UAB
Universitat Autònoma de Barcelona

Aplicacions d'aprenentatge automàtic per a la prevenció i predicció de les malalties cardiovasculars

Investigació operativa

Autors: *Toni Esteve Gené*
Gloria Tarragona Ruiz
Ainhoa Trillo Rodríguez

Data de l'entrega: 15 de desembre de 2024

Índex

1 Introducció

1.1 Aprenentatge autòmatic

L'aprenentatge automàtic és una branca de la intel · ligència artificial que se centra en el desenvolupament d'algorismes que permeten als ordinadors aprendre patrons a partir de dades i elaborar prediccions o prendre decisions de manera autònoma.

Els models d'aprenentatge automàtic es poden classificar en 3 categories:

- **Aprenentatge supervisat:** Els models d'aprenentatge supervisat tenen com a objectiu generar prediccions a partir de dades etiquetades. És a dir el model treballa amb unes dades de les quals coneixem quina és la resposta que ha de retornar.
- **Aprenentatge no supervisat:** Els models d'aprenentatge no supervisat treballen amb dades no etiquetades. Aquest tipus de models tenen com a objectiu trobar patrons inherents o correlacions en les dades en lloc de predir etiquetes específiques.
- **Aprenentatge per reforç:** Els models d'aprenentatge per reforç tenen com a objectiu aprenendre a prendre decisions interactuant amb l'entorn.

Donada una visió global sobre que és l'aprenentatge automàtic i com és podem classificar els models segons el tipus d'aprenentatge és important remarcar que en aquest treball ens centrarem en l'estudi de problemes d'aprenentatge supervisat. En la següent secció veiem com s'estructuren i quins objectius es plantegen en aquest tipus de problemes.

1.2 Problemes d'aprenentatge supervisat

En un problema d'aprenentatge supervisat tenim conjunts de dades que venen donats com la col · lecció de parelles $\{(x_i, y_i)\}_{i \in I}$ on $x_i \in \mathcal{X}$ i $y_i \in \mathcal{Y}$. Les variables x_i s'anomenen *variables explicatives* i les variables y_i s'anomenen *variables objectiu*.

Aleshores suposem que existeix una funció $f : \mathcal{X} \to \mathcal{Y}$ tal que

$$y_i = f(x_i) + \epsilon$$

on ϵ és el soroll gaussià de mitjana zero.

L'objectiu donat aquest problema és trobar la funció $\hat{f}$ que millor aproxima f, és a dir $\hat{f} : \mathcal{X} \to \mathcal{Y}$ tal que $\hat{f}(x_i) \approx y_i$

En funció del tipus de la dada de sortida tenim dos tipus de tasques:

- **Tasca de classificació:** Tenim aquesta tasca si l'espai de sortida és discret, és a dir la variable de sortida pren valors categòrics
- **Tasca de regressió:** Tenim aquesta tasca si l'espai de sortida és contínu, és a dir la variable de sortida pren valors númerics a $\mathbb{R}$.

1.3 Aprenentatge automàtic en la medicina

L'aprenentatge automàtic és una disciplina que s'aplica a una gran varietat d'àrees, en particular a la sanitat. La introducció d'eines d'aprenentatge automàtic ha revolucionat el sector sanitari permetent als professionals oferir una millor atenció als pacients. Algunes d'aquestes aplicacions són la detecció primerenca i prevenció de malalties, l'elaboració de diagnòstics personalitzats i el suport en la presa de decisions mèdiques.

A causa de la gran importància que tenen les aplicacions de l'aprenentatge automàtic en el camp de la medicina, en aquest projecte estudiarem informació clínica rellevant per l'avaluació de malalties cardiovasculars. Segons l'organització mundial de la Salut (OMS) aquest tipus d'afeccions són una de les principals causes de mort en el món, és per això que essencial prendre mesures preventives efectives i personalitzades.

A les següents seccions del treball durem a terme una anàlisi sobre un conjunt de dades clíniques usant diferents algorismes d'aprenentatge supervisat amb l'objectiu de prevenir i predir les malalties cardiovasculars. Per tal d'assolir aquest propòsit plantegem els següents objectius específics:

- Identificar quins factors de risc són més importants per predir malalties cardiovasculars
- Predir la presència o absència d'aquestes malalties
- Analitzar els resultats obtinguts per entendre com es podrien implementar aquests algorismes a entorns clínics reals

2 Descripció del conjunt de dades

2.1 Origen del conjunt de dades

El conjunt de dades utilitzat en aquest treball està disponible a Kaggle, una plataforma reconeguda de datasets i competicions d'anàlisi de dades. Es pot accedir directament mitjançant l'enllaç següent: Cardiovascular Disease Dataset, publicat per l'usuari *Cole Welkins*.

Segons la informació proporcionada a la pàgina de Kaggle, el nostre conjunt de dades té el seu origen en el *UCI Machine Learning Repository*, una font àmpliament utilitzada en la recerca científica i el desenvolupament de models d'aprenentatge automàtic, creada per investigadors de la Universitat de Califòrnia a Irvine. Aquest conjunt de dades recopila mètriques clau de pacients reals amb l'objectiu principal de predir la presència o absència de malalties cardiovasculars. Es pot accedir al conjunt de dades a UCI Machine Learning Repository - Cardiovascular Disease. Així mateix, també incorpora informació d'un altre conjunt de dades utilitzat en estudis d'investigació i creat per la mateixa plataforma de Kaggle.

És important destacar que aquest conjunt de dades està disponible sota la llicència *Open Data Commons*, cosa que en permet l'ús i distribució amb finalitats d'anàlisi i investigació, complint amb les normatives associades a l'ús de dades obertes.

A més, inclou informació sobre la metodologia de recopilació i preparació de dades, que va incloure la consolidació de conjunts de dades de diverses fonts, retenint característiques clau amb potencial predictiu. Es va realitzar una neteja de dades gestionant valors inexistents mitjançant imputació o eliminació, i tractant valors atípics basant-se en coneixement expert.

2.2 Descripció de les variables

El conjunt de dades inclou les següents variables, classificades segons el seu tipus i funció en l'anàlisi:

- **ID**: Identificador únic per a cada pacient.

Variable objectiu

- **Cardio:** És una variable binària categòrica que indica la presència o absència de malaltia cardiovascular (0: Absència, 1: Presència).

Variables explicatives

- **Contínues**
  - **Age:** Edat del pacient en dies.
  - **Age_years:** Edat del pacient en anys (derivada de age).
  - **Height:** Alçada del pacient en centímetres.
  - **Weight**: Pes del pacient en quilograms.
  - **BMI**: Índex de Massa Corporal, derivat de weight i height (unitats de kg/m^2).
  - **Ap_hi**: Pressió arterial sistòlica (en mm Hg).
  - **Ap_lo**: Pressió arterial diastòlica (en mm Hg).
- **Categòriques**
  - **Gender:** Gènere del pacient (1: Dona, 2: Home).
  - **Cholesterol:** Nivells de colesterol (1: Normal, 2: Per sobre del normal, 3: Molt per sobre del normal).
  - **Gluc:** Nivells de glucosa (1: Normal, 2: Per sobre del normal, 3: Molt per sobre del normal).
  - **Smoke:** Estat de tabaquisme (0: No fumador, 1: Fumador).
  - **Alco:** Consum d'alcohol. (0: No consumeix, 1: Consumeix).

- **Active:** Activitat física. (0: No físicament actiu, 1: Físicament actiu).
- **Bp_category:** Categoria de pressió arterial basada en ap_hi i ap_lo. Categories inclouen "Normal", "Eleva", "Hipertensio Fase ", "Hipertensio Fase 2" i "Crisi Hipertensiva".
- **Bp_category_encoded:** Variable duplicada de la variable bp_category.

2.3 Rellevància del conjunt de dades pel projecte

El conjunt de dades és rellevant en la seva alineació directa amb l'objectiu d'identificar factors de risc associats a les malalties cardiovasculars.

Un aspecte destacat del conjunt de dades és el balanç entre les instàncies de pacients amb malalties cardiovasculars i aquells que no en tenen, cosa que garanteix que els models predictius desenvolupats puguin aprendre de manera equitativa sobre ambdós grups i reduir un possible *overfitting*.

A més, les característiques incloses en el conjunt de dades representen mètriques clau per a l'estudi de les malalties cardiovasculars, tal com s'indica en estudis científics i guies clíniques internacionals. Variables com la pressió arterial, els nivells de colesterol i glucosa, l'IMC i els hàbits de vida (fumar, consum d'alcohol i activitat física) són reconegudes com a factors de risc fonamentals per a la predicció i prevenció d'aquestes malalties. Això assegura que l'anàlisi es basi en dades clínicament rellevants i recolzades per l'evidència.

L'amplitud del conjunt de dades, amb uns 70,000 registres, garanteix una anàlisi representativa i estadísticament significativa per construir models robustos, identificar patrons rellevants i explorar possibles intervencions preventives.

3 Anàlisi exploratòria del conjunt de dades

A la següent secció durem a terme una anàlisi exploratòria de les dades que consistirà en tres parts: l'anàlisi descriptiva de les variables, la detecció de valors atípics i el càlcul de correlació entre les variables.

3.1 Anàlisi descriptiva

Per obtenir una visió general de les principals característiques del nostre conjunt de dades realitzem un resum estadístic.

	age	gender	height	weight	ap_hi	ap_lo	chol	gluc
count	68205	68205	68205	68205	68205	68205	68205	68205
mean	19462.67	1.35	164.37	74.10	126.43	81.26	1.36	1.23
std	2468.38	0.48	8.18	14.29	15.96	9.14	0.68	0.57
min	10798	1	55	11	90	60	1	1
25%	17656	1	159	65	120	80	1	1
50%	19700	1	165	72	120	80	1	1
75%	21323	2	170	82	140	90	1	1
max	23713	2	250	200	180	120	3	3

	smoke	alco	active	cardio	age_years	bmi	bp_category
count	68205	68205	68205	68205	68205	68205	68205
mean	0.09	0.05	0.80	0.49	52.82	27.51	1.91
std	0.28	0.22	0.40	0.50	6.77	6.03	0.91
min	0	0	0	0	29	3.47	0
25%	0	0	1	0	48	23.88	2
50%	0	0	1	0	53	26.35	2
75%	0	0	1	1	58	30.12	2
max	1	1	1	1	64	298.67	3

Taula 1: Resum estadístic del conjunt de dades

En primera instància, atès que l'alçada mínima registrada (55 cm) es troba per sota del rang considerat normal per a la població humana, concloem que aquestes dades probablement es deuen a errors de mesura o registre. Per

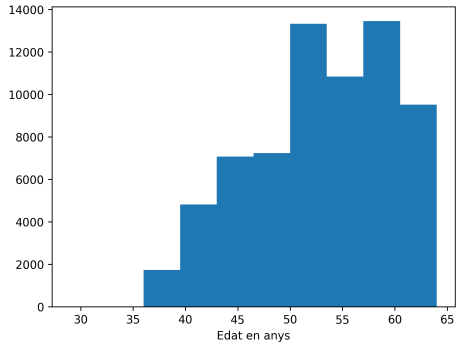
preservar la integritat de l'anàlisi i evitar biaixos als resultats, optem per excloure del conjunt de dades totes les observacions amb una alçada inferior a 140 cm.

D'altra banda, amb l'objectiu de minimitzar la multicol · linealitat i millorar la robustesa del nostre model, hem decidit eliminar les variables *altura* i *pes* del conjunt de dades, atès que el *bmi* ja captura la relació entre ambdues variables, evitant així correlacions que poguessin influir negativament en les prediccions. Addicionalment, s'han eliminat la variable *bp_category* i *bp_category_encoded*. Aquestes variables, categoritzen la pressió arterial en funció de les variables contínues *ap_hi* i *ap_lo* i, per tant, resultaven redundants comptar amb elles en el conjunt de dades. Així mateix, s'ha eliminat la variable *age* i ens hem quedat únicament amb la variable *age_years* per explicar l'edat dels pacients.

A continuació estudiarem descriptivament les variables numèriques i categòriques del nostre conjunt de dades. Durem a terme l'anàlisi de la distribució de cada característica mitjançant histogrames.

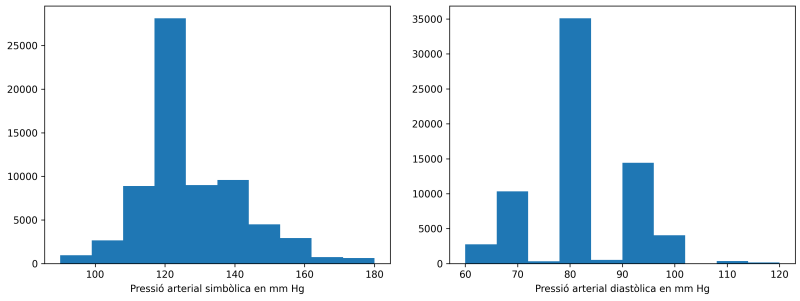
3.1.1 Variables numèriques

- **Edat:**

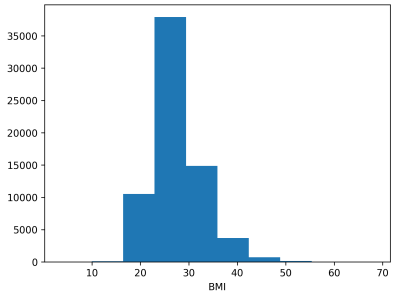


A través del següent histograma podem observar que la distribució de la característica edat està esbiaixada cap a la dreta. Aquest biaix és consistent amb el tema d'estudi, ja que les persones d'edat avançada són més propenses a patir malalties cardiovasculars.

- **Pressió arterial sistòlica i diastòlica**: En els següents histogrames trobem representades les distribucions de les dades de la pressió arterial sistòlica i la pressió arterial diastòlica. Notem que les dues distribucions són aproximadament simètriques, doncs no s'observen evidències de biaix. A més observem que les distribucions estan centrades al voltant dels valors 120 mmHg i 80 mmHg que són precisament els valors normals segons la *Sociedad Española de Hipertensión*. S'observa doncs que la distribució està centrada al voltant dels valors normals o elevats de pressió arterial sistòlica i diastòlica, cosa que és esperable en un estudi relacionat amb malalties cardiovasculars.



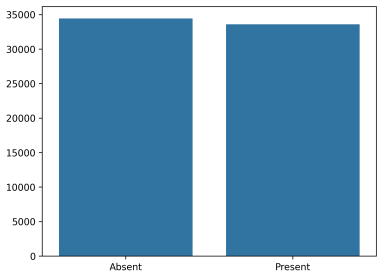
- **BMI:**



Primerament, cal notar que el BMI es classifica en tres categories: infrapès, normopès i sobrepès. Els valors menors que 18.5 es relacionen amb l'infrapès, els compresos entre 18.5 i 25 a normopès i els que són majors de 25 a sobrepès. Pel que fa a l'histograma, aquest mostra que la variable té una distribució aproximadament simètrica centrada al voltant del 25. Això indica que el conjunt de dades està format majoritariament per pacients que es troben en el límit entre un pes normal i el sobrepès.

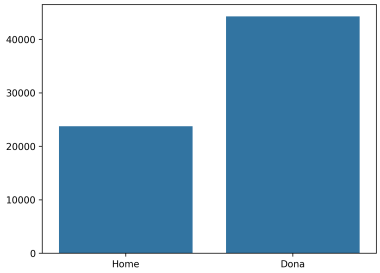
3.1.2 Variables categòriques

- **Cardio**



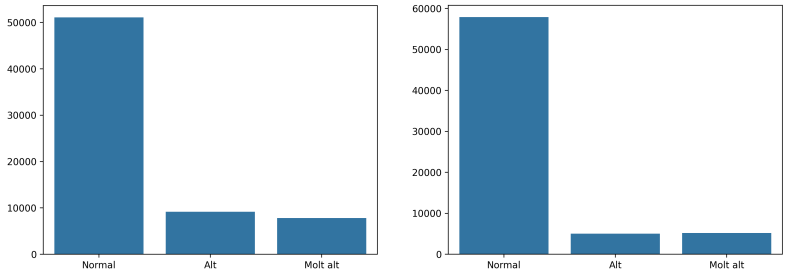
S'observa que la variable que mesura la condició de patir una malaltia cardiovascular presenta un equilibri adequat a la mostra, amb una distribució similar de casos *absent* i *present*. Aquest balanç és fonamental per garantir la validesa interna dels models predictius i minimitzar el risc de biaixos cap a una certa clase.

- **Sexe:**

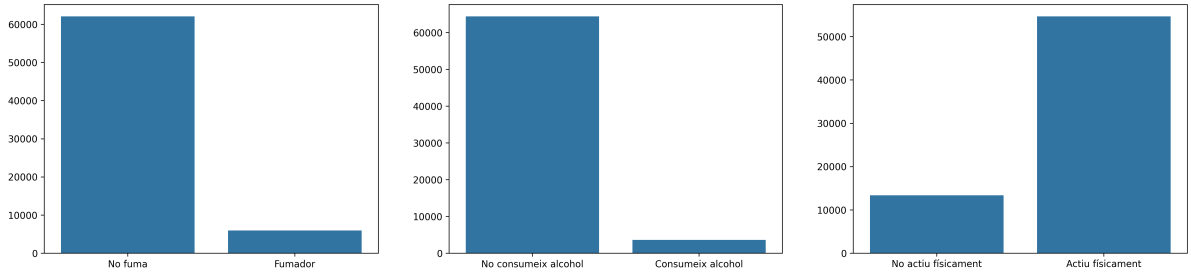


El gràfic de barres ens mostra que el nostre conjunt de dades té una proporció major de dones que d'homes. Aquest desequilibri pot introduir biaixos als resultats perquè els factors de risc poden estar distribuïts de manera diferent entre sexes. Aquesta asimetria en les proporcions pot resultar interessant, ja que històricament la gran majoria d'estudis s'han realitzat en homes.

- **Colesterol i glucosa:** Els següents gràfics indiquen que la major part dels pacients de la mostra tenen un nivell normal tant de colesterol com de glucosa. Per tant, tot i que la mostra per a pacients amb un nivell "alt" i "molt alt" és significativa, és possible que el model tingui biaix cap a aquest col · lectiu.



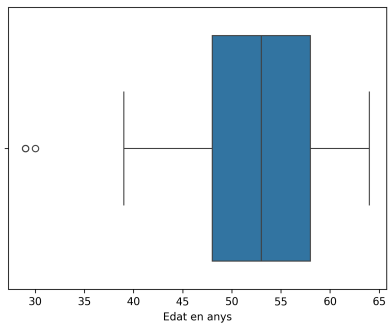
- **Fumar, alcohol i activitat:** Els següents gràfics de barres ens mostren que el nostre conjunt de dades està esbiaixat cap a persones que tenen hàbits saludables. Això es dedueix del fet que les proporcions de persones fumadores, que beuen alcohol i que no es mantenen actives són molt menors a la proporció de persones que davant les mateixes situacions escullen l'alternativa més saludable.



3.2 Detecció de valors atípics

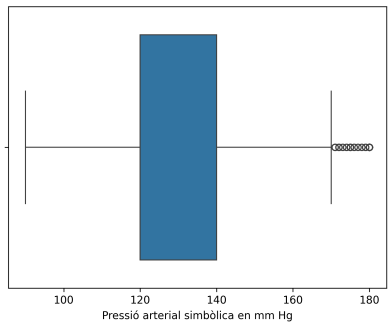
Un cop fet l'anàlisi descriptiva de les característiques, estudiarem els valors atípics de les variables contínues mitjançant boxplots.

- **Edat:**



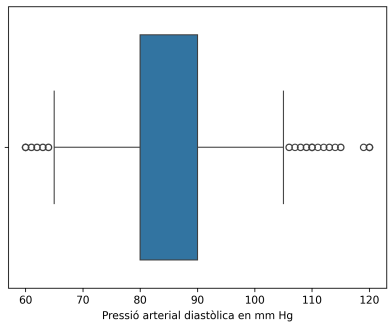
El següent boxplot mostra que la majoria dels pacients tenen edats compreses entre els 45 i 60 anys amb una mitjana d'edat d'entre 50 i 55 anys. S'observen alguns valors atípic a l'esquerre, per sota els 35 anys, cosa que indica que hi ha pocs pacients significativament més joves que la resta, però que no afecten a la distribució general, tal com veiem en el histograma anterior.

- **Pressió sistòlica:**



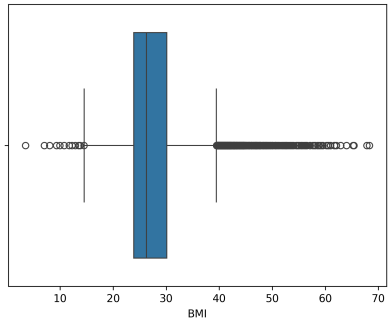
En aquest cas, la pressió sistòlica de la major part dels pacients es troba entre 120 i 14 mm Hg, tal com indica el IQR. D'altra banda, s'observen outliers cap a la dreta, amb valors superiors a 160 mmHg, cosa que és consistent amb casos d'hipertensió severa o crisi hipertensiva, que podrien tenir un impacte significatiu en la predicció i l'anàlisi de risc.

- **Pressió diastòlica:**



En el boxplot, observem que la mitjana de la pressió arterial diastòlica es troba al voltant de 80 mmHg, mentre que el IQR abasta des d'aproximadament 75 mmHg fins a 90 mmHg. Respecte als valors atípics, a la dreta hi ha pressions diastòliques superiors a 100 mmHg, cosa que suggereix casos d'hipertensió severa que afectaran en la predicció. Aquest comportament és consistent amb el nostre estudi, atès que la pressió arterial diastòlica elevada és un factor de risc important per a la detecció de malalties cardiovasculars..

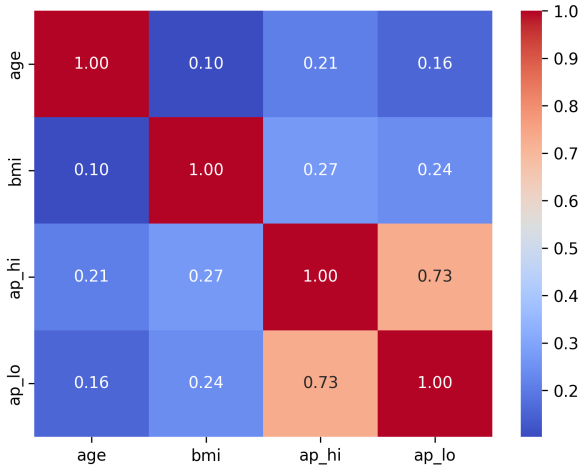
- **BMI:**



Aquest boxplot mostra valors atípics del BMI cap a la dreta. Aquests representen casos d'obessitat severa, que són clínicament rellevants per la seva relació amb el risc cardiovascular. Pel que fa als valors atípics de l'esquerra, tot i que són menys freqüents, poden ser indicatius de malnutrició o condicions associades al baix pes, que també poden influir en la salut cardiovascular.

3.3 Correlació entre les variables contínues

Realitzem un estudi preliminar per estudiar la qualitat de les nostres característiques contínues (age,BMI,ap_hi,ap_lo) en termes de correlacions. Calculem en primer lloc la matriu de correlacions.

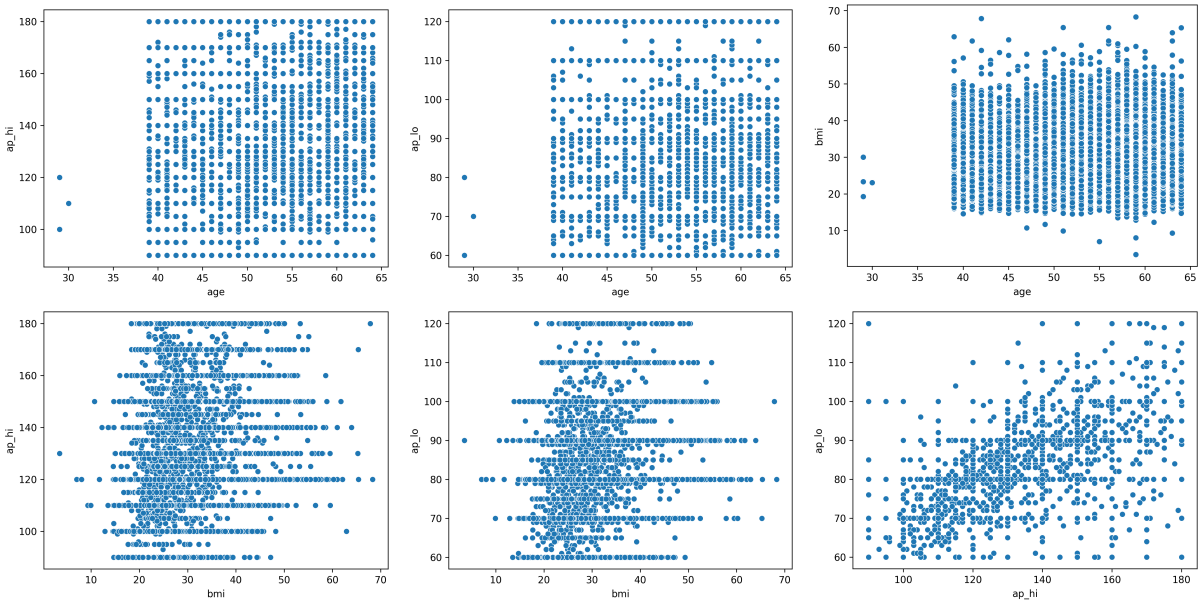


A partir de la matriu de correlació podem extreure diferents conclusions. D'entrada observem que la correlació entre totes les variables es positiva. La relació entre `ap_hi` i `ap_lo` és molt forta (0.73), la qual cosa és lògica ja que són variables interconnectades en el món clínic.

També l'anàlisi mostra que hi ha una correlació moderada entre l'índex de massa corporal i les pressions arterials. Això vol dir que, en general, les persones amb un IMC més alt tendeixen a tenir valors més elevats de pressió arterial sistòlica i diastòlica.

Per últim, cal destacar que la edat té una correlació baixa però significativa amb la pressió arterial (0.21 amb `ap_hi` i 0.16 amb `ap_lo`), reflectint que aquesta tendència és coherent amb la realitat.

A continuació es mostren els gràfics de dispersió per il · lustrar les correlacions existents:



4 Aplicació d'algorismes d'aprenentatge automàtic

Donada l'alta incidència de les malalties cardiovasculars en la salut pública, hem decidit prioritzar la maximització del *recall* en els nostres models. En identificar a tots els pacients amb risc, fins i tot si alguns resultats són falsos positius, podem dur a terme proves més exhaustives sobre aquells que tenen una probabilitat més gran de patir una malaltia. Seguint aquesta estratègia, tot i que pot generar un nombre més gran de proves diagnòstiques, considerem prioritari identificar tots els pacients possibles amb risc de patir malalties cardiovasculars.

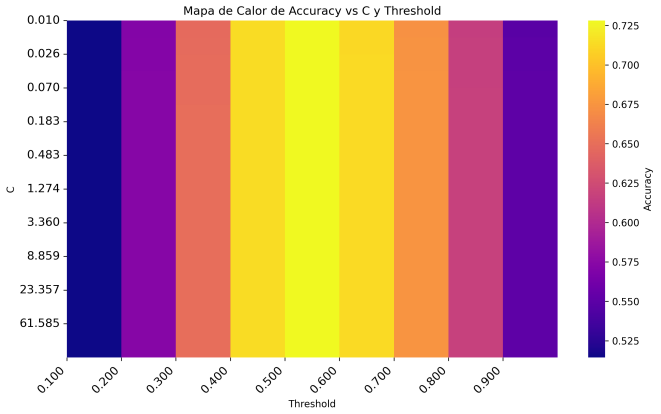
4.1 Regressió Logística

La regressió logística és un mètode d'aprenentatge automàtic supervisat utilitzat principalment per a problemes de classificació binària. L'objectiu del model és predir la probabilitat que una instància pertanyi a una classe determinada mitjançant una combinació lineal de les variables explicatives, la qual es transforma a través d'una funció sigmoide. Aquesta transformació permet que les sortides del model estiguin acotades entre 0 i 1, proporcionant així una interpretació probabilística de les prediccions.

El motiu principal de la tria d'aquest algorisme és la seva simplicitat i interpretabilitat. En l'avaluació de factors de risc cardiovasculars, la comprensió del pes de cada variable en el resultat és fonamental per prendre decisions informades. Els coeficients del model indiquen la força i la direcció de la relació entre cada variable explicativa i la probabilitat de patir una malaltia cardiovascular.

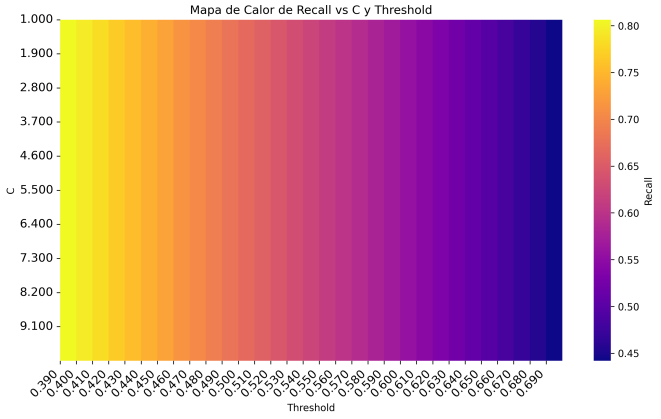
4.1.1 Entrenament del model

Per dur a terme l'entrenament del model de regressió logística hem utilitzat la funció *LogisticRegression*, de la llibreria *sklearn*, amb regularització *Ridge*. Per tal de determinar el millor valor per al *threshold* i l'hiperparàmetre C de la regularització, hem generat el següent mapa de calor:



A la figura anterior s'observa que el model no és gaire sensible als valors de C. A priori això suggereix que és suficient aplicar una regularització estàndard amb $C = 1$.

Pel que fa al *threshold*, podem afirmar que el millor valor es troba entre 0.4 i 0.7, ja que és la franja amb major *accuracy*. Per tal de determinar-ne un valor concret, tindrem en compte que la nostra prioritat és detectar la major part de casos positius possibles. Així, triarem un *threshold* dins d'aquest interval que maximitzi el *recall*.

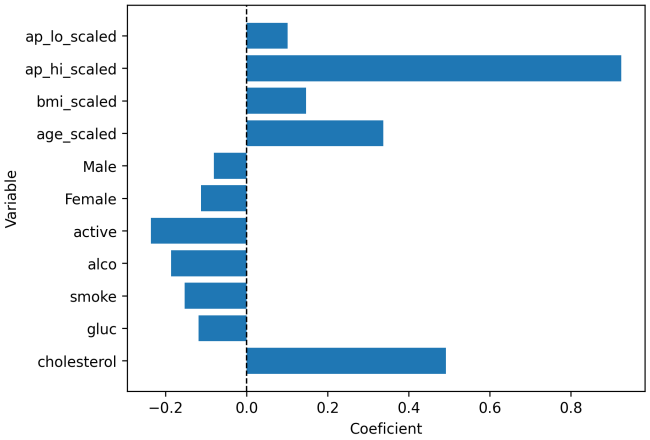


Finalment, tenint en compte els resultats mostrats en la figura anterior, definim *threshold* $= 0.45$ i $C = 1$. Així, després de l'entrenament del model, les mètriques obtingudes en el conjunt de test són:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.73	0.73	0.73	6937
1	0.72	0.73	0.72	6666
Accuracy		0.73		13603
Macro Avg	0.73	0.73	0.73	13603
Weighted Avg	0.73	0.73	0.73	13603

Classification Report

D'altra banda, la importància de cada variable explicativa en la predicció del model ve donada pel valor dels coeficients que a continuació es detallen.



Feature importance

4.2 *Gradient Boosting Tree*

Gradient Boosting Trees (GBT) és un algorisme d'aprenentatge d'ensemble que crea un model predictiu robust a partir de la combinació seqüencial d'arbres de decisió dèbils. La seva principal característica és l'optimització d'una funció de pèrdua específica mitjançant el descens del gradient.

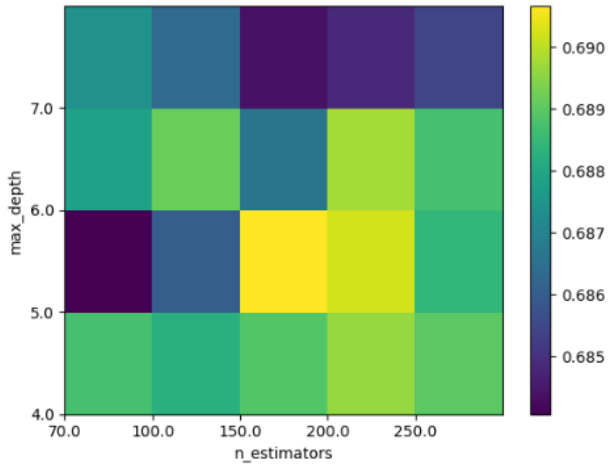
En aquest treball, hem decidit utilitzar aquest mètode d'aprenentatge automàtic per diversos motius. En primer lloc, destaquem l'ús d'aquest algorisme, ja que exhibeix una alta capacitat per modelar relacions no lineals entre variables de naturalesa diversa, incloent-hi variables numèriques i categòriques. Això el fa particularment adequat per a conjunts de dades heterogenis (variables numèriques, categòriques), tal com ocorre en el nostre cas on les nostres característiques són de diferent tipus i la correlació que mostren és d'una estructura complexa, tal com s'ha vist en els gràfics de dispersió anteriors.

D'altra banda, els *Gradient Boosting Trees* permeten calcular una mesura de la importància global de cada característica en el model, la qual cosa és crucial en problemes mèdics per identificar tendències. Aquesta informació permet identificar els indicadors de risc més rellevants per a la predicció de malalties cardiovasculars. De la mateixa manera, a nivell local, i a partir de les tècniques d'interpretabilitat que ofereix *SHAP*, també es poden proporcionar mètriques que avaluïn la importància de cada característica donada una instància particular, cosa que pot ser valuosa en problemes mèdics, personalitzant *insights* que van més enllà de la simple predicció.

4.2.1 Entrenament del model

Per entrenar el model *Gradient Boosting Classifier* seguirem una lògica similar a l'entrenament dels models anteriors. Utilitzarem la funció *GradientBoostingClassifier* de la llibreria *sklearn*, que implementa aquest mètode permetent modificar els hiperparàmetres *n_estimators* i *max_depth*, on *n_estimators* representa el nombre total d'arbres que el model construirà durant l'entrenament i *max_depth* la profunditat màxima de cada arbre.

En primer lloc, realitzarem un mapa de calor per determinar el valor dels hiperparàmetres que maximitzen el recall, la mètrica d'interès. Després d'executar un script de Python s'obté el següent mapa de calor:

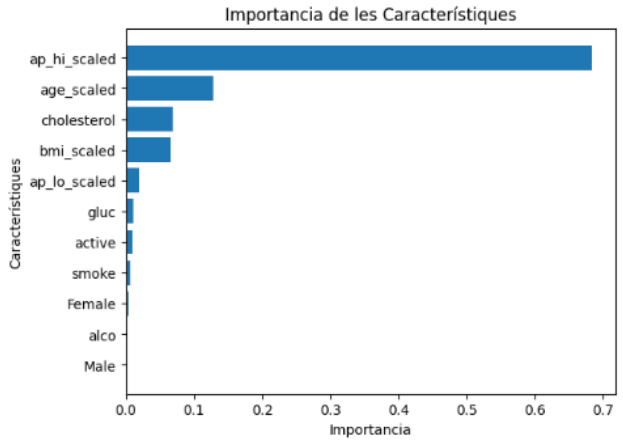


Considerant $n\_estimators = 175$ i $max\_depth = 5$, s'obté un recall de 0.69.

Entrenant un *Gradient Boosting Tree* amb aquests valors, s'obté el següent *Classification Report*.

	precision	**recall**	**f1-score**	**support**
0	0.72	0.78	0.75	6937
1	0.75	0.69	0.72	6666
accuracy			0.74	13603
macro avg	0.74	0.74	0.74	13603
weighted avg	0.74	0.74	0.74	13603

Fent un estudi de l'explicabilitat global del nostre model mitjançant l'anàlisi de la importància de les característiques (a partir de la impuresa de Gini), s'obté que les variables més rellevants per a la predicció venen donades pel següent histograma:



El nostre model revela que els principals factors de risc són la pressió arterial alta, l'edat i l'obesitat, juntament amb la presència d'un alt colesterol. A major pressió arterial sistòlica, a una edat més avançada, una obesitat i colesterol més gran, augmenta la probabilitat de patir alguna malaltia cardiovascular.

4.3 *Random Forest*

El *Random forest* és un algorisme d'aprenentatge automàtic d'ensemble que consisteix en la combinació de diversos arbres de decisió mitjançant la tècnica bagging amb l'objectiu de crear un model més estable i precís reduint la variància.

En aquest treball hem decidit utilitzar el mètode *Random Forest* per diversos motius. El *Random Forest* comparteix dues característiques amb el mètode de *Gradient Boosting*: la capacitat de capturar relacions complexes entre variables de naturalesa diversa i la possibilitat de calcular la importància global de cada característica. Tal com hem vist en la implementació de l'algorisme anterior aquestes particularitats són d'important rellevància per l'estudi del nostre conjunt de dades.

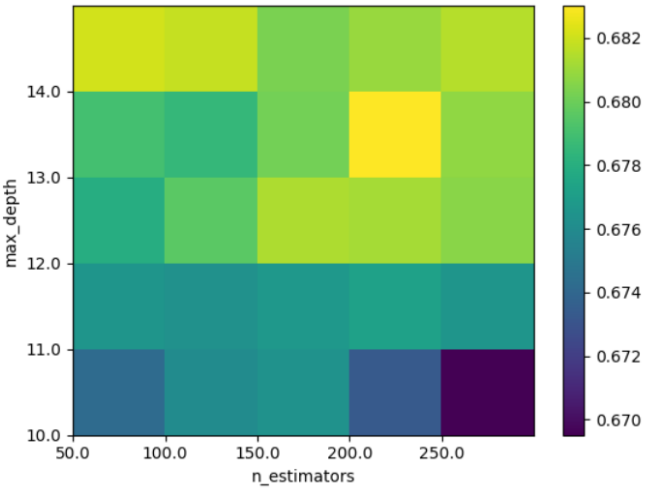
D'altra banda, gràcies a la tècnica del *bagging* el *Random forest* és un mètode amb menys risc d'overfitting. En l'àmbit mèdic és comú que els conjunts de dades continguin soroll, ja que hi ha factors tecnològics i humans que

poden limitar la precisió dels mesuraments. A causa d'això, és crucial per l'estudi que els models siguin robustos davant l'overfitting.

4.3.1 Entrenament del model

Per entrenar el model *Random Forest* utilitzarem la funció *RandomForestClassifier* de la llibreria *sklearn*, que implementa aquest mètode permetent modificar els hiperparàmetres *n_estimators* i *max_depth*, on *n_estimators* representa el nombre total d'arbres que el model construirà durant l'entrenament i *max_depth* la profunditat màxima de cada arbre.

En primer lloc, realitzarem un mapa de calor per determinar el valor dels hiperparàmetres que maximitzen el recall, la mètrica d'interès. Després d'executar un script de Python s'obté el següent mapa de calor:

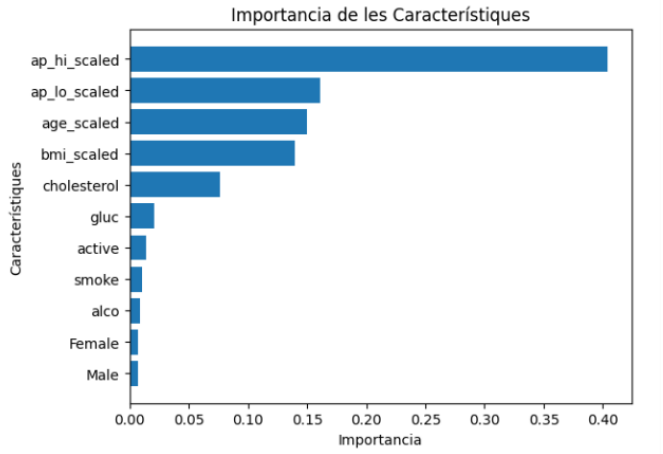


Considerant $n\_estimators = 225$ i $max\_depth = 13$, s'obté un recall de 0.68.

Entrenant un *Random Forest* amb aquests valors, s'obté el següent *Classification Report*.

	precision	**recall**	**f1-score**	**support**
0	0.72	0.78	0.75	6937
1	0.75	0.68	0.72	6666
accuracy			0.73	13603
macro avg	0.74	0.73	0.73	13603
weighted avg	0.74	0.73	0.73	13603

Fent un estudi de l'explicabilitat global del nostre model mitjançant l'anàlisi de la importància de les característiques (a partir de la impuresa de Gini), s'obté que les variables més rellevants per a la predicció venen donades pel següent histograma:



El nostre model revela que els principals factors de risc són la pressió arterial alta, la pressió arterial baixa, l'edat i l'obesitat. A major pressió arterial sistòlica i diastòlica, a una edat més avançada i una obesitat més gran, augmenta la probabilitat de patir alguna malaltia cardiovascular.

5 Resultats i interpretació

5.1 Comparació dels models en termes de rendiment

La següent taula presenta una anàlisi comparativa dels nostres tres models de classificació (*regressió logística*, *Random Forest* i *Gradient Boosting Trees*) mitjançant mètriques d'avaluació clàssiques com ara l'*accuracy*, la *precision*, el *recall* i el *F1-score*:

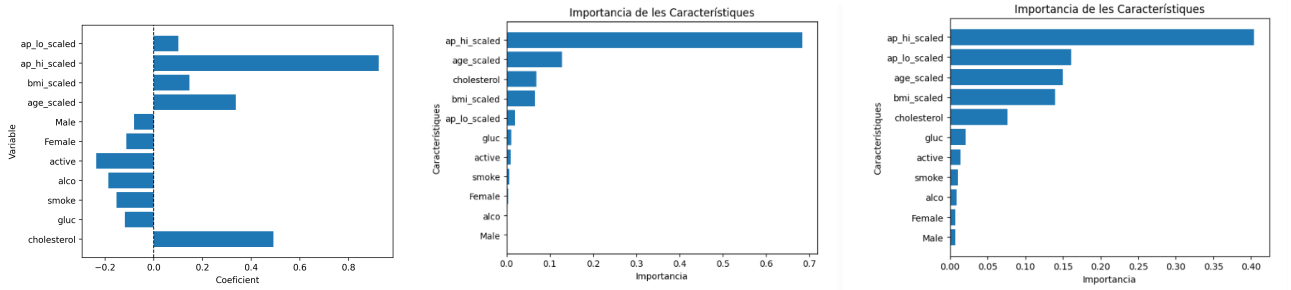
Métrica	Clase	Regressió logística	*Random forest*	*Gradient boosting*
Precision	0	0.73	0.72	0.72
	1	0.72	0.75	0.75
	Macro Avg	0.73	0.74	0.74
	Weighted Avg	0.73	0.74	0.74
Recall	0	0.73	0.78	0.77
	1	0.73	0.68	0.69
	Macro Avg	0.73	0.73	0.74
	Weighted Avg	0.73	0.73	0.74
F1-Score	0	0.73	0.75	0.74
	1	0.72	0.72	0.72
	Macro Avg	0.73	0.73	0.74
	Weighted Avg	0.73	0.73	0.74
Support	0	6937	6864	6864
	1	6666	6739	6739
	Total	13603	13603	13603
Accuracy	-	0.73	0.73	0.73

En observar el recall, podem constatar que *Random Forest* destaca en la detecció d'instàncies de la classe 0, assolint un valor notablement alt de 0.78, superant la *Regressió Logística* i el *Gradient Boosting*, que obtenen valors de 0.73 i 0.77, respectivament. No obstant això, per a la classe 1, *Random Forest* obté un recall més baix (0.68), la qual cosa indica una falta d'equilibri en la seva capacitat per identificar correctament ambdues classes. Per la seva part, *Gradient Boosting* segueix un patró similar, tot i que amb resultats una mica més equilibrats (0.72 per a la classe 1 i 0.77 per a la classe 0). En canvi, la Regressió Logística mostra un rendiment uniforme, amb un recall de 0.73 per a ambdues classes, la qual cosa ressalta la seva capacitat de tractar ambdues classes de manera equitativa.

Els tres models assoleixen una accuracy global similar, però difereixen en el seu rendiment per classe. La *Gradient Boosting* ofereix un equilibri òptim, mentre *regressió logística* que prioritza la identificació de la classe positiva a costa de la negativa. Per aquest motiu, nosaltres escollim l'ús de la Regressió Logística per a futures prediccions en el futur, atès que és el model que maximitza el *recall* i té un rendiment equilibrat en ambdues classes.

5.2 Interpretació dels resultats i implicacions dels resultats en el context del conjunt de dades utilitzat

A continuació mostrem els gràfics de *feature importance* obtinguts per cada model que mostren la importància relativa de les característiques en el conjunt de dades utilitzat.



D'esquerra a dreta: Regressió logística, *Gradient Boosting* i *Random forest*

Pel que fa a la regressió logística, un impacte positiu d'una característica està associat amb un increment en la probabilitat de patir una malaltia cardiovascular, mentre que un valor negatiu indica l'efecte contrari. Així, podem concloure que els principals factors de risc identificats són una pressió arterial alta, l'edat avançada i un índex de massa corporal elevat (sobrepès).

D'altra banda, el gràfic del *Gradient Boosting* mostra quines són les característiques més importants quant al guany en la impuresa de cada node. De manera consistent amb els resultats de la regressió logística, s'observa que les variables més rellevants són la pressió arterial alta, l'edat, el colesterol i el sobrepès, destacant així el seu paper com a factors determinants en el model.

Finalment, en línia amb els resultats dels models anteriors, el gràfic corresponent al *Random Forest* destaca com a variables més rellevants la pressió arterial (tant la sistòlica com la diastòlica), l'edat, l'índex de massa corporal i el colesterol. No obstant això, cal considerar que la forta correlació habitual entre les mesures de pressió arterial alta i baixa podria estar influint en la percepció d'importància atribuïda a aquestes variables, degut a la redundància d'informació que aporten al model.

En resum, podem concloure que els factors clau per determinar el risc de patir una malaltia cardiovascular són la pressió arterial alta, l'edat i l'índex de massa corporal (BMI). En un context mèdic, aquests resultats suggereixen que la prevenció hauria de centrar-se en controlar la pressió arterial alta i en promoure un pes saludable com a estratègies prioritàries per reduir el risc cardiovascular. Per contra, característiques com la glucosa, l'activitat física i el consum d'alcohol semblen tenir un pes menys significatiu en les prediccions realitzades pels models.

6 Conclusions i Treball Futur

6.1 Conclusions generals sobre l'ús d'aprenentatge automàtic per al conjunt de dades analitzat

L'aplicació de tècniques d'aprenentatge automàtic en el conjunt de dades analitzat, que inclou variables com el gènere, els nivells de colesterol i glucosa i factors d'estil de vida com el tabaquisme o l'activitat física, ha demostrat ser útil per modelar i predir el risc de malalties cardiovasculars. En particular, la utilització de la regressió logística, *Random Forest* i *Gradient Boosting* ha permès establir relacions clares entre els factors de risc i el resultat esperat, proporcionant una base sòlida per a la interpretació clínica.

L'anàlisi exploratori de les dades ha revelat patrons complexos i no lineals entre les variables. L'aprenentatge automàtic, gràcies a la seva capacitat per identificar patrons en grans conjunts de dades, ha demostrat ser una eina potent per modelar aquestes relacions de manera efectiva. Els resultats obtinguts donen suport a la hipòtesi que les tècniques d'aprenentatge automàtic són capaces de capturar la complexitat que presentava el nostre conjunt de dades.

Així doncs, l'aplicació de diversos models d'aprenentatge automàtic revela que l'índex de massa corporal, l'edat i la pressió sistòlica emergeixen com els factors de risc més significatius associats a la condició estudiada. Això es correspon amb la realitat i mostra la potència de l'aprenentatge automàtic per identificar relacions complexes que podrien no estar corroborades.

6.2 Limitacions del treball i possibles millores futures

Un possible factor limitador és que, tot i que els resultats obtinguts en aquest estudi suggereixen que els models d'aprenentatge automàtic poden ser útils per predir malalties cardiovasculars, és important destacar que el nostre conjunt de dades presenta un biaix cap a individus amb factors de risc més baixos. Com vam veure en l'anàlisi exploratòria de les dades, hi ha un biaix cap a individus no fumadors, amb nivells de glucosa normals i hàbits de vida actius, la qual cosa pot limitar la generalització dels resultats a poblacions amb perfils de risc més elevats.

Per aquest motiu, la recollida de dades mèdiques hauria de centrar-se en poblacions més diverses i en l'avaluació de la capacitat dels models per identificar individus amb múltiples factors de risc.

6.3 Propostes de treball futur amb altres tècniques d'aprenentatge automàtic o altres conjunts de dades

Tot i que aquest estudi ha proporcionat evidència sòlida de la utilitat dels models d'aprenentatge automàtic en la predicció de malalties cardiovasculars, és crucial destacar que considerem que la participació activa de professionals de la salut en el desenvolupament i validació d'aquests models és essencial per garantir que les conclusions obtingudes siguin clínicament rellevants i útils.

Una direcció important per a la feina futura seria la implementació d'aquests models en entorns clínics reals, avaluant el seu impacte en la pràctica mèdica mitjançant estudis prospectius. Això ajudaria a determinar no només la seva precisió, sinó també la seva utilitat pràctica per millorar la salut dels pacients.

Es proposa l'ús d'aquest model com una eina inicial en l'avaluació del risc cardiovascular, permetent identificar aquells pacients que podrien requerir una avaluació més detallada. Els resultats obtinguts poden servir com a punt de partida per a una estratificació del risc i la implementació d'estratègies de prevenció personalitzades.