Projecte final curs Data Science

Javier Serrat Sorolla

Setembre 2023

Introducció

1. Obtenció de dades
2. Descripció de les dades
3. Visualització de dades
4. Test d'hipòtesi
5. Regressió lineal
6. Models de classificació
7. Models d'agrupació

Conclusions

Introducció

En aquest treball fi de curs es pretèn posar en pràctica els coneixements i les tècniques apreses en el transcurs del curs Data science de l'ITAcademy mitjançant la seva aplicació pràctica a un conjunt de dades públiques.

Em plantejo dos objectius principals: per una part consolidar l'us de les tècniques i l'interpretació del resultats i per l'altra desenvolupar les habilitats necessaries que permeten l'aplicació d'aquestes eines en el mon real i en casos reals de negoci.

Entenc aquest segon objectiu com la constatació efectiva de l'aprofitament dels coneixements adquirits. Més enllà l'amplitud de coneixements i de la destressa en l'ús de tècniques i mètodes en l'àmbit del Data Science, aquests no són més que el mitjà per permetre'ns coneixer millor l'escenari en que desenvolupem la nostra activitat i oferir una pressa de decisions adequada als objectius marcats en el negoci.

Dividirem aquest document en varies parts. Per un costat les preceptives 'Introducció' i 'Conclusions' que com inici i final de la memòria delimiten el contingut del que s'ha treballat.

En l'apartat d'obtenció de dades comentarem quin ha estat l'origen de dades, una breu descripció sobre el conjunt de dades escollit i les preparacions bàsiques realitzades per afavorir el posterior tractament i anàlisis que en farem.

La part de preparació de dades farà un recorregut exhaustiu pels diferents atributs analitzant el seu contigut i distribució i fent , si s'escau, adequacions per evita influències no desitjades en els anàlisis posteriors.

En la visualització de dades representarem gràficament posibles relacions entre els atributs del conjunt de dades, cosa que ens permetrà aprofundir en el seu coneixement.

El capítol del test d'hipòtesis es basarà en l'aplicació d'alguns dels tests d'hipòtesi estudiats per treuren alguna conclusió sobre les dades que tenim.

La part de regressió lineal ens permetra establir alguns models entre atributs dels dataset que identificarem mitjançant la matriu de correlació i el mapa de calor que poden estar relacionades.

Els models de classificació que desenvoluparem ens serviran per catalogar mostres del valors dels atributs d'aquestes. Com a model d'arenentage supervisat, uns dels stributs el considerarem com a resultat per definir i entrenar a model de classificació.

Els models d'agrupació, com a models d'aprenentatge no supervisat, ens oferiran la possibilitat d'agrupar les instàncies de les nostres dades en funció de les seves característiques.

1. Obtenció de dades

Després de la cerca i selecció de un grapat de datasets que servís com base d'aquests treball m'he decidit per un dataset extret de www.kaggle.com. A quest conte quasi un milió de registres de persones amb la catalogació del seu gènere i els hàbits de fumar i beure així com

tot un seguit d'indicadors de salut i descriptius del seu fisic. La finalitat d'aquest dataset és l'anàlisi dels indicadors de salut i la classificació de grups de fumadors i bevedors

En aquesta primera part s'adequen els nom dels atributs per facilitar-ne la familiarització i es disposen el atributs categòrcs tant en format numèric com descriptiu, ja que algunes de les tècniques que aplicarem precisen que aquest suguin de format numéric, pero al mateix temps la forma descriptiva ens facilita la interpretació de resultats.

També es prendran les decisions sobre que fer amb els valors nuls i duplicats.

2. Descripció de les dades

En aquest apartat es recorreran cadascun dels atributs del dataset. La finalitat es identificar el rang de valors i la resta de descriptors estadístics, així com identificar l'existència d'outliers i decidir que fer amb ells en funció de l'influència que estimem poden tenir en quan a distorsions en l'interpertació dels resultats. Per aquesta tasca ens recolzarem en visualizacions simples com boxplot i histogrames básicament.

3. Visualitzacio de dades

La visualització de dades es basara en gràfics multiatributs i ens complementarà l'aprenentage individual dels atributs amb les interrrelacions que siguem capaços de identificar. Utilitzarem mapes de calor, diagrames de barres, histogrames, boxplots, grafics de línies i de violí.

4. Test d'hipòtesi

En aquest aparat utlitzarem test d'hipotesi per veure si podem "demostrar" amb un cert grau de confiança si algunes deduccions, sospites i sugerències (en definitva hipotesis)s que hem fet ara quan hem conegut en profunditat les dades les podem cosnsiderar certes. Aplicarem test de shapiro, d'Agostini i de t de Student en funció de l'escenari del que volguem constatar.

Comprovarem si la distribució d'homes i dones es igual en tota la mostra, i si tambe ho és per edats. Tot aixo motivat per les percepcions que ens mostren les visualitzacions gràfiques del punt anterior.

5. Regressió lineal

La regressió lineal ens proporcionarà alguns models d'aprenentatge supervisat que desenvoluparem per veure de quina maera una variable ens explica el comportament d'una altra. Veurem casos de regresió lineal simple, multiple y polinomial.

Estudiarem una potencial relació o dependencia si més no curiosa entre l'alcada i els nivells d'hemoglobina i altress més lògiques com la relación entre els nivells dels diferents tipus de colesterol. Tamnbe estudirem el casa amb un model de Random Forest Regression i compararem tots els models en base a R -square i el MSE. Aprofitarem el model Random Forest Regression per veure quin es el pes de cada atribut en la relació amb el resultat i quin resulta més determinant.

6. Models de classificació

Seguim amb les models d'aprenetage supervisat, aplicarem models de Decission Tree, KNN, SVM i Logistic regression per intentar classificar els elemets del nostre dataset segons la columna de resultat que hem decidit.

És potser el capitol més extens. Farem una transformació a les dades del tipus MinMaxScaler per tenir-les totes en la mateixa escala y evitar influències no desitjades que distorsionin els resultats.

Intentarem determinar en quin mesura els nivells de transaminases poden determinar si una persona és bevedora o no.

Tots els models es creen, es desenvolupen amb un grup de dades d'entrenament i es comprovem sobre un conjunt de dades de prova. A mes a mes també s'aplicaa un mecanisme de validació creuada i es comparen els resultats del diferents models per triar el que proporcioni millors resultats. Aquest comparacio es fa en base a la mattriu de confusió, l'accuracy, la precision, el recall i el f1.

Emprarem mètodes per trobar el valor òptim de k i tambe utlitzarem metoies avançats d'avaluació del rendiment com la corva ROC i l'area AUC

Tot es fa acompanyat dels resultats dels cálculs i de la representació gràfica quan és escaient.

8. Models d'agrupació

En aquest apart dedicat al'aprenetage no supervisat desenvoluparem els metodes K-means i Aglomerative Clustering per fer una agrupació de les instàncies en clusters de manera que les que es trobin en el mateix cluster estiguin entre elles més a prop del que estarien amb instàncies d'altres clusters.

Utilitzarem l'análisi principal de components per veuren's la seva bondat i com reduir els atributs a considerar a l'hora de fer una agrupació

A més de la definició, entrenament i prova del model, veurem diferents maneres de calcular el valor òptim de clusters com pot ser el mètode del colze, el dendograma i el métode de Silhouette.

Conclusions

A l'acabament d'aquest treball queda la constancia del llarg camí recorregut amb tot un seguit de idees, tècniques, mètodes, trucs i estrategies que s'acumulen al sarrò pèro amb una visió a l'encara més llarg camí que queda per recòrrer. Si mes no, i no menys important que allò que ens emportem pòsat i que queda en la ment, és tot el conjunt de recursos als que ara sóc capaç d'accedir i que sens dubte facilitaran i enriquiran els projectes a futur.