Data Analysis and Data Mining Project

Heart disease

Introduzione

Di tutte le applicazioni di **machine learning**, diagnosticare una qualsiasi malattia grave usando una **black box** sarà sempre difficile da vendere. Se l'output di un modello è il particolare corso del trattamento (potenzialmente con effetti collaterali), o la chirurgia, o l'assenza di trattamento, le persone vorranno sapere perché.

Questo set di dati fornisce un numero di variabili insieme a una condizione obiettivo di avere o non avere malattie cardiache. Di seguito, i dati vengono utilizzati per la prima volta in un modello di **random forest** semplice, quindi il modello viene esaminato utilizzando gli strumenti e le tecniche di spiegabilità di **machine learning**.

Vediamo nel dettaglio la struttura del data set:

* **age**: L'età della persona in anni
* **sex**: Il sesso della persona (1 = maschio, 0 = femmina)
* **cp:** Il dolore al petto sperimentato (Valore 1: angina tipica, Valore 2: angina atipica, Valore 3: dolore non anginoso, Valore 4: asintomatico)
* **trestbps:** Pressione sanguigna a riposo della persona (mm Hg al momento del ricovero in ospedale)
* **chol:** La misurazione del colesterolo della persona in mg / dl
* **fbs:** Glicemia a digiuno (> 120 mg / dl, 1 = vero; 0 = falso)
* **restecg:** Misurazione elettrocardiografica a riposo (0 = normale, 1 = con anormalità dell'onda ST-T, 2 = mostra ipertrofia ventricolare sinistra probabile o definita secondo i criteri di Estes)
* **thalach:** La frequenza cardiaca massima della persona raggiunta
* **exang:** Angina indotta dall'esercizio (1 = si; 0 = no)
* **oldpeak:** Depressione ST indotta dall'esercizio relativo al riposo
* **slope:** la pendenza del segmento di picco dell'esercizio ST (Valore 1: salita, Valore 2: piano, Valore 3: discesa)
* **ca:** Il numero di vasi principali (0-3)
* **thal:** Disturbo del sangue chiamato talassemia (3 = normale; 6 = difetto fisso; 7 = difetto reversibile)
* **target:** Cardiopatia (0 = no, 1 = si)

Obiettivo

L’obiettivo è di creare un modello per la predizione di malattie cardiache, utlizzando Random Forest Classifier il quale verrà configurato mediante la variazione degli Iper-parametri e ne verrà misurata la perfomance mediante la Cross Validation.

Procedimento

Analisi Dataset

Abbiamo utilizzato I tools della libreria sklearn per vedere:

1. Lo stato dei dati, cioè che i tipi di dati siano coerenti con il loro significato e rilevare la presenza di possibili valori nulli.
2. Sono stati fatti gli istogrammi di tutti i campi del dataset per valutarne la distribuzione.
3. Si trattano le categorical features (dati categorici) utilizzando le dummy features.
4. Guardiamo la Correlation Heatmap per vedere la correlazione tra le features

Tuning Iper-parametri