Predicció de Malalties Cardíaques mitjançant Tècniques d'Aprenentatge Automàtic

Víctor Moreno Borràs
Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona
Universitat Politècnica de Catalunya
victor.moreno.borras@estudiantat.upc.edu

***Abstract*—Aquest document presenta l'estudi del desenvolupament d'un sistema predictiu per a malalties del cor utilitzant tècniques d'aprenentatge automàtic. Per l'entrenament, s'ha utilitzat un conjunt de dades d'una regió de Sud-àfrica i s'han explorat diversos mètodes de classificació, incloent-hi la regressió logística, el Gaussian Naive Bayes i el Random Forest. No obstant això, els resultats obtinguts no són gaire bons, possiblement a causa de la petita mida i la distribució desequilibrada del conjunt de dades.**

I. Introducció

Aquest document presenta un treball realitzat en el marc del seminari d'Aprenentatge Automàtic. L'objectiu es basa a desenvolupar un sistema predictor de malalties cardíaques a partir de la informació clínica dels pacients, plantejat com una competició a la plataforma Kaggle [3]. Tot el codi d'aquest projecte es pot trobar a [2].

Per entrenar i avaluar el sistema es disposa d'una base de dades provenen d'una regió de Western Cape, a Sud-àfrica. La base de dades consta de 462 mostres, cada una representada per un vector de 9 característiques. D'aquest total, 302 mostres corresponen a subjectes sans i 160 a subjectes amb malaltia coronària:

Les 9 característiques són les següents:

- sbp: pressió arterial sistòlica
- tobacco: tabac acumulat (kg)
- ldl: colesterol de lipoproteïnes de baixa densitat
- adiposat: adipositat
- famhist: antecedents familiars de malalties cardíaques
- typea: comportament de tipus A
- obesitat: obesitat
- alcohol: consum actual d'alcohol
- edat: edat d'inici
- La classe associada a cada subjecte es defineix com a malaltia coronària (1) o sa (0).

II. Mètodes de Classificació Proposats

En aquesta secció, es descriuen els mètodes de classificació proposats per abordar el problema de classificació. Després de provar diversos mètodes vistos a classe i al laboratori, s'ha decidit aprofundir l'estudi en tècniques que han donat uns resultats inicials més bons. Durant tot l'estudi, s'han utilitzat les classes proporcionades per la llibreria Scikit-learn [4] de Python.

A. Regressió Logística

La Regressió Logística [5] és un mètode de classificació lineal que modela la relació entre les característiques clíniques i la probabilitat de pertànyer a una classe específica. Sigui **X** la matriu que representa les característiques clíniques dels pacients i **y** el vector de classes binàries (0 o 1) que indica si un pacient té o no malaltia coronària. La Regressió Logística estima la probabilitat condicional $Pr(y_i = 1|\mathbf{X}_i)$ utilitzant la funció logística:

$$\hat{p} = Pr(y_i = 1|\mathbf{X}_i) = \frac{1}{1 + exp(-\mathbf{X}_i\mathbf{w} - w_o)} \quad (1)$$

on w_0 és biax i el vector **w** són els coeficients de regressió. Com a problema d'optimització, la regressió logística amb el terme de regularització $r(w)$ minimitza la funció de cost següent:

$$\min_w C \sum_{i=1}^{n} \left(-y_i \log(\hat{p}) - (1 - y_i) \log\left(1 - \hat{p}\right)\right) + r(w) \quad (2)$$

A més, la classe **LogisticRegression** [4] proveeix de tres termes de regularització:

penalització	**r(w)**
None	0
ℓ_1	$\|w\|_1$
ℓ_2	$\frac{1}{2}\|w\|_2^2 = \frac{1}{2}w^T w$

B. Gaussian Naive Bayes

Els mètodes de Naive Bayes són un conjunt de tècniques d'aprenentatge automàtic basades en la probabilitat i el teorema de Bayes. Aquests mètodes són utilitzats principalment per a problemes de classificació i es basen en l'assumpció de la independència condicional entre les característiques. Una de les variants populars de Naive Bayes i l'aplicada en aquest estudi és el Gaussian Naive Bayes, que assumeix que les dades segueixen una distribució Gaussiana (normal) per a cada classe.

$$Pr\left(x_i \mid y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \quad (3)$$

Com s'ha dit, en el nostre cas es considera que cada $Pr(x_i|y)$ segueix una distribució Gaussiana (3) amb una mitjana μ_y i una desviació estàndard σ_y específiques per a cada classe chd. Això significa que cal estimar aquestes mitjanes i desviacions estàndard per a les classes no malalt i malalt utilitzant les dades d'entrenament.

C. Random Forest

Com a últim mètode, estudiarem el **Random Forest** [6] un mètode d'aprenentatge supervisat que combina múltiples arbres de decisió per a la classificació. Cada arbre s'entrena amb una submostra aleatòria del conjunt de dades original, així com una submostra aleatòria de les característiques. L'ús de diverses submostres i l'agregació de les prediccions dels arbres resultants permet reduir la variància i millorar la generalització.

Cada arbre de decisió en el Random Forest pren decisions basades en les característiques clíniques i divideix el conjunt de dades en subconjunts més petits. La classificació final es realitza per majoria de vots entre els arbres. Aquesta combinació de múltiples arbres proporciona una classificació més robusta i pot tenir en compte relacions no lineals entre les característiques clíniques i la presència de malaltia coronària.

D. Voting (Votació)

La votació [6] és una tècnica que combina les prediccions de diversos classificadors per arribar a una predicció final. En aquest cas, es proposa utilitzar un Voting que combina les prediccions de la Regressió Logística, el Gaussian Naive Bayes i el Random Forest. El Voting pot prendre diferents formes, com ara un vot majoritari (Hard Voting) o una mitjana ponderada de les probabilitats predites (Soft Voting).

III. Mesures de la Qualitat de la Solució

A banda de treballar amb la **F1-score** (4), la requerida per el Kaggel [3], és important considerar altres mesures de qualitat. Per exemple el **Recall** (3), també conegut com a sensibilitat. Aquesta mesura és especialment rellevant en el context de les dades clíniques, on és crucial detectar adequadament tots els casos d'enfermetat (chd = 1).

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\textbf{F-score} = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision} \tag{5}$$

El Recall mesura la proporció de casos positius (TP) que són correctament detectats pel model. És a dir, indica la capacitat del sistema per identificar tots els pacients amb malaltia cardíaca. En aquest context, és molt més important tenir un Recall alt que un Precision alt. Si un model té un Recall baix, significa que hi ha un elevat nombre de falsos negatius, és a dir, pacients amb malaltia cardíaca que el model no identifica. Això pot tenir conseqüències greus per als pacients, ja que es perden o es retarden oportunitats de diagnòstic i tractament.

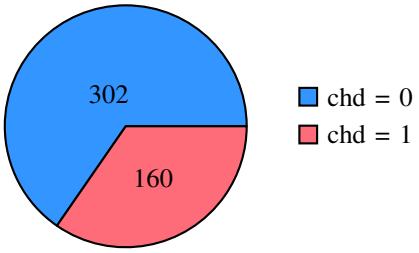

Fig. 1. Distribució dels casos a la base de dades

A més, cal tenir en compte que la base de dades utilitzada per al sistema predictor no està balancejada (Figure 1). Hi ha aproximadament dos casos sans per cada cas de malaltia coronària. Aquest desequilibri pot afectar les mesures de qualitat, ja que un model pot optar per una predicció més inclinada cap a la classe majoritària per obtenir millors resultats en termes de precisió global. No obstant això, aquest enfocament pot resultar en un baix Recall per a la classe minoritària (malaltia cardíaca), que és precisament la classe que ens interessa identificar correctament.

Per tant, en aquesta tasca, és essencial considerar tant la F1-score com el Recall com a mesures de qualitat. La F1-score proporciona un equilibri entre el Recall i la Precisió, mentre que el Recall ens dóna una indicació directa de la capacitat del sistema per detectar els casos positius. Així, ambdós ajuden a avaluar l'efectivitat i la fiabilitat del sistema predictor de malalties cardíaques.

IV. Descripció dels Experiments

A. Anàlisis de Dades

Prèviament a l'entrenament dels models, s'ha elaborat una anàlisi exploratòria de les dades [1] per comprendre millor les relacions entre les variables. A banda d'això, un cop entrenats els models, s'ha fet un estudi de la importància de cada característica en el mètode de Regressió Logística.

1) Correlació de Pearson: En primer lloc, s'ha calculat la correlació de Pearson entre les variables. La correlació de Pearson mesura la força i la direcció de la relació lineal entre dues variables. Es va utilitzar aquesta mètrica per avaluar la relació entre les característiques del conjunt de dades

La Figura 2 mostra els coeficients de correlació de Pearson per parelles de columnes. Com es pot observar, s'aprecia una alta correlació entre les variables "obesity" i "adiposity". Aquesta correlació suggereix que aquestes dues variables estan estretament relacionades. En aquest sentit, pot ser interessant prescindir d'una de les característiques a l'hora d'entrenar els models, per evitar la multicolinealitat.

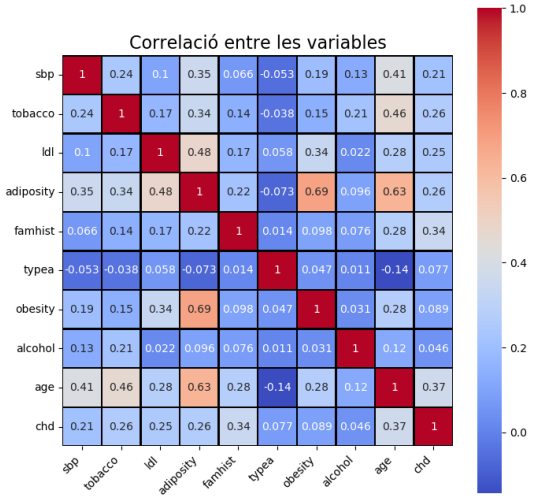

Fig. 2. Correlació de Pearson entre les característiques clíniques

2) Importància de les Característiques en la Regressió Logística: Un cop entrenat el model de Regressió Logística es va realitzar un estudi de la influència de cada característica en aquest mètode. Aquesta anàlisi es va basar en els coeficients del vector de pesos **w** del model obtingut.

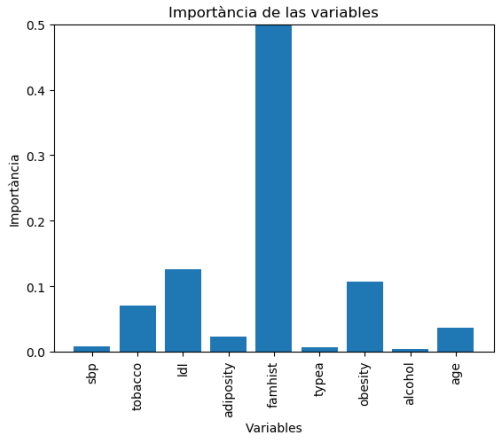

Fig. 3. Importància de les característiques per al Logistic Regression

Posteriorment, a la secció V, es tenen en compte aquests resultats.

B. Millors Models

Prèviament a l'estudi exhaustiu d'uns pocs models, s'ha realitzat una prova amb un conjunt ampli de models per identificar aquells que ofereixen millors resultats inicials. Aquesta prova inclou models vists a teoria i al laboratori. Mitjançant les mètriques de rendiment vistes en l'anterior secció, s'han seleccionat els models més bons per a una anàlisi més detallada. Cal destacar que aquesta prova inicial no garanteix la idoneïtat dels models seleccionats, però proporciona una orientació inicial per a l'estudi posterior. A la Figura 5 es poden veure els resultats d'aquesta prova.

C. Hiperparàmetres

Un cop seleccionats els models més prometedors, s'ha realitzat un estudi detallat d'aquests. Per a això, s'ha dividit la base de dades en diferents conjunts: 20 % per a les dades de prova (test) i el 80 % restant per a l'entrenament (train). En alguns casos, per a models específics, s'ha realitzat una divisió addicional del 15 % de les dades d'entrenament per a la validació i el 65 % restant per a l'entrenament principal. Les dades de validació s'han utilitzat per a l'optimització d'altres hiperparàmetres, com la frontera de decisió γ, en el **Logistic Regression**.

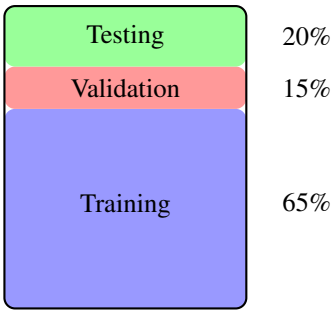

Fig. 4. Partició de la base de dades per a la Regressió Logística

A més, s'ha utilitzat la tècnica de validació creuada "K-Folds", per a trobar els millors hiperparàmetres per a cada model a partir de les dades d'entrenament. Concretament, s'ha fet servir l'eina "Grid Search" de la llibreria Sklearn [4], la qual explora de manera sistemàtica diferents combinacions d'hiperparàmetres i retorna el millor estimador.

V. Resultats

Com podem veure a la Figure 4, primer s'ha fet una primera avaluació de diferents mètodes d'aprenentatge supervisat. Els tres millors mètodes s'han escollit a partir de la mètrica **F1-score**. Aquests son els enumerats en la secció II:

- Regressió Logística: **LogisticRegression**
- Gaussian Naive Bayes: **GaussianNB**
- Random Forest: **RandomForestClassifier**

El cas del model de Regressió Logística, és el que s'ha estudiat en més profunditat. Concretament, s'ha treballat dos aspectes: la influència de les característiques en el model i l'optimització de la frontera de decisió.

En primer lloc, s'ha trobat que el comportament del sistema millorava a partir d'ometre característiques de la base de dades amb poca influència. A partir d'aquests resultats (Figura 5), s'ha decidit ometre les següents característiques: 'sbp','typea' i 'alcohol'. A banda d'això, també s'ha buscat optimitzar la frontera de decisió γ, que per defecte és 0.5, maximitzant la F1-score a partir de les dades de validació. Aquest son els resultats:

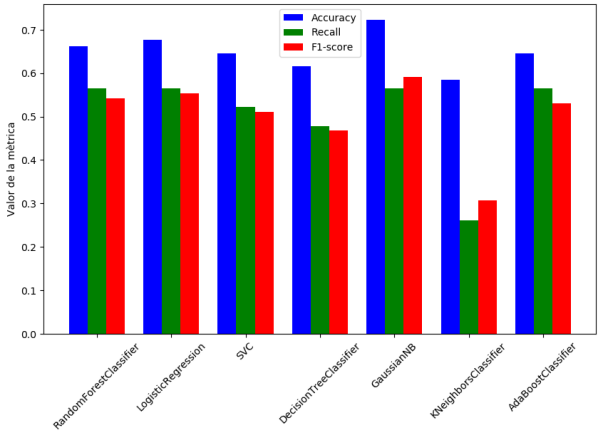

Fig. 5. Mètriques per a diferents models d'aprenentatge supervisat

Logistic Regression	**Accuracy**	**F1-score**
Normal	0.662	0.522
Menys Variables	0.694	0.565
γ **óptima**	0.708	0.578

TABLE I

Es pot observar a la Taula 1, que ometen certes característiques i optimitzant γ s'ha pogut millorar el comportament del Regressor Logístic.

Per altra banda, en el cas del mètode Gaussian Naive Bayes el sistema no s'ha pogut optimitzar més, ja que no conté hiperparàmetres. En canvi, el Random Forest, fen ús del Grid Search s'ha pogut millorar lleugerament el seu comportament. Per acabar, s'han provat tots els mètodes, inclosos el Hard and Soft Voting. Els resultats finals sobre les dades de test són els següents:

	Accuracy	**Recall**	**F1-score**
Logistic Regressor (γ opt.)	0.708	0.555	0.578
Gaussian NB	0.723	0.565	0.591
Random Forest	0.708	0.478	0.537
Hard Voting	0.723	0.523	0.558
Soft Voting	0.723	0.609	0.609

TABLE II

Els resultats presentats es basen en les dades de prova, tot i que no mostren valors similars als que s'han obtingut finalment a Kaggle. Per prendre una decisió final sobre quin model utilitzar per a l'enviament a Kaggle, s'ha intentat trobar el model que proporcionava els millors resultats tant a Kaggle com a les dades de prova. Aquest model és la Regressió Logística amb la frontera de decisió optimitzada, amb un F-score de 0.666 en la puntuació pública i una puntuació privada de 0.5614.

VI. Conclusions

Després d'analitzar els resultats obtinguts en aquest estudi, es pot concloure que els valors mètrics presentats són relativament baixos. Això pot ser degut a diversos factors, com ara la mida de la base de dades, que és relativament petita, i la seva distribució desequilibrada, amb una majoria de casos negatius i només uns pocs casos positius. Això pot afectar la capacitat dels models per detectar correctament els casos positius, ja que poden optar per una predicció més inclinada cap a la classe majoritària per obtenir millors resultats en termes de precisió global. A més, cal tenir en compte que la base de dades només es va obtenir en una regió, la qual cosa pot limitar la seva generalització a altres poblacions. Això pot afectar la capacitat dels models per detectar correctament els casos positius en altres.

Un altre factor important a destacar és que no s'ha aconseguit un Recall alt per a la classe minoritària (malaltia cardíaca), que és precisament la classe que ens interessa identificar correctament. Això és important perquè el Recall ens dona una indicació directa de la capacitat del sistema per detectar els casos positius, i en aquest cas, és crucial per a la detecció precoç de la malaltia cardíaca.

En resum, tot i que s'han realitzat diversos experiments per optimitzar els models i s'ha utilitzat una tècnica de votació per combinar les prediccions de diversos classificadors, els resultats obtinguts són relativament dolents. Això pot ser degut a diversos factors relacionats amb la base de dades, com ara la seva grandària, la seva distribució desequilibrada i la seva limitació a una sola regió.

References

[1] Víctor Moreno Borràs. Data analysis. https://colab.research.google.com/drive/1TEqBswlXmDBxqQvwWgFr4mf12Vl5dGx8?us 2023.
[2] Víctor Moreno Borràs. Predicció de malalties cardíaques. https://colab.research.google.com/drive/1rbRAr6mwsY6aKOVwyayP8Iu7jg-1LURp?usp=sharing, 2023.
[3] verovilaplana Josep Vidal. Hearth disease prediction, 2023.
[4] 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.
[5] Verónica Vilaplana. Lab 3 - logistic regression, 2023.
[6] Verónica Vilaplana. Lab 5 - decision trees, random forests, other ensembles, 2023.