utilitza totes les dades per computar el gradient, cosa que en bases de dades molt grans pot ser molt costós. També, és complicat escollir un valor d'aprenentatge eficient, i com s'ha explicat, és un paràmetre essencial.

És per això que existeixen optimitzadors que regulen el seu comportament de manera interna per aconseguir una major adaptabilitat al problema, guardant dades dels anteriors gradients calculats per agilitzar el procés i fent modificacions segons calgui per arribar al mínim com més aviat millor.

4.6.2 Mètodes de Regularització

Hi ha diferents mètodes per evitar que les xarxes neuronals facin overfit quan entrenen a més a més dels optimitzadors. Algunes d'aquestes són:

- **Dropout**: Consisteix a eliminar una selecció aleatòria de neurones a cada capa. Com més es descartin, més regularització s'aconsegueix, ja que sovint, a l'haver-hi moltes connexions, aquestes poden esdevenir soroll i afectar negativament al model.
- **Early Stopping**: Consisteix a acabar l'entrenament abans d'hora. El loss es calcula comparant amb dades que es guarda la xarxa per validar els resultats i perquè l'algoritme funcioni correctament aquest ha de disminuir amb les iteracions. Tanmateix, pot ser que després de fer un descens, el valor torni a pujar perquè l'algoritme es passi el mínim local, és en aquest moment quan l'early stopping atura l'entrenament per evitar un pitjor resultat.
- **Bias**: Els valors dels pesos de les neurones en inicialitzar el model són aleatoris. Però en cas de saber quina distribució tenen les dades, és una bona pràctica canviar aquests pesos de manera que tinguin en compte les proporcions de la base de dades. Això és especialment útil en dades no balancejades.

4.6.3 Estructura de la Xarxa Neuronal

En aquest apartat s'exposarà l'arquitectura de la xarxa neuronal utilitzada, il · lustrada a la **9**. Com es pot observar conté com a mínim 17218 paràmetres, per la qual cosa es pot considerar una xarxa neuronal petita.

Taula 9: Estructura de la xarxa neuronal

Tipus de capa	**Forma**	**#Paràmetres**
Input	64	64 +(64 * #variables)
Dropout	64	0
Densa	256	256 + (256*64) = 16640
Output	2	2 + (256*2) = 514
Total de paràmetres		**17218 + (64 * #variables)**

Altres de les seves característiques són:

- Dropout: S'utilitza un dropout de 0.5
- Funcions d'activació: S'utilitza ReLU en totes les capes menys l'última, on s'aplica una funció sigmoidal per tal de fer la predicció probabilística.
- Optimitzador: S'utilitza NAdam.
- Bias: S'inicialitza la xarxa amb un careful bias calculat com:

$$bias = \log(\frac{\#positius}{\#negatius})$$

- Early Stopping: S'implementa un early stopping amb una paciència de 50 epochs.

5 Tècniques de validació i processament de dades

5.1 Metodologies de validació

En aquest apartat s'explicaran diferents metodologies que s'han utilitzat per tal de refinar al màxim els diferents models que s'han escollit.

5.1.1 Grid-search Cross-Validation

El cross-validation (CV) és una metodologia de validació de resultats que consisteix a efectuar una divisió de les dades, de manera que el model s'entrena amb unes i s'avalua amb unes altres que no ha vist per tal de simular un cas real. Aquest és el mètode més clàssic, el de divisió de train i test. Tanmateix, és sensible a les particions que es facin, és a dir, si el set de test no té les mateixes distribucions que el de train el resultat pot variar. També pot passar que l'entrenament que s'hagi fet amb les dades de train hagi estat coincidència i que no sigui general per totes les dades.

No obstant existeixen altres estratègies per tal de validar l'efectivitat del model que es vulgui utilitzar. Un d'aquests és el k-fold cross validation, que consisteix a dividir de manera aleatòria les dades en k grups o en anglès *folds* d'una mida aproximadament igual. El primer fold s'utilitza com a conjunt de validació (com si es tractés de la primera explicació) i es calcula una mètrica per tal d'avaluar el resultat. Aquest procés es repeteix k vegades, en els quals es torna a començar l'aprenentatge i el conjunt de variació va canviant amb cada iteració.

Així doncs, s'aconsegueix que el resultat del conjunt de proves aproximi millor l'error que es pugui generar a test, ja que s'han fet diferents proves amb distribucions semblants de les dades aconseguint així generalització.

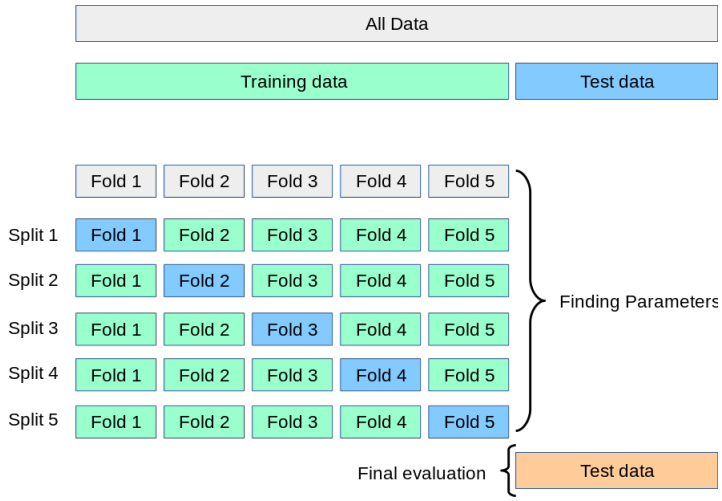
A tot això, cal sumar que tots els models tenen uns paràmetres que poden fer canviar els resultats de manera dràstica, per tant és important buscar el millor set per cada problema nou. Per exemple, com s'ha explicat abans, els SVMs poden estar treballant en un espai de dades on aquestes són separables, o no. Els paràmetres que el model ha de fer servir per a cada cas són molt diferents, però quins són els millors?

Aquí entra en joc el Grid-Search, que és una metodologia que se suma al procés de CV per trobar el millor set de paràmetres pel model que s'estigui entrenant. Per començar és necessari definir quins són els paràmetres que es volen tenir en compte per trobar el millor predictor i els valors que es vulguin provar. Això es pot realitzar fent un estudi bibliogràfic de les possibilitats o simplement seleccionar-los tots i atribuir-los-hi valors aleatoris dintre d'uns límits. Després a cada partició de dades del CV, es provaran totes les combinacions possibles amb els paràmetres que se'ls hi passin i s'agafa la combinació que doni els millors resultats. A la **figura 27** es pot apreciar el funcionament de la GridSearch - CV.

Aquests dos processos tenen un alt cost computacional, però asseguren reduir al

màxim l'error de test i que els resultats de train siguin fiables.

Figura 27: Cross validation, extreta de la *documentació de scikit-learn.*



5.1.2 Mètriques

Com que no existeix un mètode estadístic per mesurar si el resultat d'un model és correcte o no, existeixen mètriques que computen que tant bé treballa un model a partir dels seus resultats. Aquests es mesuren en:

- **TP**: True positive, que són aquelles variables positives que han estat ben classificades.
- **FP**: False positive, que són aquelles variables negatives que han estat classificades com a positives
- **TN**: True Negative, que són aquelles variables negatives que han estat classificades correctament.
- **FN**: False Negative, són aquelles variables positives que han estat mal classificades.

Entenent variable positiva com la variable objectiu, és a dir, en el cas del problema que s'intenta solucionar, seran els episodis cardiovasculars, mentre que la classe negativa serà no patir-ne un.

A continuació s'explicaran les diferents mètriques que s'utilitzen per avaluar els diferents models explicats al **capítol 4**:

- **Eficàcia, accuracy o AC**: És la proporció del nombre total de prediccions correctes:
$$AC = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Sensibilitat, recall o RVP**: És la proporció de casos positius que s'han

identificat correctament:

$$RVP = \frac{VP}{VP + FN}$$

- **Especificitat o RVN**: És la ràtio de casos negatius que han estat identificats correctament:

$$RVN = \frac{VN}{VN + FP}$$

- **Precisió o VPP**: És la proporció de casos positius ben predits entre el total de casos positius predits:

$$VPP = \frac{VP}{VP + FP}$$

- **F1-score**: És la mitja harmònica de la precisió i la sensibilitat:

$$F1 = 2 * \frac{VPP * RVP}{VPP + RVP}$$

- **Eficàcia Balancejada**: És una versió de l'accuracy original, però adaptada per datasets que tinguin dades desequilibrades.

$$BACC = \frac{RVP + RVN}{2}$$

5.2 Preprocessament de dades

5.2.1 Normalització de les dades

Es tenen dades que provenen d'una anàlisi de sang, i que per tant són numèriques i contínues, mentre que d'altra banda es poden tenir diagnòstics o condicions com poden ser la hipertensió, la diabetis o el tabaquisme, que són dades categòriques i finites.

Quan es passa el vector de dades a un model, aquest llegeix els seus components i intenta establir uns pesos a partir d'aquests. Però si li arriben components del vector que segueixen distribucions totalment diferents, li causarà confusió. És aquí on és necessari fer un preprocessament de les dades, per tal de fer-li el treball el més senzill possible.

Molts algoritmes utilitzen la distància euclidiana per a calcular diferències entre vectors de característiques, de manera que si existeixen dues variables amb distribucions molt diferents, com per exemple, l'edat i el colesterol total, la segona sempre tindrà valors superiors a l'edat, i per tant sempre tindrà més pes a la predicció. Però com s'ha vist a l'anàlisi de la base de dades, l'edat és un biomarcador molt important.

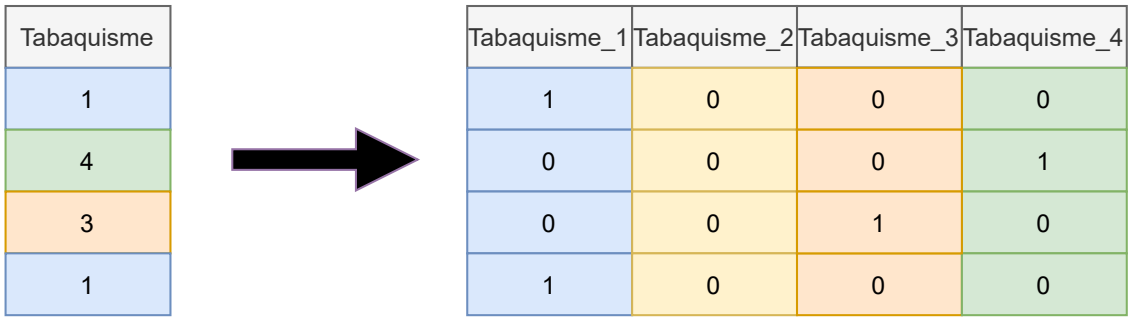
És per això que és necessari normalitzar les dades, és a dir, encabir-les dintre d'un rang d'entre [0,1] per tal que totes les característiques dels vectors siguin comparables.

5.2.2 Codificació categòrica

No obstant això, tot i que en les dades numèriques l'exercici de normalització és suficient, amb les categòriques és necessari replantejar el concepte. Mentre que a les numèriques és cert que un valor més alt o més baix és comparatiu, en el cas de les categòriques això no té per què ser cert, és a dir, si es té que els valors per definir l'estat del tabaquisme del participant són 1,2,3 i 4, l'algoritme sempre donarà més importància al valor més alt, però interessa que tingui en compte les 4 classes per igual.

Per corregir aquesta problemàtica s'utilitza un mètode anomenat *One Hot Encoding* en el qual els valors de la variable categòrica passen a convertir-se en un conjunt de vectors binaris per cada valor que tingui, on hi haurà un 1 en aquells vectors que corresponguin a una classe de la variable. A la **figura 28** es pot observar com funcionaria per l'exemple del tabaquisme.

Figura 28: One Hot Encoding



5.2.3 Outliers

Una altra cosa a tenir en compte amb les dades, són aquelles que es troben als límits de la distribució normal gaussiana. Aquestes dades poden ser introduïdes a la base de dades per culpa d'errors de mesura o simplement errors humans.

A l'hora de calcular les distribucions per normalitzar el conjunt de les dades, és important tractar aquestes dades, ja que la distància entre el valor màxim dintre de la normal i el valor d'un outlier pot ser molt elevada.

Hi ha diferents maneres de treballar aquest tipus de dades, es poden eliminar directament de la base de dades, utilitzar mètodes d'imputació per a calcular un nou valor per aquesta entrada, via regressió o simplement aplicant la mitja de la variable en concret.

De totes maneres, primer s'han de detectar i després tractar. Un mètode útil en aquests casos és la Z-score, definida com:

$$Zscore = \frac{X_i - \overline{X}}{\sigma_x}$$

La Z-score és un valor sensible als valors extrems. Aquesta score donarà un valor que es trobarà entre el -100 i 100, que indicarà com de diferent de la resta de mostres és. A partir d'aquí només s'ha de definir una sensibilitat de quin grau de diferència es permet. Per exemple si es decidís que aquest valor és 90, totes aquelles entrades que siguin diferents del 90% de les altres observacions, seran tractats com outliers.

5.2.4 Dades Desconegudes

Les dades desconegudes tenen un origen semblant als outliers. Són aquelles entrades a la base de dades que no tenen cap valor, això pot ser degut a un error humà o que simplement no s'ha pogut obtenir la dada. Per exemple, si a un dels voluntaris de l'estudi de REGICOR no se li va poder fer l'analítica, els valors que s'haurien d'extreure d'aquesta queden buits.

A un model no se li poden passar dades nul · les, és per això que és necessari aplicar mètodes per substituir-les com als outliers. Novament es pot buscar la mitja de la columna de les dades o buscar un mètode més refinat. Aquest pot ser, per exemple, aplicar una regressió lineal a les dades del voluntari per tal d'aproximar al màxim el valor de la dada nul · la.

5.3 Dades desequilibrades

Una base de dades no balancejada suposa un problema de partida pels models, que estan preparats per treballar amb conjunts de dades que tinguin les mateixes distribucions.

Davant un dataset desequilibrat, l'algoritme tendirà a afavorir de manera errònia la classe majoritària, per exemple, en el cas de la database de REGICOR, hi han 4069 persones sense episodi i 48 que si en presenten,

$$\%episodis = \frac{48}{4069} * 100 = 1.18\%$$

És a dir, els episodis suposen un 1.18% de les dades totals, si es calcula l'accuracy ignorant totalment aquesta classe:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{0 + 4069}{0 + 4069 + 0 + 48} = 0.988$$

S'està aconseguint una eficàcia a priori molt bona, però que en realitat està ignorant per complet una de les classes a predir. Hi ha diferents estratègies per tal de compensar aquest error.

Però en aquest tipus de problema, no només existeix la problemàtica del nombre d'instàncies i tal com diu a [**22**] existeixen els següents escenaris:

- **Descompensació de classes**: És el problema amb què s'ha obert aquest apartat de la memòria, es poden utilitzar mètodes per ajustar les ràtios de cada classe o aproximar els pesos.

- **Superposició de classes**: Quan les classes a predir es troben molt a prop de les altres cosa que fa que els classificadors tinguin problemes per diferenciar entre classes.
- **Disjunció de classes**: Les instàncies de les classes a predir es troben totalment repartides per l'espai de les dades.

A continuació es plantegen algunes de les solucions que s'han trobat.

5.3.1 Recàlcul del pes de les classes

Es pot intentar calcular un pes per cada una de les classes, dividint per 2 (cas binari) es manté la pèrdua a una magnitud similar:

$$pespositiu = \frac{4117}{4069 * 2} = 0.51$$
$$pesnegatiu = \frac{4117}{48 * 2} = 42.39$$

Ara es torna a calcular l'accuracy, però multiplicant per aquests valors:

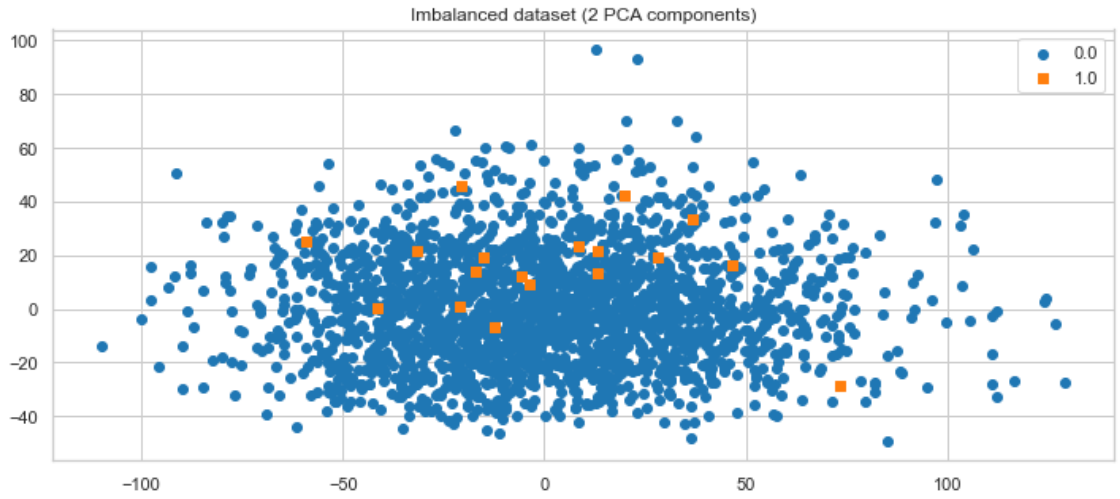
$$AC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{0 + (4069 * 0.51)}{0 + (4069 * 0.51) + 0 + (48 * 42.39)} = 0.50$$

Al predir de manera incorrecta la meitat de les classes, l'accuracy resulta en 0.50 gràcies al càlcul de pes.

5.3.2 Resampling

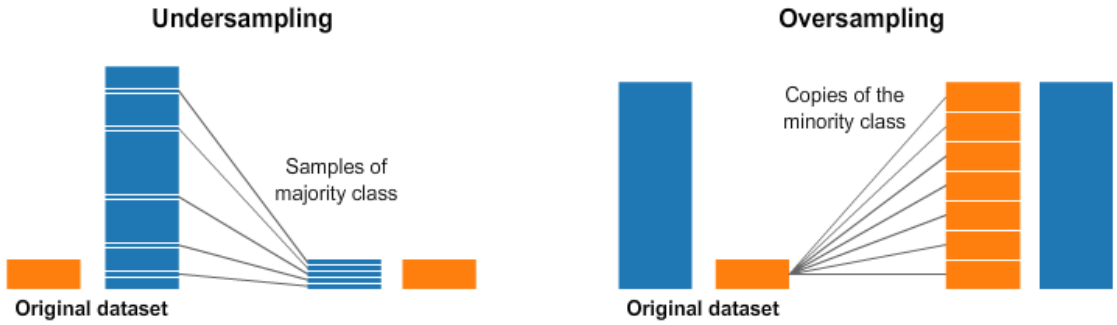
La **figura 29** és una representació d'un fragment de les dades després de fer una PCA de dos components. Com es pot observar hi ha moltes instàncies de la classe 0 i només unes poques de la 1, també presenta problemes de disjunció i superposició.

Figura 29: Dataset no balancejat d'exemple



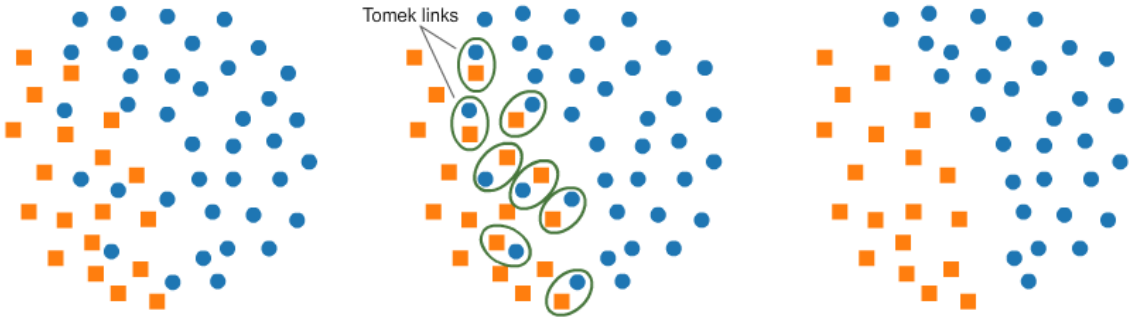
Es poden aplicar mètodes d'upsampling o downsampling per equilibrar el conjunt de dades. La primera consisteix a crear dades artificials de la classe minoritària mentre que la segona es basa a rebaixar al classe majoritària tal com s'aprecia a la **figura 30**. No obstant això, no són mètodes que funcionin a la perfecció i hi ha diferents maneres d'utilitzar-los. La més senzilla és o duplicar instàncies de la classe minoritària, cosa que pot produir sobreestimació o eliminar dades de manera aleatòria de la classe majoritària, cosa que pot fer que es perdi molta informació.

Figura 30: Mètodes de resampling, extretes de *kaggle*



Tomek Links El Tomek links es un algoritme que busca reduir les instàncies de la classe majoritària, tal com es pot veure a la **figura 31**. L'algoritme elimina aquells nodes que es troben a prop de la classe minoritària. Funciona de manera semblant a un KNN. S'anomena Tomek link a aquells dos veïns més pròxims que formen part de dues classes diferents, eliminant el que pertany a la classe majoritària, s'aconsegueix reduir el soroll i fer que les instàncies de la classe desequilibrada es puguin detectar més fàcilment.

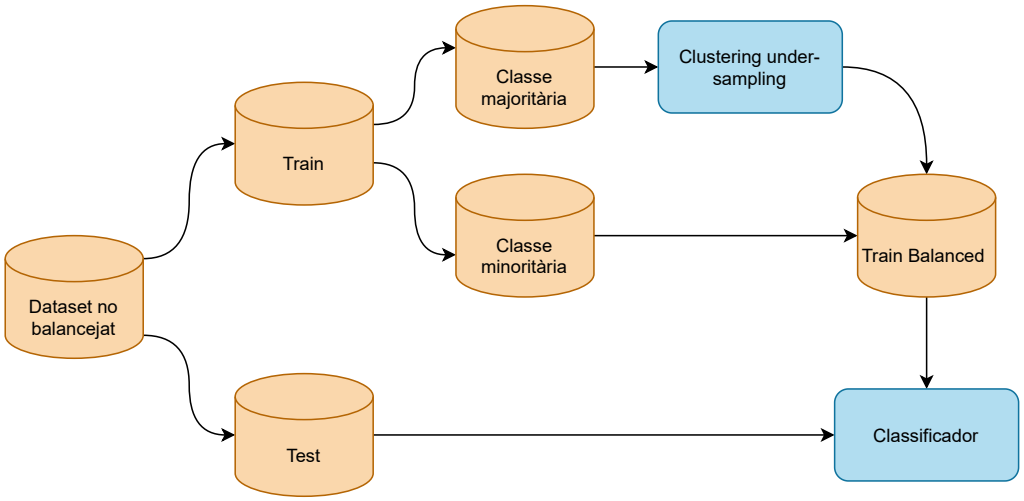
Figura 31: Tomek links, extreta de *kaggle*



Cluster centroids Donat un dataset que es troba desequilibrat, primer s'ha de dividir en train i test, i a continuació dividir el primer en classe minoritària i majoritària. A aquesta última se li aplicarà l'algoritme de clustering per reduir els seus números i es tornarà a combinar amb la classe desequilibrada, resultant en un conjunt de train més equilibrat. Per tal d'utilitzar aquesta metodologia es poden emprar diferents tipus d'algoritmes de clustering. Un d'ells podria ser, que el nombre de clústers sigui el total d'instàncies de la classe minoritària, de manera que un algoritme com el k-means trobi els diferents centres d'aquests clústers, que acabaran substituint el conjunt anterior. D'altra banda, també es pot utilitzar l'aproximació que en comptes de substituir les dades reals pels centroides, s'agafi el veí més pròxim del centre, de manera que les dades resultants siguin reals.

Cal destacar que no és necessari que el nombre de clústers sigui igual al nombre de mostres de la classe minoritària, posat que s'estaria perdent molta informació pel camí.

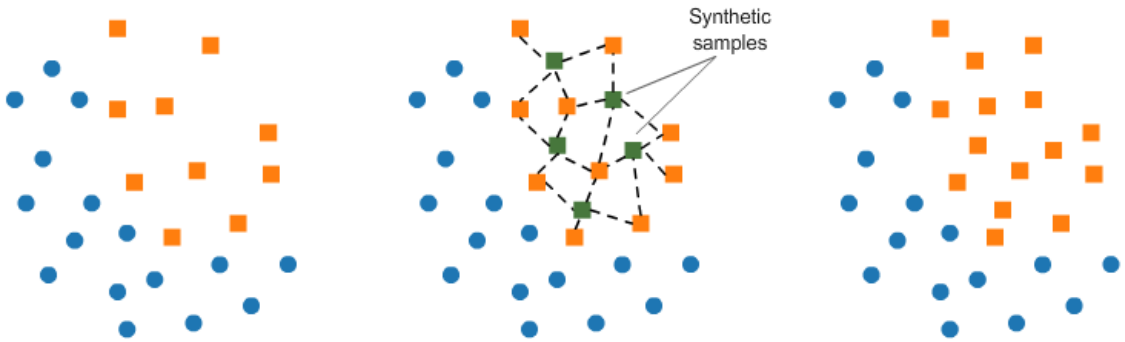
Figura 32: Cluster centroid, extreta de [**22**]



SMOTE SMOTE ve dels termes Minority over-sampling with replacement [7], i consisteix a crear mostres sintètiques de la classe minoritària al mig del camí

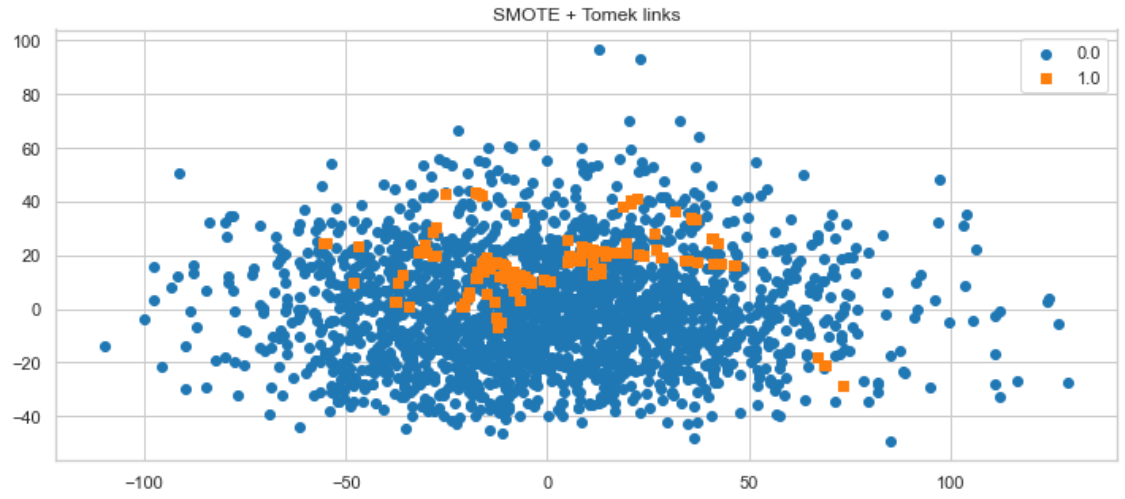
entre els veïns més pròxims. S'agafa la diferència del vector de característiques dels dos veïns, es multiplica per un valor entre 0 i 1 i se suma al valor que s'estigui considerant. D'aquesta manera s'aconsegueix ampliar l'espai de decisió de la classe minoritària. Com major sigui la ràtio entre les dues classes més veïns es consideraran a l'algoritme de cerca.

Figura 33: Smote, extreta de *kaggle*



Oversampling i Undersampling Per últim, existeixen mètodes que combinen SMOTE amb undersampling, **figura 34**. La lògica darrere això diu que, si es redueix el nombre d'instàncies de la classe majoritària i després es fa un augment de la classe minoritària, el biaix de l'algoritme pot canviar, ja que inevitablement els dos processos afavoreixen a classe desequilibrada.

Figura 34: Resultat d'aplicar Smote i Tomek al dataset d'exemple



5.4 Detalls d'implementació

En aquesta secció es definirà el programari que s'ha utilitzat per desenvolupar la part pràctica d'aquest treball. La totalitat d'aquesta part ha estat realitzada en

Python 3.8 en entorns de jupyter notebook. A la **taula 10** s'expliquen de manera detallada les llibreries de python utilitzades.

Taula 10: Programari utilitzat

Lliberia	**Versió**	**S'ha utilitzat per**
Numpy[14]	1.19.2	Algebra i tractament de llistes.
Pandas[28]	1.2.4	Tractament de dades i visualitzacions dels datasets.
Matplotlib[15]	3.3.2	Visualitzacions de dades.
Seaborn[36]	0.11.0	Visualitzacions de dades (hi ha alguna gràfica que utilitza una versió anterior)
Plotly[16]	4.14.3	Visualitzacions de dades interactives.
Scikit-learn[29, 6]	0.24.1	Models de machine learning i clustering, metodes de calcul de métriques, cross-validation, grid-search, separació de dades, k-folds, i preprocessament de dades.
XGBoost[8]	1.4.2	Implementació optimitzada dels algoritmes de boosting.
Imblearn[21]	0.7.0	Tractament de dades no balancejades, mètodes de downsample i upsample, implementació de pipelines que integren el resampling de les dades, models que tracten les dades de manera equilibrada
Tensorflow.keras[1, 9]	2.4.0	Deep learning (creació de xarxes neuronals, métriques, callbacks)

D'altra banda, el codi ha estat executat entre dos ordinadors amb CPUs diferents:

- Processador Intel core i7-10750H, 6 nuclis, 12 fils.
- Processador AMD Ryzen 5 5600X, 6 nuclis, 12 fils.

El fet de realitzar un grid-search cross-validation per cada mètode necessita paral · lelisme per tal de realitzar els càlculs ràpidament. Això és directament proporcional al nombre de nuclis del processador, i en executar els experiments de la **secció 6.3**, s'ha notat falta de potència que s'ha traduït en intervals llargs de temps entre resultats. Aproximadament cada notebook tardava 30-36 minuts en executarse, per tant, al haver-hi 8, en fer una prova sencera es tardaven aproximadament entre 4 i 5 hores.

6 Experiments i resultats

6.1 Plantejament

S'han proposat diversos experiments per intentar trobar una millora a la funció de REGICOR mitjançant algoritmes de machine learning. L'objectiu consisteix a predir si un participant tindrà un episodi cardiovascular o no. A partir d'aquí s'han organitzat una sèrie de proves, a cada una s'afegeixen variables diferents per veure com afecten el comportament de la predicció:

1. El primer set de dades correspon a les 8 variables que utilitza la funció de REGICOR.
2. El segon són les 8 més les 3 trobades a literatura.
3. El següent les 8 més els fenotips
4. Un conjunt amb totes les dades mencionades.

D'altra banda, també s'han provat estratègies d'aprenentatge no supervisat per avaluar si aquests eren capaços de trobar relacions que els altres algoritmes no troben.

Els models que s'utilitzaran per fer les proves són els següents:

- K-means
- SVM
- Random Forest
- Regressió Logística
- Extreme Gradient Boosting
- Model d'ensemble híbrid de random forest i regressió logística
- Xarxa Neuronal Senzilla (**secció 4.6.3**)

6.1.1 Preprocessament de dades

A l'hora de dividir les variables i els predictors, es fan tres passos, explicats al **capítol 5.2**, tenint en compte els tipus de dades que hi ha a les columnes, categòriques o numèriques:

1. Pels outliers s'utilitza una Z-score amb sensibilitat de 95 i als valors fora d'aquesta normal se'ls hi posa el valor més pròxim que entri dintre de la distribució.
2. Per les dades categòriques es fa un One Hot Encoding.
3. Per les dades numèriques s'estandarditzen utilitzant un Standard Scaler.
4. Per a les dades desconegudes, s'aplica una regressió en totes aquelles que no són fenotips. En el cas que ho siguin, simplement s'eliminen les entrades,

posat que hi han molts valors nuls i fer regressió pot ser contraproduent i afectar negativament a les dades.

A més a més, per tal de fer l'algoritme més eficient és necessari fer un pas més en el preprocessat de dades. Després d'estandarditzar les dades, s'utilitzen mètodes d'undersample per disminuir la classe majoritària, que en aquest cas és el grup que no presenta episodi. S'ha valorat utilitzar SMOTE per crear dades sintètiques, però era molt possible que produís overfit, a més a més, al tenir un set de dades en el que els episodis es troben molt separats entre ells, **figura 29** l'estratègia perd eficàcia.

S'utilitzen dues estratègies, TOMEK links, per tal de fer més fàcil al model diferenciar entre participants amb episodi i sense, i Cluster Centroids, per tal de no perdre informació al reduir dades. S'aplica un downsample de manera que la classe minoritària suposi un 0.05% del total. Aquests mètodes estan integrats dintre del pipeline del model, ja que es recomana que només es facin les transformacions a les dades de train. Com aquest pas es fa de manera interna, s'utilitza una funció de la llibreria imblearn que executa els passos de resample de les dades només quan es fa el fit de la pipeline.

6.1.2 Estratègia de Cross-Validation

Com no existeix un set de dades per validar els resultats obtinguts, es realitza un cross-validation amb tota la Base de dades.

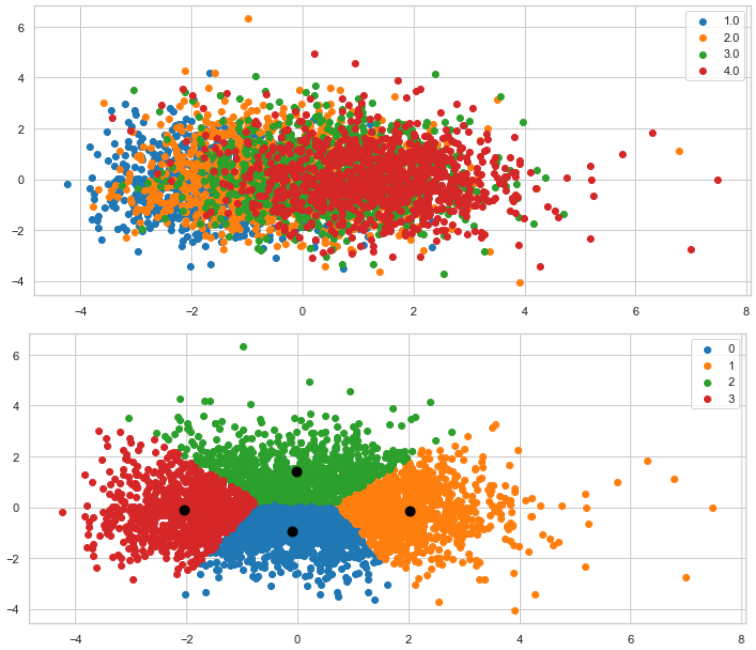
La partició de les dades es fa utilitzant un stratified K-Fold, que bàsicament divideix els sets a la CV de manera que a tots els conjunts de test hi hagi la mateixa distribució de la classe a predir. Per validar els resultats, s'han utilitzat 5 folds a les dades que entrarien a la funció de REGICOR i de 3 al conjunt total. Això és causa del fet que el cost computacional d'un pas a l'altre augmenta i la màquina on s'han fet les proves té limitacions.

6.2 Cerca no Supervisada

Els primers resultats que s'explicaran són les comparatives entre clústers extrets del K-means i els valors d'estratificació de la funció de REGICOR. En aquest experiment s'utilitzen les dades seguint els barems de cribratge de la funció, **secció 2.4**, però pot ser que afegint més variables i dades la distribució pugui canviar.

A la primera gràfica de la **figura 35** s'observa la distribució de les dades d'estratificació, mentre que a la segona gràfica el resultat de l'algoritme K-means.

Figura 35: Distribució dels grups d'estratificació originals a dalt i clústers trobats pel K-means a baix

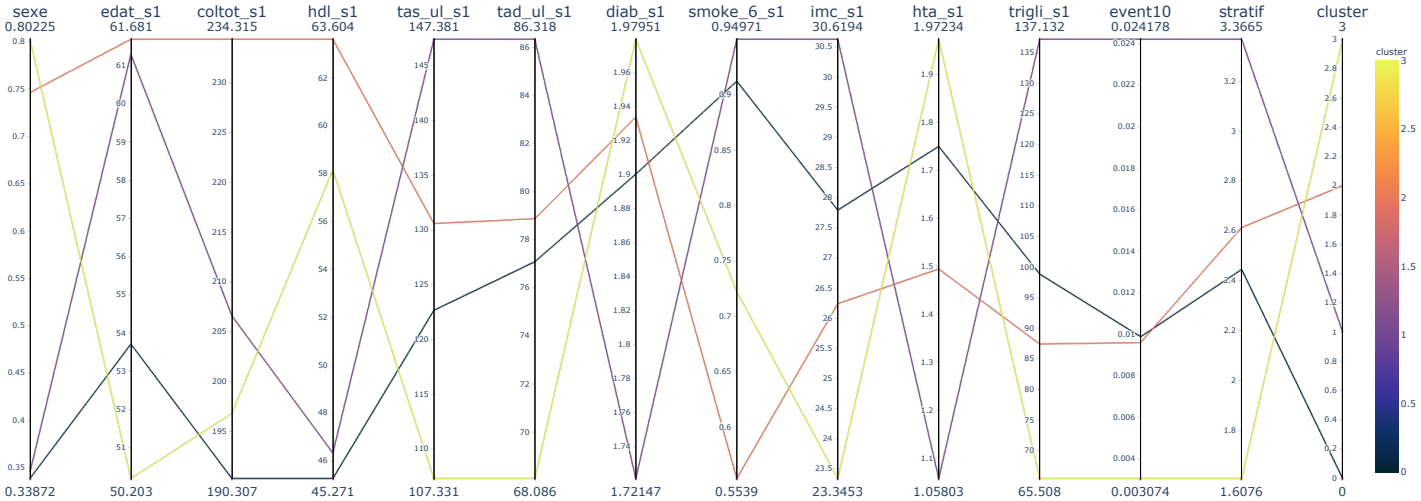


A la **figura 36** es pot comparar quins valors mitjans tenen els diferents clústers trobats. A l'última columna es pot observar el número de clúster i a la penúltima quin valor mitjà d'estratificació tenen, és a dir, dels participants que s'han classificat a cada clúster, a quin grup de risc pertanyien segons la funció de REGICOR.

Seguint la línia del clúster número 1, es pot apreciar que té la mitja d'estratificació més alta, a més a més, de tenir els valors més elevats en els factors de risc. Per tant, es podria dir que aquest clúster coincideix amb el grup de "risc alt". Per un altre costat, el clúster 3, té el comportament contrari i per tant correspondria al grup de "risc baix".

Per últim, els grups de risc mitjos són més difícils d'identificar, ja que no tenen diferències tan clares a nivell d'estratificació. S'observa disparitat a les variables, siguin de risc o de protecció, per exemple, el clúster 2 té els valors d'edat i colesterol més alts, però el mateix passa amb el sexe i el colesterol HDL.

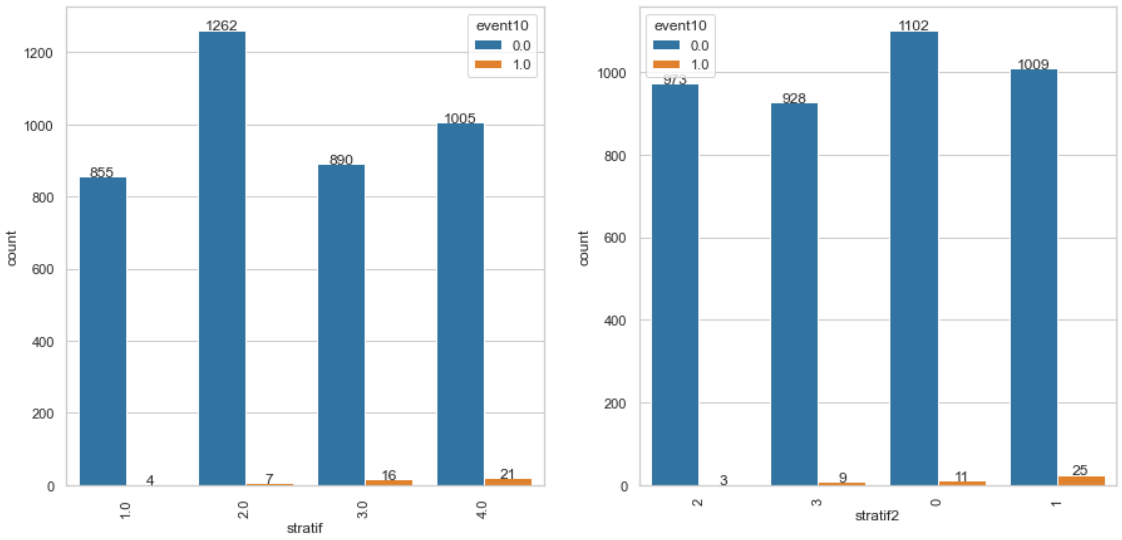
Figura 36: Mitges de les variables per cada clúster



A la **figura 37** es troba una comparativa entre la distribució dels episodis cardiovasculars segons la funció de REGICOR i segons els clústers. Tal com s'ha extret de la **figura 36**, el clúster 1 de major risc, té la majoria dels casos, però el clúster 3, tot i tenir els valors de risc més baixos, no és el que té menys incidència d'episodis, sinó que és el 2.

En clustering els grups de risc del mig acumulen 20 episodis, un 41% d'aquests, mentre que els de REGICOR 23, un 47% del total. Hi ha una diferència del 6% que no és suficient per dir que el clustering ha millorat la funció de REGICOR.

Figura 37: Comparativa de distribucions d'episodis, a l'esquerra funció de REGICOR i a la dreta clústers



6.3 Predicció d'episodi cardiovascular

Per les proves de classificació binària s'han utilitzat 3 mètriques que serviran per avaluar els resultats dels diferents models: eficàcia, sensibilitat i especificitat. Es valorarà positivament un equilibri entre especificitat i sensibilitat, posat que voldrà dir que l'algoritme és capaç de diferenciar entre la classe positiva i la classe negativa, és a dir, entre episodi cardiovascular o no.

Per tal de trobar els millors paràmetres per cada model s'ha tingut en compte la sensibilitat, de manera que el millor conjunt serà aquell que aconsegueixi un número més elevat de classe positiva predita correctament. Tot i que això pot causar un major nombre de falsos positius, s'ha cregut que és un error tolerable, posat que l'objectiu principal és millorar el diagnòstic d'episodis cardiovasculars que poden arribar a ser mortals. Tots els models utilitzen la versió adaptada per dades no balancejades, que calculen el pes de cada classe de manera interna com s'explica a la **secció 5.3.1**.

S'han fet dos estudis, un on s'utilitzen només les dades que entren dintre de les restriccions de la funció de REGICOR, i un altre per tal d'afegir el conjunt de dades no classificat per la funció, després de fer imputació dels valors desconeguts, per veure com es comporta davant els episodis que provenen de població més envellida.

6.3.1 Dades dintre de les restriccions de la funció REGICOR

Primer es presenten els resultats de les dades que es troben dintre les restriccions de la funció de REGICOR, **taula 11**. En negreta hi han els millors resultats per cada mètrica, mentre que en vermell es remarquen els pitjors resultats. Destaca que en general, les 8 variables que s'utilitzen a la funció de REGICOR més els 3 factors de risc afegits aconsegueixen el millors resultats. Concretament els millor model és el SVM, amb un 0.71 d'eficàcia equilibrada, un 0.82 de sensibilitat i un 0.61 d'especificitat.

D'altra banda, les imatges sembla que desplomin els resultats dels models de predicció, empitjorant molt els resultats en tots els casos, es discuteix el perquè a la **secció 6.4**

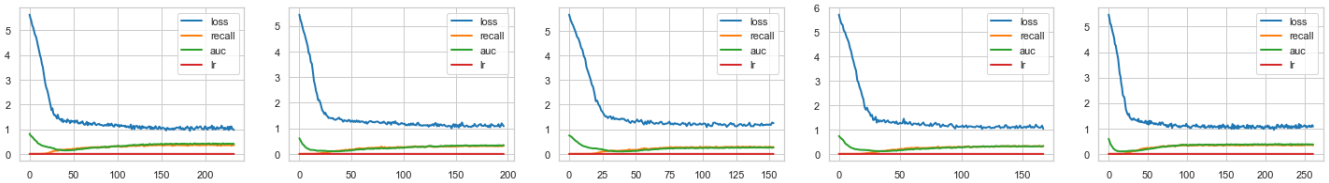
També, s'exposen les gràfiques de cada fold d'entrenament del cross-validation de les xarxes neuronals a la **figura 38**.

Taula 11: Resultats de la CV amb dades incloses als límits de la funció de Regicor

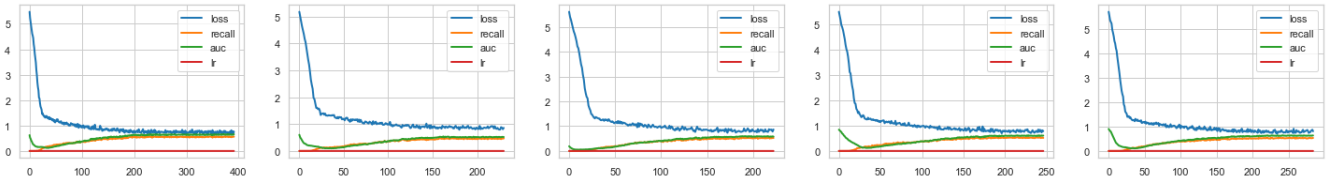
Model	Variables	Eficàcia Equilibrada	Sensibilitat	Especificitat
Support Vector Machine	Original	0.67	0.77	0.57
	Original + HTA, IMC, Trigli	**0.71**	**0.82**	**0.61**
	Original + fenotips imatge	**0.55**	**0.65**	**0.44**
	Original + HTA, IMC, Trigli + fenotips imatge	0.59	0.70	0.48
Random Forest	Original	0.63	0.60	0.62
	Original + HTA, IMC, Trigli	**0.65**	**0.65**	**0.66**
	Original + fenotips imatge	**0.48**	**0.59**	**0.36**
	Original + HTA, IMC, Trigli + fenotips imatge	0.52	0.62	0.41
Logistic Regression	Original	0.62	0.56	0.67
	Original + HTA, IMC, Trigli	**0.67**	**0.65**	**0.69**
	Original + fenotips imatge	**0.54**	**0.46**	0.63
	Original + HTA, IMC, Trigli + fenotips imatge	0.58	0.54	**0.62**
XGBoost	Original	0.55	0.60	0.50
	Original + HTA, IMC, Trigli	**0.57**	0.79	0.35
	Original + fenotips imatge	0.56	**0.51**	**0.61**
	Original + HTA, IMC, Trigli + fenotips imatge	**0.52**	**0.84**	**0.20**
HRFLM	Original	0.63	0.62	0.63
	Original + HTA, IMC, Trigli	**0.65**	**0.62**	**0.68**
	Original + fenotips imatge	**0.56**	0.54	**0.57**
	Original + HTA, IMC, Trigli + fenotips imatge	0.59	**0.51**	0.67
Xarxa Neuronal Simple	Original	0.50	0.60	**0.40**
	Original + HTA, IMC, Trigli	**0.57**	**0.62**	0.51
	Original + fenotips imatge	0.51	0.30	0.72
	Original + HTA, IMC, Trigli + fenotips imatge	**0.48**	**0.22**	**0.74**

Figura 38: Gràfiques d'entrenament de cada fold de la NN en dades restringides

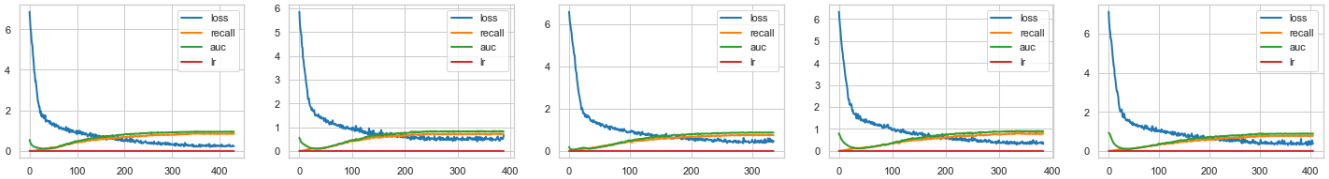
(a) Experiment amb només les dades originals



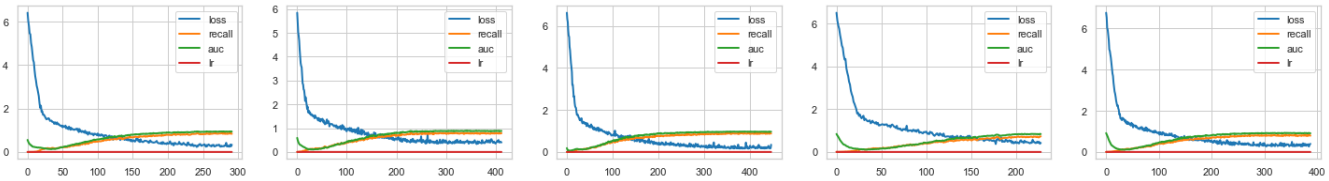
(b) Experiment amb les dades originals i els 3 biomarcadors nous



(c) Experiment amb les dades originals i fenotips



(d) Experiment amb les dades originals, 3 biomarcadors nous i fenotips

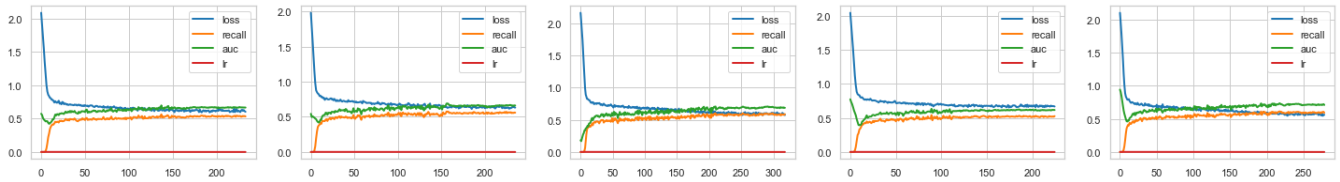


Taula 12: Resultats de la CV amb totes les dades

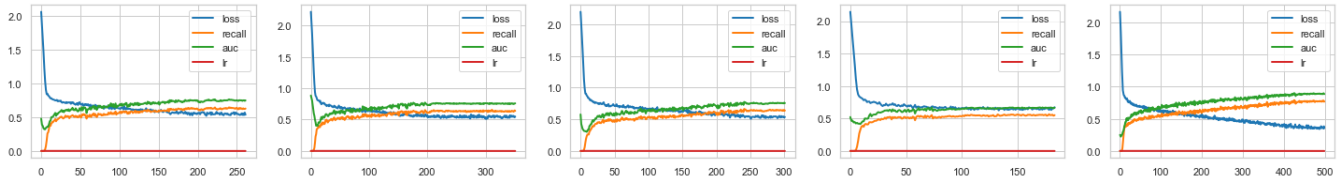
Model	Variables	Eficàcia Equilibrada	Sensibilitat	Especificitat
Support Vector Machine	Original	0.70	0.75	**0.65**
	Original + HTA, IMC, Trigli	**0.71**	**0.77**	0.66
	Original + fenotips imatge	0.70	0.70	0.69
	Original + HTA, IMC, Trigli + fenotips imatge	**0.69**	**0.68**	**0.70**
Random Forest	Original	0.68	0.69	**0.67**
	Original + HTA, IMC, Trigli	**0.70**	**0.72**	**0.68**
	Original + fenotips imatge	**0.66**	**0.64**	**0.68**
	Original + HTA, IMC, Trigli + fenotips imatge	0.68	0.69	**0.67**
Logistic Regression	Original	**0.71**	**0.72**	**0.70**
	Original + HTA, IMC, Trigli	**0.71**	**0.72**	0.71
	Original + fenotips imatge	0.67	0.62	0.72
	Original + HTA, IMC, Trigli + fenotips imatge	**0.66**	**0.59**	**0.74**
XGBoost	Original	**0.62**	0.52	0.72
	Original + HTA, IMC, Trigli	**0.62**	**0.54**	**0.70**
	Original + fenotips imatge	**0.61**	**0.41**	0.80
	Original + HTA, IMC, Trigli + fenotips imatge	**0.62**	0.43	**0.81**
HRFLM	Original	**0.69**	**0.70**	**0.68**
	Original + HTA, IMC, Trigli	**0.69**	**0.70**	**0.68**
	Original + fenotips imatge	0.68	0.66	0.71
	Original + HTA, IMC, Trigli + fenotips imatge	**0.67**	**0.61**	**0.72**
Xarxa Neuronal Simple	Original	**0.49**	**0.42**	**0.55**
	Original + HTA, IMC, Trigli	**0.49**	0.30	0.68
	Original + fenotips imatge	**0.49**	0.28	0.69
	Original + HTA, IMC, Trigli + fenotips imatge	**0.48**	**0.25**	**0.70**

Figura 39: Gràfiques d'entrenament de cada fold de la NN en dades no restringides

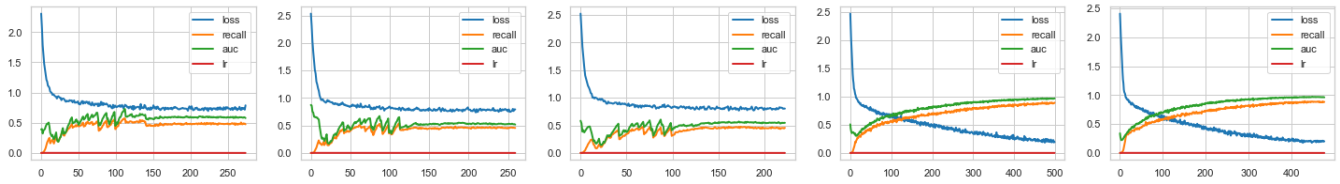
(a) Experiment amb només les dades originals



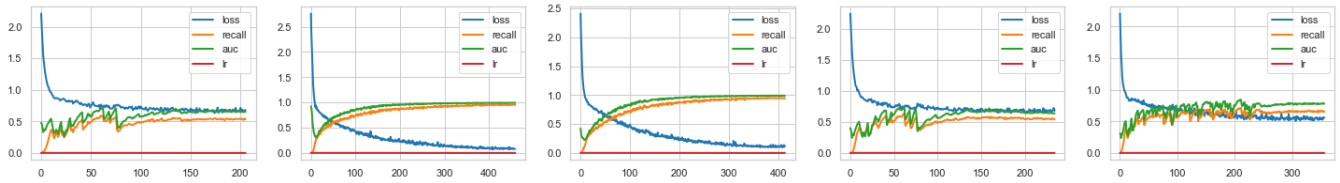
(b) Experiment amb les dades originals i els 3 biomarcadors nous



(c) Experiment amb les dades originals i fenotips



(d) Experiment amb les dades originals, 3 biomarcadors nous i fenotips



6.3.2 Dades fora de les restriccions de la funció de REGICOR

A continuació es presenten els resultats dels models al aplicar totes les dades, **taula 12**. Novament en negreta els millors resultats per cada mètrica, i en vermell els pitjors resultats. En aquest cas, es troba una millora considerable de les mètriques respecte l'experiment anterior, novament, en el cas de les 8 originals , més els 3 factors nous, el SVM torna a ser el millor model, amb un 0.71 d'eficàcia, 0.77 de sensibilitat i 0.66 d'especificitat, també destaca la regressió logística, que obté un 0.71 en eficàcia, 0.72 en sensibilitat i 0.71 d'especificitat en el mateix cas. D'aquest últim destaca l'especificitat mínima, que es de 0.70.

Tot i que en aquest cas les imatges també empitjoren els resultats, sembla que el fet de tenir més dades fa que els resultats es mantinguin més estables i que l'efecte negatiu de les noves dades sigui més suau.

Novament, s'exposen les gràfiques de cada fold d'entrenament del cross-validation de les xarxes neuronals a la **figura 39**.

6.4 Interpretació dels resultats

En general els resultats no aconsegueixen predir d'una manera precisa l'episodi, i una sensibilitat elevada va acompanyada d'una especificitat molt baixa.

Possiblement això sigui degut al fet que les classes, a més a més d'estar desequilibrades, tenen poques dissimilituds entre elles. És a dir, un voluntari amb episodi i un sense poden tenir característiques semblants, tot i que això s'intenta millorar aplicant TOMEK, no sembla ser suficient.

En estar les dades tan acoblades i utilitzar les versions dels models que afavoreixen les classes desequilibrades, aquests tendeixen a agafar els límits on més classes positives es troben sense tenir en compte els falsos positius. No obstant això, prenent de referència la distribució dels voluntaris amb episodis de la gràfica esquerra de la **figura 37**, es pot veure com a la classe de risc més alt hi ha aproximadament 1000 persones que no pateixen episodi.

Això no vol dir que la funció de REGICOR s'hagi equivocat fent la classificació, sinó que ha estimat que aquestes persones tenen més probabilitats que les demés de patir un episodi. Un dels punts a millorar d'aquesta funció, és evitar que la majoria dels episodis cardiovasculars caiguin a grups d'estratificació mitjana, per tant, sempre que el model entrenat aconsegueixi una sensibilitat superior al 0.50 i una especificitat raonable, millorarà la funció de REGICOR. Això és degut al fet que el model estarà predient de manera correcta la possibilitat d'infart de més de la meitat dels casos i en una estratificació s'afegiran al grup de risc més alt.

De fet, el millor dels resultats, el SVM amb un 0.82 de sensibilitat, **taula 11**, es tradueix en el fet que un 0.82 dels episodis han estat ben estimats pel model de predicció, i podrien ser afegits al grup de "risc alt"en una estratificació. D'aquesta manera, només hi hauria repartits en els grups de risc menor un 0.18 dels episodis. En el cas de REGICOR, prenent la **figura 37 esquerra**, només hi ha un 0.44 dels episodis al grup d'estratificació alt i el 0.66 restant es troba repartit entre les altres classes. Per tant, el model d'aprenentatge automàtic sobrepassa la funció de REGICOR en aquest aspecte.

L'altra cosa que s'ha de tenir en compte és l'especificitat, en el cas del SVM, és de 0.61, cosa que vol dir que hi ha un 0.39 de falsos positius. A l'estratificació de la funció REGICOR, el total de voluntaris que no presenten episodi i es troben al grup de risc alt, és d'un 0.25, per tant s'estan classificant "pitjor" un 0.13 més de voluntaris. Tanmateix, és molt més favorable millorar la predicció d'un episodi cardiovascular que pot arribar a ser mortal, que reduir els falsos positius.

S'ha de tenir en compte, però, que aquest valor ha d'estar controlat, per exemple, el XGboost, **taula 11**, té un resultat en el qual la sensibilitat és de 0.84, però l'especificitat és de 0.20. Això vol dir que està classificant erròniament un 0.80 dels participants, i per tant no és un resultat òptim.

Una altra conclusió que es pot treure, és que les dades extretes d'imatges no ajuden al model a predir millor els episodis. Això pot ser degut al fet que s'estan afegint moltes dades que poden confondre al model. És possible que algunes d'aquestes

siguin soroll i s'hagi de descartar alguna. No obstant això, les proves realitzades representen una primera aproximació per avaluar el rendiment de les imatges i es poden estudiar altres maneres d'incorporar-les a la predicció.

Per últim, i per acabar aquesta secció, els models utilitzats han estat capaços de millorar el seu rendiment afegint les dades d'aquells voluntaris que quedaven fora dels llindars de la funció de risc original. Per tant, es pot assumir que els models d'aprenentatge automàtic poden incloure els participants de més avançada edat, obrint portes a utilitzar les dades cerebrovasculars que es mencionaven a la **secció 2.4** i millorar la diagnosi d'aquestes malalties.

7 Conclusions i treball futur

7.1 Conclusions

En aquest treball s'ha fet un estudi de diversos models d'aprenentatge automàtic per intentar trobar una alternativa millor a una funció de risc mèdica mitjançant dades clíniques. Una vegada acabat aquest estudi es conclou que:

- Existeix molta literatura que tracta el tema de la predicció de risc cardiovascular. Molta d'aquesta se centra en imatges i extracció de característiques d'aquestes. S'ha aconseguit utilitzar les eines de recerca científica per a trobar articles relacionats amb el tema a tractar, i així extreure informació valuosa sobre la contextualització i abordament del problema per aplicar-ho al treball.
- Davant la base de dades, s'han estudiat de manera eficient els diversos biomarcadors escollits, associant i trobant relacions entre aquests que han afavorit la comprensió del problema. També es creu que mitjançant les representacions visuals presentades, s'ha fet un bon anàlisi a primera vista de la situació de les dades. No obstant això, és possible que una exploració més profunda dels fenotips de les imatges podria haver ajudat a reduir el soroll d'aquestes dades i per tant millorar la predicció d'episodis.
- S'ha fet un estudi dels models utilitzats, permetent així una major comprensió del seu funcionament que ha ajudat a definir les proves realitzades. Tanmateix, envers l'ús de les xarxes neuronals, es creu que es podria haver fet una millor exploració d'estructures per tal d'aconseguir un millor resultat.
- També s'ha aprofundit en el coneixement sobre les dades no equilibrades, fent exploracions sobre els mètodes que s'utilitzen a la bibliografia tècnica i extraient metodologies útils pel desenvolupament del treball. També, s'han hagut de definir les mètriques per avaluar les dades, però en un problema tan poc equilibrat, és complicat trobar una validació absoluta.
- Envers els experiments, han existit limitacions tècniques, **secció 5.4**, que poden haver afectat de manera negativa als resultats. No obstant això, s'ha aconseguit definir un conjunt d'experiments en els quals s'han pogut fer comparacions entre models totalment diferents. S'han extret resultats vàlids, que demostren que els mètodes d'aprenentatge automàtic tenen un millor rendiment que els mètodes clàssics utilitzats a medicina, a més a més, aquests models predictius tenen espai de millora, tal com es comenta a la **secció 7.2**.

7.2 Treball futur

En aquesta secció es discutiran diversos canvis que es poden realitzar als experiments per millorar els resultats obtinguts.

Primer de tot, Tal com s'ha avançat a la **secció 6.2** es poden seguir fent diverses proves amb l'algorisme de classificació no supervisada. Per cada dimensió que s'afegeixi, la forma de les dades canviarà i l'agrupament per clústers també, descobrint noves relacions entre dades.

D'altra banda, pel que fa als models de predicció d'episodi cardiovascular, es poden adoptar diferents estratègies per intentar millorar els seus resultats. En aquest cas s'ha utilitzat una xarxa neuronal bàsica i amb molt pocs paràmetres, estudiar-ne una de més profunda podria extreure resultats millors.

Una altra prova que es podria realitzar és extreure les probabilitats de patir episodi en comptes de retornar directament la predicció. Amb aquestes, es podria realitzar una nova estratificació, fer una comparació directa amb REGICOR i avaluar on cauen els casos de risc. Tots aquells que l'algoritme doni possibilitats de patir episodi es classificarien directament al grup de "risc alt" i després caldria aprendre on es posa el límit entre les mostres restants per classificar entre risc baix i mitjà.

També, s'hauria de fer un feature selection entre les dades d'imatge per tal d'intentar reduir el soroll d'aquestes. A més a més, es podrien afegir més biomarcadors seguint aquesta estratègia, posat que en els resultats s'aprecia com, incorporant més dades mèdiques, els models milloren.

Per últim, en aquest treball s'han utilitzat les dades extretes de les imatges, però en un futur es podrien afegir les imatges en una fusió entre una xarxa neuronal convolucional i algun dels models de predicció estudiats. De manera que, en comptes d'afegir totes les dades que es poden extreure de les imatges, sigui la xarxa qui decideixi quina informació és la més important.

Referències

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Charu C. Aggarwal. *Data Mining*. Springer, 2015.

[3] Charu C. Aggarwal. *Neural Networks and Deep Learning*. Springer, 2018.

[4] Charlotte Andersson, Matthew Nayor, Connie W. Tsao, Daniel Levy, and Ramachandran S. Vasan. Framingham heart study: Jacc focus seminar, 1/8. *Journal of the American College of Cardiology*, 77(21):2680–2692, 2021.

[5] Małgorzata Bach, Aleksandra Werner, and Mateusz Palt. The proposal of undersampling method for learning from imbalanced datasets. *Procedia Computer Science*, 159:125–134, 2019. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.

[6] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[9] François Chollet et al. Keras. `https://keras.io`, 2015.

[10] Maria del Mar Vila, Beatriz Remeseiro, Maria Grau, Roberto Elosua, and Laura Igual. *Recent Advances on Automatic Carotid Artery Analysis in Ultrasound Images*.

[11] Maria del Mar Vila, Beatriz Remeseiro, Maria Grau, Roberto Elosua, Àngels Betriu, Elvira Fernandez-Giraldez, and Laura Igual. Semantic segmentation

with densenets for carotid artery ultrasound plaque segmentation and cimt estimation. *Artificial Intelligence in Medicine*, 103:101784, 2020.

[12] Ivo D. Dinov. *Data Science and Predictive Analytics*. Springer, 2018.

[13] Jiri Frohlich and Ahmad Al-Sarraf. Cardiovascular risk and atherosclerosis prevention. *Cardiovascular Pathology*, 22(1):16–18, 2013.

[14] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[15] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[16] Plotly Technologies Inc. Collaborative data science, 2015.

[17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[18] Stephanie M. Kochav, Yoshihiko Raita, Michael A. Fifer, Hiroo Takayama, Jonathan Ginns, Mathew S. Maurer, Muredach P. Reilly, Kohei Hasegawa, and Yuichi J. Shimada. Predicting the development of adverse cardiac events in patients with hypertrophic cardiomyopathy using machine learning. *International Journal of Cardiology*, 327:117–124, 2021.

[19] A. Kondababu, V. Siddhartha, BHK. Bhagath Kumar, and Bujjibabu Penumutchi. A comparative study on machine learning based heart disease prediction. *Materials Today: Proceedings*, 2021.

[20] Sonia Kunstmann and Int. Fernanda Gainza. Herramientas para la estimaciÓn del riesgo cardiovascular. *Revista Médica Clínica Las Condes*, 29(1):6–11, 2018. Tema central: Fronteras de la cardiología.

[21] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[22] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409-410:17–26, 2017.

[23] Jaume Marrugat, Joan Vila, José Miguel Baena-Díez, María Grau, Joan Sala, Rafel Ramos, Isaac Subirana, Montserrat Fitó, and Roberto Elosua. Validez relativa de la estimación del riesgo cardiovascular a 10 años en una cohorte poblacional del estudio regicor. *Revista Española de Cardiología*, 64(5):385–394, 2011.

[24] Rosa-María Menchón-Lara, José-Luis Sancho-Gómez, and Andrés Bueno-Crespo. Early-stage atherosclerosis detection using deep learning over carotid ultrasound images. *Applied Soft Computing*, 49:616 – 628, 2016.

[25] Carol Mitchell, Claudia E. Korcarz, Adam D. Gepner, Joel D. Kaufman, Wendy Post, Russell Tracy, Amanda J. Gassett, Nanxun Ma, Robyn L. McClelland, and James H. Stein. Ultrasound carotid plaque features, cardiovascular disease risk factors and events: The multi-ethnic study of atherosclerosis. *Atherosclerosis*, 276:195 – 202, 2018.

[26] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[27] Rine Nakanishi, Piotr J. Slomka, Richard Rios, Julian Betancur, Michael J. Blaha, Khurram Nasir, Michael D. Miedema, John A. Rumberger, Heidi Gransar, Leslee J. Shaw, Alan Rozanski, Matthew J. Budoff, and Daniel S. Berman. Machine learning adds to clinical and cac assessments in predicting 10-year chd and cvd deaths. *JACC: Cardiovascular Imaging*, 14(3):615–625, 2021.

[28] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] Massimo Piepoli, Arno Hoes, Stefan Agewall, C. Albus, Carlos Brotons, Alberico Catapano, Marie-Therese Cooney, Ugo Corrà, Bernard Cosyns, Christi Deaton, Ian Graham, Michael Hall, F.D. Hobbs, Maja-Lisa Løchen, Herbert Loellgen, Pedro Marques-Vidal, Joep Perk, Eva Prescott, Josep Redon, and Slavomira Filipova. 2016 european guidelines on cardiovascular disease prevention in clinical practice: The sixth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the european association for cardiovascular prevention & rehabilitation (eacpr). *European Heart Journal*, 37:ehw106, 05 2016.

[31] Kenichi Sakakura, Masataka Nakano, Fumiyuki Otsuka, Elena Ladich, Frank D. Kolodgie, and Renu Virmani. Pathophysiology of atherosclerosis plaque progression. *Heart, Lung and Circulation*, 22(6):399–411, 2013.

[32] Santi Seguí. Apunts de l'assignatura aprenentatge automàtic, Setembre-Gener 2020-2021.

[33] Balaji K. Tamarappoo, Andrew Lin, Frederic Commandeur, Priscilla A. McElhinney, Sebastien Cadet, Markus Goeller, Aryabod Razipour, Xi Chen, Heidi Gransar, Stephanie Cantu, Robert JH. Miller, Stephan Achenbach, John Fried-

man, Sean Hayes, Louise Thomson, Nathan D. Wong, Alan Rozanski, Piotr J. Slomka, Daniel S. Berman, and Damini Dey. Machine learning integration of circulating and imaging biomarkers for explainable patient-specific prediction of cardiac events: A prospective study. *Atherosclerosis*, 318:76–82, 2021.

[34] Gerrit L. ten Kate, Stijn C.H. van den Oord, Eric J.G. Sijbrands, Aad van der Lugt, Nico de Jong, Johan G. Bosch, Antonius F.W. van der Steen, and Arend F.L. Schinkel. Current status and future developments of contrast-enhanced ultrasound of carotid atherosclerosis. *Journal of Vascular Surgery*, 57(2):539 – 546, 2013.

[35] Bin Wang, Zhiyun Zhao, Shanshan Liu, Shuangyuan Wang, Yuhong Chen, Yu Xu, Min Xu, Weiqing Wang, Guang Ning, Mian Li, Tiange Wang, and Yufang Bi. Impact of diabetes on subclinical atherosclerosis and major cardiovascular events in individuals with and without non-alcoholic fatty liver disease. *Diabetes Research and Clinical Practice*, 177:108873, 2021.

[36] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[37] Chenhui Xiao, Zhenzhou Li, Jianfeng Lu, Jinyan Wang, Haoteng Zheng, Zuyue Bi, Mengyang Chen, Rui Mao, and Minhua Lu. A new deep learning method for displacement tracking from ultrasound rf signals of vascular walls. *Computerized Medical Imaging and Graphics*, 87:101819, 2021.

[38] Salim Yusuf, Steven Hawken, Stephanie Ounpuu, Tony Dans, Alvaro Avezum, Fernando Lanas, Matthew McQueen, Andrzej Budaj, Prem Pais, John Varigos, and Liu Lisheng. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study. *The Lancet*, 364(9438):937–52, 2004.

[39] Min Zeng, Beiji Zou, Faran Wei, Xiyao Liu, and Lei Wang. Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pages 225–228, 2016.