Visualització de dades

Pràctica (primera part)

El conjunt de dades que he seleccionat s'anomena **"Indicadors personals clau de la malaltia cardíaca"**, el nom està traduït de l'anglès, no es tracta d'un conjunt de dades mundial, sinó que està basat en un estudi als Estats Units d'Amèrica del Nord. La llicència del conjunt de dades és CC0: Public Domain i la freqüència d'actualització prevista és anual.

Respostes:

1. | Justifiqueu breument la vostra selecció, sigui per motius personals o professionals.

La meva selecció està basada en la comprensió de les dades, ja que entenc perfectament que s'està estudiant i perquè. Per tant,

- Es tracta d'una selecció per motius personals, ja que ara mateix no treballo ni en el camp de la medicina, ni en l'àmbit de la ciència de dades.
- El conjunt de dades està format per variables lògiques o booleanes, de text i numèriques, fet que em permetrà fer una visualització correcta des del punt de vista avaluatiu de la segona part de la pràctica, i de qualitat visual i comprensible per al "lector" de la visualització.
- Per últim, té un nombre de variables, 18, suficients per estudiar-lo, i milers de registres verificats, ja que es tracta d'un conjunt de dades netejat per l'autor, KAMIL PITLAK.

2. | La rellevància del conjunt de dades en llur context. Són dades actuals? Tracten un tema important per algun col·lectiu concret? S'ha tingut en compte la perspectiva de gènere?

Les dades provenen de l'enquesta anual dels CDC (Centers for Disease Control and Prevention) i és de l'any de 2020. La responen uns 400.000 adults i està relacionada amb el seu estat de salut general, tot i que l'objectiu principal de l'enquesta és arribar a detectar els indicadors clau de les malalties del cor.

La rellevància del conjunt de dades en el context tant actual, com en el del 2020 és alta, no únicament pels Estats Units d'Amèrica, m'ha semblat que és un conjunt de dades extrapolable al continent Europeu i també i en general, a occident. Segons els CDC, les malalties del cor és una de les principals causes de mort per a persones de la majoria de races als EUA (afroamericans, indis americans i nadius d'Alaska i blancs), així doncs, gairebé la meitat de tots els nord-americans (47%) tenen almenys un dels tres factors de risc clau per a malalties del cor: pressió arterial alta, colesterol alt i tabaquisme. Aquesta conclusió em va fer pensar en que la majoria de gent del meu voltant o que conec per referències de tercers, tenen els mateixos factors de risc.

D'altres indicadors clau que inclouen els CDC són l'estat diabètic, l'obesitat (IMC, o índex de massa corporal, alt), no fer prou activitat física o beure massa alcohol, fet que em fa reafirmar en que al menys a l'estat espanyol, fa molt de temps que s'està posant el focus d'atenció i s'està alertant sobre aquests factors, segurament deguts a la alimentació incorrecta, a la cultura de l'alcohol que tenim a aquest país, al sedentarisme o pel contrari a l'obsessió per una impecable estètica de tonificació muscular, a la rapidesa en la que volem viure, a la falta de descans i temps per a gaudir d'àpats de qualitat, etc.

Així doncs, detectar i prevenir els factors que tenen un major impacte en les malalties del cor és molt important, no només pels humans, sinó per l'assistència sanitària de la que gaudim, gratuïta i fins ara, gairebé excel·lent.

Es tracta d'un conjunt de dades prou actualitzat, tenint en compte que portem dos anys capficats amb les dades de la pandèmia de la COVID. Pel que he llegit la última actualització es de fa dos mesos (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease)

El col·lectiu interessat en aquest conjunt de dades som gairebé tota la humanitat, ja que estudia els factors decisius que poden provocar una malaltia cardíaca.

En el conjunt de dades es té en compte si la persona entrevistada és un home o una dona, així com alguns aspectes com l'embaràs o la dosis d'alcohol que es determina que és excessiva per un home o una dona, tot i així, no crec que s'hagi tingut en compte del tot la perspectiva de gènere, ja que oblida el col·lectiu que no es considera binari i que la quantitat d'alcohol excessiva si ets home o dona no està determinada pel sexe, sinó per d'altres factors com el IMC, estat de salut general, etc.

3.| La complexitat (mida, variables disponibles, tipus de dades, etc.). Té de l'ordre de centenars o milers de registres? Té de l'ordre de desenes de variables? Combina dades categòriques i quantitatives? Inclou altres tipus de dades?

En aquest moment, el conjunt de dades té 319.796 files i 18 variables disponibles (9 booleanes, 5 cadenes de text i 4 numèriques), combinant tres tipus de dades, booleanes, categòriques i numèriques. Com que es tracta del resultat d'una enquesta, la mostra està basada en respostes a preguntes molt concretes. Les variables són les següents:

1. **HeartDiseasesort (variable booleana)**

   En aquesta variables, els/les enquestats/des informen de si han sofert, en algun moment de la seva vida, una malaltia coronària (CHD) o infart de miocardi (MI)

2. **BMIsort (variable numèrica)**

   Índex de massa corporal (BMI) de l'enquestat o enquestada.

3. **Smokingsort (variable booleana)**

   La pregunta és si la persona ha fumat almenys 100 cigarrets en tota la teva vida? **Nota:** 5 paquets = 100 cigarrets

4. **AlcoholDrinking (variable booleana)**

   És bevedor o bevedora l'enquestat/da? En el conjunt de dades es consideren bevedors/es intensos/ses els homes adults que prenen més de 14 begudes alcohòliques per setmana i dones adultes que prenen més de 7 begudes alcohòliques per setmana.

5. **Stroke (variable booleana)**

   Aquesta variable ens indica si l'enquestat/da ha sofert un ictus en algun moment de la seva vida.

6. **PhysicalHealth (variable numèrica)**

   Aquesta variable inclou si la persona pateix alguna malaltia física i/o lesions. La pregunta és si en els últims 30 dies, quants d'aquests dies la salut física de l'enquestat/da no va ser bona.

7. **MentalHealth (variable numèrica)**

   Aquesta variable inclou si la persona pateix alguna malaltia mental. La pregunta és si en els últims 30 dies, quants d'aquests dies la salut mental de l'enquestat/da no va ser bona.

8. **DiffWalking (variable booleana)**

   Dificultat greu per caminar o pujar escales.

9. **Sexe (variable categòrica)**

   Home o dona (Male o Female)

10. **AgeCategory (variable categòrica)**

    Categoria d'edat dividida en catorze nivells.

11. **Race (variable categòrica)**

    El que es considera la raça de l'enquestat o enquestada (6 nivells)

12. **Diabetic (variable categòrica)**

    Si l'enquestat o enquestada és diabètic o ho ha sigut durant un període determinat de la seva vida (embaràs), o gairebé ho és perquè està al límit dels indicadors de diabetis.

13. **PhysicalActivity (variable booleana)**

    Té l'enquestat o enquestada una activitat física? Fa esport del tipus que sigui? Caminar, córrer, etc.

14. **GenHealth (variable categòrica)**

    Estat general de la salut de l'enquestat o enquestada (cinc nivells)

15. **SleepTime (variable numèrica)**

    Hores de son diàries de l'enquestat o enquestada.

16. **Asthma (variable booleana)**

    Té l'enquestat o enquestada, asma?

17. **KidneyDisease (variable booleana)**

    L'enquestat o enquestada pateix d'una malaltia al ronyó?

18. **SkinCancer (variable booleana)**

    L'enquestat o enquestada té càncer de pell?

4.| L'originalitat. Combinar o millorar visualitzacions existent. Hi ha altres visualitzacions basades en aquest conjunt de dades? És una evolució o actualització d'un conjunt anterior? Heu enriquit un conjunt de dades ja existent?

Aspectes importants sobre el conjunt de dades seleccionat:

El conjunt de dades original de prop de 300 variables i es va reduir a només unes 20. Com ja he dit abans, el conjunt de dades prové del CDC i aquest és una part important del sistema de vigilància de factors de risc conductuals (BRFSS Behavioral Risk Factor Surveillance System), que realitza enquestes telefòniques anuals per recopilar dades sobre l'estat de salut dels habitants dels Estats Units d'Amèrica. El BRFSS es va establir l'any 1984 en 15 estats i en actualment, recull dades als 50 estats dels EUA, així com al districte de Columbia i dos territoris més dels Estats Units, com són Guam i Puerto Rico.

Així doncs, BRFSS realitza més de 400.000 entrevistes a adults cada any, la qual cosa la converteix en l'enquesta sobre salut, realitzada continuadament en el temps, més gran del món.

La font original de les dades, ja he dit abans, que l'autor les ha tractat per millorar-les, és https://www.cdc.gov/brfss/annual_data/annual_2020.html, els centres de control i prevenció de malalties que conté dades i documentació de l'enquesta BRFSS de l'any 2020.

Metodologia aplicada al conjunt de dades original:

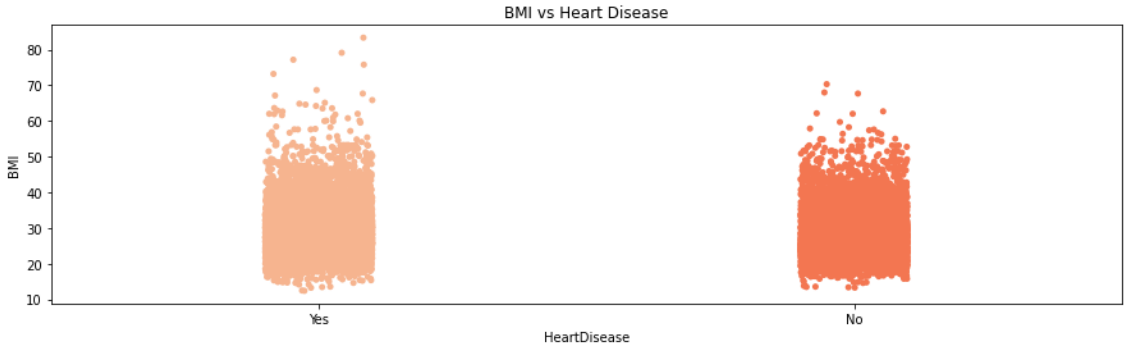
Segons l'autor, l'organització i la neteja del conjunt de dades es van aplicar mitjançant pandes i polars a Python. Aquesta va ser la transformació i neteja de les dades originals:

- Primer, el conjunt de dades original es va convertir de format SAS a CSV.
- Després es van seleccionar les variables amb un efecte directe o indirecte sobre les malalties del cor.
- Els valors de les variables categòriques es van convertir de tipus numèric a tipus de text per facilitar-ne l'anàlisi.
- I per últim es van eliminar les files amb registres buits.

El que ha fet que hem decidís per aquest conjunt de dades és que en hi ha de més completes dins dels CDC, dins de Kaggle[1], d'altres anys, d'altres organismes, però sobretot hi ha estudis sobre predicció de dades en python i també visualitzacions que tenen a veure en la prevenció de les malalties cardíaques. Si bé aquestes visualitzacions no són sofisticades, ni animades, ni tan sols interactives, m'ha semblant que sí que es podria explorar i realitzar alguna infografia o visualització de dades impactant, per a la prevenció de les malalties cardíaques.

Més aviat, el que m'ha semblat que puc fer és prevenció a través de la visualització de dades, ja que evitant comportaments que tenim molt interioritzats en la nostra vida diària, es poden millorar els ratis de patiment, d'hospitalització i en definitiva de desenvolupar una malaltia del cor, que finalment no només pot condicionar la nostra vida i dels que tenim al voltant i que ens estimen, sinó evitar una mort sobtada i prematura.

Per exemple, un dels gràfics que he trobat interessant és el següent:



Autor: MATVEI ALEKSANDROVICH. "The BMI vs Heart Disease". Estret de l'estudi en phyton de la malaltia cardíaca: visualització i classificació. Publicat en Kaggle: https://www.kaggle.com/datasets, L'autor comprova que l'índex de massa corporal (IMC), tot i que pugui semblar que té un gran impacte en la salut i sobretot en el cor, no sembla que hi hagi cap correlació forta. Disponible a Internet a: https://www.kaggle.com/code/ricksan4ez/heartdisease-visualisation-classification

[1] Kaggle és una empresa que s'encarrega, entre d'altres activitats, a organitzar concursos per a l'anàlisi de grans quantitats de dades i aconseguir amb els resultats fer prediccions. Es tracta d'una web oberta on port participar qualsevol i tota la informació recopilada està a l'abast de tothom. És tracta d'un projecte obert, participatiu i, en general, solidari, en quan que cobreix, en molts casos, la falta de científics de dades i de dades per fer anàlisis científics.

5.| Les qüestions que respondreu amb la visualització de dades, tenen en compte els punts anteriors? Estan ben plantejades? Són adequades pel conjunt de dades triat?

El que m'agradaria fer en aquest conjunt de dades és difusió sobre la prevenció de malalties cardíaques, responent a preguntes sobre si són realment factors de risc real o exagerats, algunes de les conductes i actituds que formen part de la nostra vida quotidiana. També m'agradaria ampliar aquesta informació i fer-la extensible a Europa i si és possible a la resta de continents. Seria fantàstic trobar més indicadors de risc i revisar d'altres estudis per poder afegir més informació al conjunt de dades.

Una de les preguntes que voldria poder contestar és: quines variables tenen un efecte significatiu en la probabilitat de patir malalties del cor?

També voldria esbrinar si el fet de tenir o haver nascut en una malaltia de tipus cardíac és un factor de risc ineludible que condicionarà de facto com s'acabarà aquesta vida, tenint en compte que l'autor del conjunt de dades ens avisa de que la variable **HeartDiseasesort**, tot i que per millorar l'estudi, s'ha convertit en booleana, les classes no estan equilibrades, de manera que l'enfocament clàssic d'anàlisis exploratòria de dades no és el més adient.

A més de l'EDA clàssic, tot i que no estigui recomanat, aquest conjunt de dades es pot utilitzar per aplicar mètodes d'aprenentatge automàtic, sobretot models de classificació (regressió logística, SVM[2], random forest[3], etc.)

2 SVM o support-vector Machines, és un conjunt d'algorismes que són capaços d'analitzar dades i reconèixer patrons mitjançant l'ús de mètodes d'aprenentatge supervisat.

3 Random forest és una combinació d'arbres predictors on cada arbre depèn dels valors d'un vector aleatori provat independentment i amb la mateixa distribució per a cadascun d'aquests.

Bibliografia i referències digitals:

- Personal Key Indicators of Heart Disease, 2020. Annual CDC survey data of 400k adults related to their health status. Disponible a Internet a: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
- U.S. Department of Health & Human Services. National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health, 2021. Disponible a Internet a: https://www.cdc.gov/brfss/annual_data/annual_2020.html