PRAC2: Neteja i Anàlisi de Dades

Membres

- Sergio Costa Planells
- Lorena Casanova Lozano

Descripció

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. S'ha triat un conjunt de dades extret de la plataforma Kaggle penjat en aquest repositori (indian_liver_patient.csv) i accessible a través de l'enllaç ("Indian Liver Patient Records | Kaggle," n.d.). El joc de dades conté els resultats d'anàlsis de mostres biològiques d'un total de 583 individus, 416 pacients amb malaltíes hepàtiques i 167 sense patologia, procedents del nordest d'Andhra Pradesh, India. En total inclou 10 atributs, tots numérics menys el sexe de l'individu i un 11è atribut classe binari que indica si l'individu presenta o no patologia de fetge.

Variables

1. Age: Edat del pacient

2. Sex: Sexe del pacient

3. Tot_bil: Nivell de bilirrubina total

4. Dir_bil: Nivell de bilirrubina directa

5. Alkphos: Nivell de fosfotasa alcalina

6. Alamine: Nivell d'aminotransferasa alanina

7. Aspartate: Nivell d'aminotransferasa aspartat

8. Tot_Prot: Nivell de proteïnes totals

9. Albumin: Nivell d'albunima

10. A_G_Ratio: Ratio albumina-globulina

11. Disesase: Estat de malaltia

Objectius

Atès que les malalties hepàtiques acostumen a ser greus l'interès del nostre objectiu rau en avaluar la capacitat que tenim de detectar aquestes patologies en pacients al més aviat possible. Plantegem la hipòtesis que els valors obtinguts a través de les anàlisis de sang de què disposem poden ser suficients per a detectar individus malalts malgrat encara no hagin desenvolupat una simptomatologia evident o, si més no, podem desenvolupar un model que ens permeti fer una primera selecció dels pacients candidats a anàlisis posteriors. Volem donar resposta, doncs, a les següents preguntes:

- Existeix una diferencia estadísticament significativa en la proporció de malalts de fetge en funció del sexe?
- Existeix una diferència estadísticament significativa en la mitjana d'edat entre els malalts i els individus sans?
- Quina és la influència dels diferents factors (atributs de què disposem) en la probabilitat de desenvolupar una patologia de fetge?
- Quina capacitat tenim de predir si un individu pateix o no una malaltia de fetge a partir dels atributs que disposem?

Metodologia

Per tal de donar resposta a aquestes preguntes s'apliquen diferents mètodes d'anàlisi estadístic sobre el conjunt de dades tot comprovant si es compleixen les condicions necessàries per a la validesa dels resultats. Atès que els objectius definits giren al voltant d'un problema de classificació i disposem d'una variable classe de control (Diesease) farem ús d'algoritmes d'aprenentatge supervisat per tal de construir el model final de classificació. Els diferents mètodes que s'utilitzaran per a realitzar l'anàlisi estadístic són els següents:

- Test d'hipòtesis -> Comparar mostres d'Homes/Dones i Malalts-Sans per donar resposta a les preguntes plantejades.
- Regressió logística -> Avaluar la influència de cada variable en la probabilitat de tenir afecció de fetge i predir el valor resultant.
- Random Forest -> Construir un model de classificació que ens permeti de predir la variable Disease a partir de la resta d'atributs. ("Cómo implementar bosques aleatorios en R | R-bloggers," n.d.; "RPubs - Árboles de decisión y métodos de ensemble," n.d.) ("Tune Machine Learning Algorithms in R (random forest case study)," n.d.)

Contribucions

Contribucions	Firma
Investigació Prèvia	SCP, LCL
Redacció de les respostes	SCP, LCL
Desenvolupament Codi	SCP, LCL

Resultats

Després de realitzar tot el procés d'anàlisi de les dades, disponible a ("Lorena96/PRAC2_Neteja-i-Analisi-de-Dades," n.d.) s'han pogut respondre a les preguntes inicialment plantejades.

- Existeix una diferencia estadísticament significativa en la proporció de malalts de fetge en funció del sexe?

  Hem realitzat un test d'hipòtesis sobre la diferència de proporcions entre la mostra de homes i dones i em pogut concloure amb un nivell de confiança del 95% que la proporció de malalts en homes és significativament superior a les dones.
  Si bé això no indica que sigui realment la distinció de gènere la que genera la diferència, sí que és un factor a tenir en compte per al diagnòstic.

- Existeix una diferència estadísticament significativa en la mitjana d'edat entre els malalts i els individus sans?

  Hem realitzat un test d'hipòtesis sobre la diferència de mitjanes d'edat entre la mostra d'individus sans i malalts i hem pogut concloure amb un nivell de confiança del 95% que la mitjana d'edat dels malalts és significativament superior a la dels individus sans el que indica certa rellevància del factor edat en el diagnòstic de patologies de fetge. Aquesta relació s'ha estudiat amb més deteniment durant la interpretació del model de regressió logística.

- Quina és la influència dels diferents factors (atributs de què disposem) en la probabilitat de desenvolupar una patologia de fetge?

  De la resta de variables de què disposem i en base a la mostra estudiada em obtingut uns odds ratio molt propers a la unitat, fet que indica que el seu efecte sobre la variable objectiu és minso, tot i ser factors de risc. Podem concloure que el conjunt de variables de que disposem es poc rellevant quan a la inferència de patologies de fetge fet que limita la capacitat de diagnòstic.

- Quina capacitat tenim de predir si un individu pateix o no una malaltia de fetge a partir dels atributs que disposem?

  Els resultats del model corroboren que la capacitat per a predir la patologia de fetge a partir de les dades disponibles és força limitada. Ajustant els paràmetres del model em aconseguit una precisió (Accuracy) del 71,84% vers el 71,26% que ens ofereix si seleccionem directament sempre l'individu com a malalt (No Information Rate). Comprovem que, tot i tenir una sensibilitat força alta (~88%) la sensibilitat del model és molt baixa (~30%) el que indica un taxa alta del falsos positius.

  Podem destacar dos factors principals que juguen un paper important en els resultats obtinguts a la predicció, d'una banda el joc de dades tractat és força desequilibrat amb molt més nombre de malalts i d'altra banda, com hem vist a partir del model logístic, les variables de què disposem no són especialment rellevants per a l'objectiu.

En resum, al llarg d'aquesta pràctica hem realitzar diferents anàlisis sobre el joc de dades seleccionat que ens han permès d'extreure'n coneixement i donar resposta les preguntes plantejades. Com a continuació del treball considerem que podria ser interessant l'aplicació de tècniques específiques per contrarestat el desequilibri present en les dades mitjançant un submostreig dels casos de malalts i una sobre-mostreig per repetició aleatòria dels casos d'individus sans amb l'objectiu d'equiparar ambdós classes i verificar-ne l'efecte sobre el model.

Cal especificar que en l'apartat del document resultant del codi que desenvolupa l'anàlisi de les dades (Prac2.pdf) hi han els resultats explicats amb els visuals extrets dels resultats dels models i les estadístiques aplicades.

Bibliografia

Cómo implementar bosques aleatorios en R | R-bloggers. (n.d.). Retrieved January 5, 2021, from https://www.r-bloggers.com/2018/01/how-to-implement-random-forests-in-r/

Indian Liver Patient Records | Kaggle. (n.d.). Retrieved January 5, 2021, from https://www.kaggle.com/uciml/indian-liver-patient-records

Lorena96/PRAC2_Neteja-i-Analisi-de-Dades. (n.d.). Retrieved January 5, 2021, from https://github.com/Lorena96/PRAC2_Neteja-i-Analisi-de-Dades

RPubs - Árboles de decisión y métodos de ensemble. (n.d.). Retrieved January 5, 2021, from https://rpubs.com/Cristina_Gil/arboles_ensemble

Tune Machine Learning Algorithms in R (random forest case study). (n.d.). Retrieved January 5, 2021, from https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/

Greenwell, B. (2019, October 19). Hands-On Machine Learning with R. Retrieved December 4, 2019, from https://bradleyboehmke.github.io/HOML/.

I. H. Witten, Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank, Mark A. Hall.Burlington, MA : Morgan Kaufmann Publishers/Elsevier, 2011.

J. Gironés Roig, Data mining / Jordi Gironès Roig ; [el encargo y la creación de este material docente han sido coordinados por los profesores: Jordi Conesa Caralt, David Masip Rodó].

Barcelona : Universitat Oberta de Catalunya, 2013.J. Gironès Roig, Minería de datos : modelos y algoritmos / Jordi Gironés Roig [i 3 més]. Barcelona : Editorial UOC,2017.

J. Long i P. Teetor, R Cookbook, 2nd Edition [recurs electrònic]; J. Long i P. Teetor, 2019.https://rc2e.com

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168755/

https://rpubs.com/bpoulin-CUNY/338004

https://www.kaggle.com/uciml/indian-liver-patient-records

http://www.statisticalassociates.com/logistic10.htm

https://event-mohs.gov.mm/wp-content/uploads/2019/12/Logistics-regression.pdf

https://www.statology.org/assumptions-of-logistic-regression/