Predicció de malalties cardíaques mitjançant Machine Learning

Alba Puig Font

Universitat Autònoma de Barcelona

07/12/2023

Abstract -- *Les malalties cardiovasculars són la principal causa de mort als països desenvolupats i un dels problemes actuals de salut pública més rellevants. Aquest tipus de malaltes poden afectar qualsevol persona, independentment de l'edat, el gènere o l'estatus social.*

Amb aquest estudi es demostra el potencial de l'ús del Machine Learning per a la predicció de malalties del cor, mitjançant els algorismes Logistic Regression i XGBoost en models binaris o multiclasse respectivament. Aquest anàlisi es transforma en un aplicatiu web amb dos perfils d'usuari (models) destinat a la presa de decisions en base a situacions reals.

1. Introducció

El cor és un dels òrgans més importants del cos humà. Un dels casos més complicats i complexos en el camp de la medicina és la predicció de les malalties cardíaques. La predicció de les malalties del cor és una tasca difícil i arriscada i en la que la precisió és un factor rellevant, ja que afecta directament la salut de les persones. Si no es prediu amb precisió, es poden cometre errors amb conseqüències indesitjables.

Aquesta investigació se centra, per tant, en la comparació de diferents tècniques de classificació de dades per tal de predir l'existència de malalties cardiovasculars. En el nostre cas hem utilitzat la Regressió Logística, Random Forest i XGBoost, ja que són algunes de les tècniques més utilitzades en la determinació de malalties.

L'objectiu d'aquest projecte és la detecció de malalties cardíaques mitjançant l'ús d'un sistema d'aprenentatge automàtic. En l'àmbit de la sanitat, l'aprenentatge automàtic pot ajudar els metges en la realització de prediccions més precises per als pacients i, a més, pot incrementar la velocitat de processament i anàlisi de les dades.

El producte final d'aquesta anàlisi s'engloba dins d'un aplicatiu web que pot ser usat com a eina de predicció de malalties cardiovasculars.

Amb la intenció d'explotar al màxim les possibilitats d'aquesta eina predictiva s'ha analitzat l'estudi a dos nivells. Un primer model (model I) de tipus binari, que classifica entre les classes 0 i 1, on 0 correspon a una persona sana i 1 a una persona amb malaltia cardíaca. El segon model (model II), es tracta d'un model multiclasse que classifica entre les classes 0, 1, 2, 3 i 4 corresponents a diferents nivells de malaltia.

2. Metodologia

La metodologia proposada (Figura 1) inclou diversos passos.

En primer lloc, partint de les dades adquirides de la base de dades "Heart Disease UCI"[(1)] s'ha dut a terme una exploració inicial (Exploratory Data Analysis) per obtenir una visió de la seva distribució, abast i qualsevol valor atípic. Amb aquesta anàlisi es pot identificar qualsevol problema potencial amb les dades

(1) Base de dades obtinguda de Kaggle, una plataforma en línia per a científics de dades que ofereix conjunts de dades, competicions de ciència de dades, quaderns de codi i cursos d'aprenentatge.

i orienta les decisions sobre la neteja i el preprocessament de dades.

En segon lloc, s'ha fet una neteja de dades,

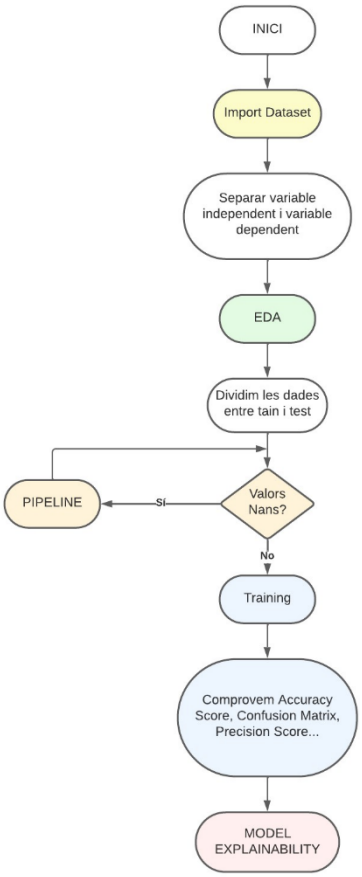

Figura 1

un pas fonamental en qualsevol projecte d'aprenentatge automàtic, per garantir que les dades siguin precises i de confiança per al modelatge i l'anàlisi. Quan es tracta de predicció de malalties cardíaques, la neteja de dades implica detectar i corregir qualsevol error o inconsistència en el conjunt de dades que podria afectar la precisió del seu model predictiu.

En tercer lloc, com a pas crític en aquest procés d'aprenentatge automàtic s'ha fet la selecció del millor model. Després de la preparació de les dades, es procedeix a entrenar i avaluar diversos models amb l'objectiu de trobar aquell que ofereixi el millor rendiment en la predicció de malalties cardíaques. En aquest cas, s'han avaluat tres models diferents: Logistic Regression, Random Forest i XGBoost. Cada model es configura amb diferents hiperparàmetres i es mesura el seu rendiment utilitzant una validació creuada (cross-validation) amb una mètrica personalitzada. En el cas del model I s'ha usat la precisió màxima quan la Recall és superior a 0,95, metre que en el model II s'ha utilitzat una mètrica molt més estàndard, la F1-Score.

Finalment, es realitza un "model explainability", crucial per a comprendre el funcionament intern del model i conèixer les característiques que més contribueixen a les prediccions. Una eina comuna per al "model explainability" és SHAP (SHapley Additive exPlanations), que proporciona una mesura unificada de la importància de les característiques.

Aquesta metodologia s'ha aplicat en els dos models estudiats.

3. Resultats

La seqüencia d'estudi dels resultats ha estat la següent: gràfic de la corba PR i gràfic de prediction vs probability, matriu de confusió, classification report i model explainability.

Model I:

Per al model I els tres algorismes tenen un bon funcionament, el Logistic Regression té una puntuació de la mètrica lleugerament més alta, seguit del Random Forest, i el XGBoost té la puntuació més baixa. Això indica que, per a aquest conjunt de dades específic i la mètrica de puntuació utilitzada,

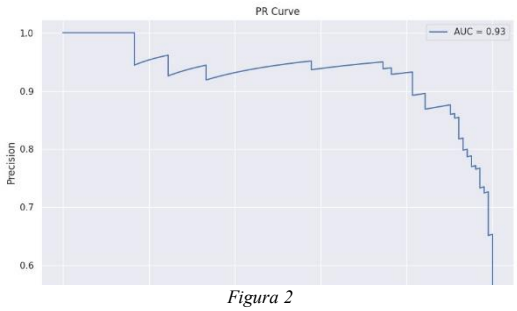

Figura 2

la Regressió Logística està funcionant millor que els altres dos algoritmes.

La Figura 2 mostra la corba de Precision-Recall (PR) que representa la relació entre la Precision (proporció de veritables positius entre les instàncies classificades com a positives) i la Recall (mesura la proporció de veritables positius que han estat correctament identificats pel model) per a diferents llindars de probabilitat. La corba PR té un AUC (àrea sota la corba) de 0,93. Això indica que el model té un bon rendiment, ja que un AUC d'1,0 seria perfecte.

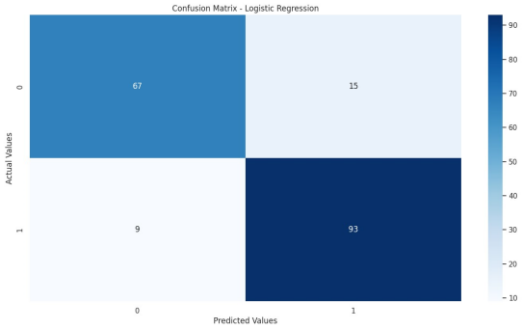

Figura 3

D'acord amb la matriu de confusió obtinguda per aquest model (Figura 3), observem que el model ha identificat correctament 67 instàncies com a negatives (TN) i 93 instàncies com a positives (TP). No obstant això, s'han produït alguns falsos positius (FP = 15) i falsos negatius (FN = 9), indicant algunes prediccions equivocades. Tot i això, aquests errors són baixos en comparació amb les instàncies correctament classificades.

El "classification report" mostra un rendiment general bastant bo en la predicció de la presència de malalties cardíaques. Amb una precisió global del 87%, el model té una capacitat notable per classificar correctament les instàncies entre les dues classes (0 i 1). En resum, el model sembla tenir un bon rendiment, amb una precisió, Recall i F1-Score al voltant del 87%.

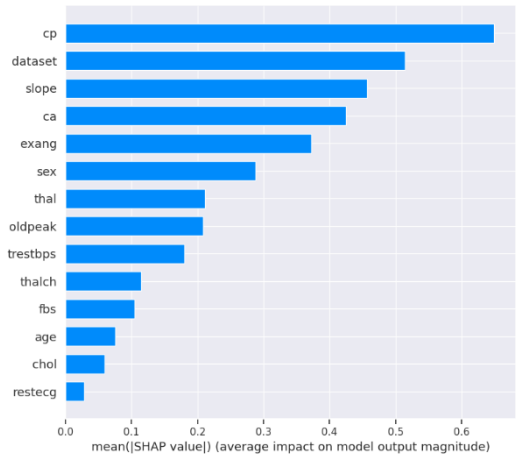

Figura 4

Els gràfics SHAP (SHapley Additive exPlanations) proporcionen una visió interpretativa dels resultats del model, destacant la contribució de cada característica a les prediccions del model. La Figura 4 permet veure quines característiques són més importants per a les prediccions del model. Les característiques amb barres més llargues tenen un major impacte en la sortida del model. Això pot ser útil per entendre millor com el model està fent les seves prediccions.

Els resultats indiquen que "cp"(chest pain), "dataset" i "slope" tenen un paper destacat en la determinació dels resultats, mentre que "restecg" té el menor impacte.

Model II:

Per al Model II, XGBoost mostra una puntuació d'F1-Score lleugerament superior, seguit pel Random Forest, mentre que el Logistic Regression presenta la puntuació més baixa. Aquestes dades suggereixen que, per a aquest conjunt específic de dades i la mètrica F1-Score, el XGBoost està funcionant millor que els altres dos algoritmes.

Després d'aplicar el mateix procediment que en el model anterior s'ha detectat un desequilibri entre les classes. Conseqüentment, s'ha optat per aplicar la tècnica SMOTE (Synthetic Minority Over-sampling Technique) al conjunt de dades

d'entrenament, amb l'objectiu de crear mostres sintètiques per a les classes minoritàries. Després de calcular els resultats amb les classes balancejades, la precisió global (accuracy) disminueix fins al 0,52. Davant d'aquesta disminució de la precisió, s'ha pres la decisió de revertir els resultats al seu estat inicial, sense l'aplicació de SMOTE, és a dir quan els valors de la F1-Score eren superiors.

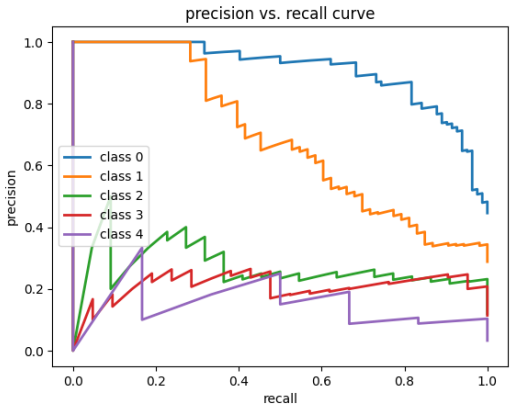

Figura 5

Les corbes PR (Figura 5) són irregulars i no segueixen un patró suau, ja que les classes 2, 3 i 4 tenen moltes menys instàncies que la classe 0 i 1. Això afecta la Precision i a la Recall, ja que el model té dificultats per aprendre les característiques de les classes menys representades.

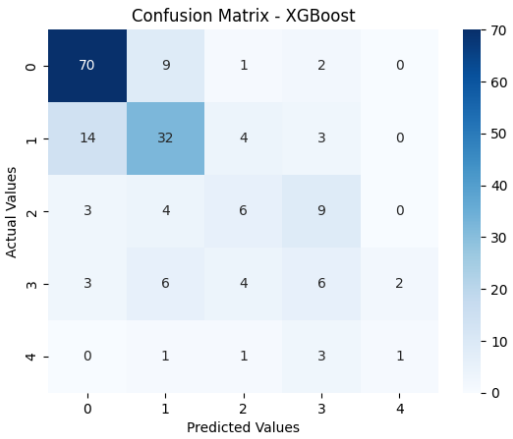

Figura 6

La matriu de confusió obtinguda per al model II (Figura 6), evidencia que les classes 1 i 2 tenen una presència més destacada que les altres classes, mentre que la classe 4 sembla ser la menys freqüent i també presenta dificultats a l'hora de classificar-la, amb una alta quantitat de falsos positius i falsos negatius.

El "classification report" del model proporciona una anàlisi detallada del seu rendiment en la classificació de diverses classes. La Precision destaca la capacitat del model per predir correctament les instàncies positives, mostrant variacions entre les classes. Per exemple, la classe 0 té una Precision elevada (0.78), mentre que la classe 2 i la classe 3 mostren precisions més baixes (0,38 i 0,26, respectivament). El Recall, que mesura la capacitat del model per identificar correctament les instàncies positives, presenta resultats més alts per a la classe 0 i més baixos per a la classe 2 i la classe 4. L'F1-score proporciona una visió equilibrada de Precision i Recall, i tot i que varia entre les classes, destaca el bon rendiment global del model. L'accuracy general del model és del 62%.

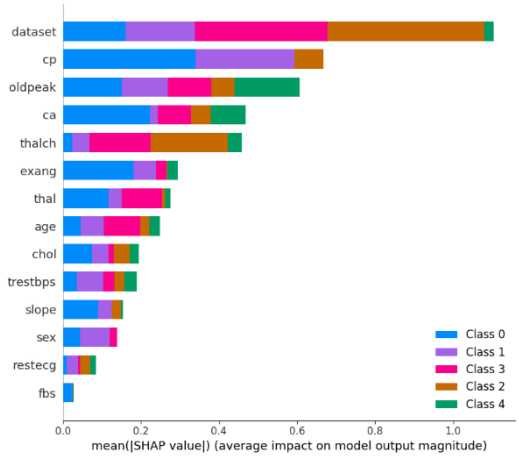

Figura 7

Com es pot observar a la Figura 7, el gràfic de SHAP ens mostra les característiques que més contribueixen en la predicció de cada classe en diferents colors.

En l'anàlisi dels resultats també s'ha utilitzat la llibreria LIME (Local Interpretable Model-agnostic Explanations) per tal de proporcionar una representació visual de la importància relativa de les característiques

per a la predicció de la instància seleccionada. Les característiques més destacades es mostren amb les seves contribucions, indicant com cada variable influeix en la predicció final del model.

4. Disseny de la Interfície

Per tal de permetre que els usuaris accedeixin al model de predicció de malalties cardíaques d'acord amb les dades que vulguin consultar, s'ha desenvolupat una interfície web.

Es defineixen dues pàgines principals: la pàgina d'usuari (pacient, model I) i la pàgina d'accés a professionals (metge, model II). Aquesta primera selecció condueix a l'usuari de l'aplicació a un formulari específic per a cada model. La interfície proporciona controls d'entrada com sliders, botons i caixes de selecció per permetre als usuaris introduir les seves dades de salut. Quan es fa clic al botó "Finish", les dades d'entrada es processen a través del pipeline i es realitza una predicció utilitzant els models pertinents.

5. Conclusions

En aquest estudi, introduïm el sistema de predicció de malalties cardíaques. Tot i que el conjunt de dades no és molt extens, ens ha permès crear dos models relativament senzills, un de binari i l'altre multiclasse. Es demostra que l'ús de diverses eines i tècniques d'explicabilitat de Machine Learning són vàlides en la predicció de malalties cardíaques.

Els resultats que obtenim d'aquest estudi de recerca mostren que l'algorisme Logistic Regression és l'algorisme amb major precisió en un model de classificació binari, amb una puntuació del 87% en la predicció de malalties del cor. Quan s'expandeix l'abast del model per predir entre més de dues classes (multiclasse), l'algorisme amb major precisió és XGBoost.

L'exactitud del model multiclasse ha resultat ser bastant baixa, això repercuteix en les prediccions del model. Les classes minoritàries són més difícils de predir a causa del baix nombre d'instàncies.

Les futures aplicacions de les eines de Machine Learning en medicina podrien revolucionar la predicció i diagnòstic de malalties en el món de la medicina.

6. Futures millores

El projecte es podria millorar implementant suggeriments de medicina al pacient juntament amb els resultats. Podem implementar un feedback dels metges experimentats que poden donar els seus punts de vista i opinions sobre determinats medicaments o proves realitzades pel metge sobre el pacient. Es podria implementar una opció de xat en directe on el pacient pot xatejar amb un metge disponible per tal de fer les preguntes necessàries sobre el resultat obtingut. A més, si s'implementa a gran escala es pot utilitzar en hospitals o clíniques on consultat l'aplicació web un pacient obtindria un diagnòstic inicial sense haver d'esperar en llargues cues si té símptomes relacionats amb malalties del cor.