MODEL CARD

Detalls del model

Autora

Zhiqian Zhou, zhiqian.zhou@estudiantat.upc.edu

Data i versió

data: 28 - 12 - 2023
version_name: svm_zqz_10086

Tipus de model

El model és un Support Vector Machine entrenat amb finalitat de predir la variable resposta demanada.

Informació

La base de dades original és proporcionada per la UCI (la Universitat de Califòrnia a Irvine) anomenada "Cirrhosis Patient Survival Prediction" amb la qual s'ha entrenat el model consta de 20 variables sobre característiques clíniques del pacient. El preprocessament està composta per l'eliminació de variables menys informatius, eliminació d'outliers, imputació amb knn i normalització per les variables numèriques i imputació simple afegint una modalitat i transformació a numèrica per les variables categòriques. Els hiperparàmetres del model són: 'C': 3, 'gamma': 'scale', 'kernel': 'linear'.

Citacions

Més informació a: Fleming, Thomas R. i David P. Harrington. Processos de recompte i anàlisi de supervivència. Vol. 625. John Wiley & Sons, 2013.

Ús previst

Ús principal

Amb una finalitat educativa, el present model està creat per a la pràctica de les tècniques i algoritmes apreses a les classes de IAA del departament de FIB de UPC.

Usuaris destinats

- Estudiants de IAA de la FIB de UPC
- Professors de IAA de la FIB de UPC

Ús fora de l'abast

El model creat no està pensat per ser utilitzat amb la finalitat que es proposa en cas real, és a dir, per realitzar prediccions de l'estat de supervivència dels pacients amb cirrosi hepàtica. Un ús no recomanat podria portar al risc vital del pacient.

Factors

Els factors que poden alterar la funcionalitat o el rendiment del model, serien possibles grups de pacients amb les característiques recol·lectades a les variables alterades per altres malalties. Cal una especificació per part de professional mèdics.

Mètriques

Les mètriques que s'han tingut en compte per modelar i avaluar el model són el F1 weighted i el Balanced accuracy. La raó d'aquesta tria és per resoldre els problemes que porta el desequilibri de les dades.
El threshold que s'ha considerat per eliminar les dades outliers és de 3, per tal de millorar el rendiment del model sense perdre grans quantitats de dades, ja que el dataset original en té poques.

Dades d'avaluació

Les dades emprats per l'avaluació del model han estat escollits de forma aleatòria dintre del data set i conté el 20% de les mostres. Dintre d'aquesta partició de test s'ha eliminat les variables que durant el preprocessament s'han considerat com no informatius. Ha sigut recodificat, ja que el dataset original contenia expressions per identificals els missing values que no es detectaven. També ha sigut normalitzat (les variables numèriques) i transformat a numèrica (les variables categòriques) abans de testejar-lo al model.

Dades d'entrenament

Les dades emprats per l'entrenament del model han estat escollits de forma aleatòria dintre del data set i conté el 80% de les mostres. Dintre hi conté les variables que especifiquen les característiques clíniques del pacient per poder predir la variable 'Status', una variable categòrica amb les modalitats D igual a mort, C igual a censurat i CL igual a censurat per trasplantament hepàtic. Les variables que conté són: ID, N_Dies, Estat, Droga, Edat, Sexe, Ascites, Ascites, Aranyes, Edema, Bilirubina, Colesterol, Albúmina, Coure, Alk_Phos, SGOT, Triglicèrids, Les plaquetes, Protrombina i Etapa.

Rendiment del model

A continuació són els resultats obtinguts sobre la partició de test:

- Balanced accuracy: 0.75
- F1 weighted: 0.84

Consideracions ètiques

Les dades que es conté en la base són informacions de persones reals, per tant, són dades sensibles i no s'ha d'utilitzar per males intensions.

Recomanacions

No utilitzar el model en cap cas sense que el professorat responsable l'hagi validat.