UNIVERSITAT DE BARCELONA

Treball final de grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Detecting discrimination through Suppes-Bayes Causal Network

Autor: Blai Ras Jimenez

Director: Oriol Pujol Vila
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 19 de Gener de 2019

Abstract - Català

Avui en dia confiem en models matemàtics o algoritmes per a realitzar decisions tan importants com qui obté la llibertat condicional o quant ha de pagar una família pel seu benestar. Usem aquestes pràctiques o *Machine Learning* pensant que eliminaran el biaix humà, però oblidem que els creadors d'aquests algoritmes som nosaltres, i, per tant, actuen com a miralls. En conseqüència, confiar cegament amb un model de *Machine Learning* per a decisions importants és un problema.

En aquest treball s'intenta tractar aquest problema basant-nos en l'article *"Exposing the probabilistic causal structure of discrimination"* [1], on es descriu una tècnica per al descobriment de discriminació basada en la inferència bayesiana i la teoria de Patrick Suppes de la causalitat probabilística.

A partir d'aquí s'ha adaptat i millorat el seu algoritme, el qual ens permet extreure models gràfics probabilístics que després de patir un entrenament per *likelihood-fit* es transformen en *Suppes-Bayes Causal Networks* (SBCN). La nostra SBCN generada és explotada amb algoritmes estocàstics de recorregut de grafs que ens permeten classificar i valorar la gravetat de la discriminació en cas de ser-hi present.

Finalment, s'ha programat i dissenyat una aplicació web que permet expressar els resultats de l'algoritme d'una manera visual i intuïtiva. Aquesta pàgina web es troba de manera totalment *open-source* en un núvol d'*AWS* pel seu ús i desenvolupament de futures extensions.

Experiments amb dades reals sobre judicis i demandes realitzats amb aquest mètode i en aquesta plataforma demostren el poder de l'enfocament causal pel descobriment de la discriminació.

Abstract - Castellano

Hoy en día confiamos decisiones tan importantes como quién obtiene la condicional o cuánto debe pagar cada miembro de una familia por su seguro médico a modelos matemáticos o algoritmos. Usamos estas prácticas o *Machine Learning* pensando que eliminará el prejuicio humano, pero olvidamos que los creadores de estos algoritmos somos nosotros mismos, y, por lo tanto, actúan como espejos.

En este trabajo se intenta tratar este problema basándonos en el artículo "*Exposing the probabilistic causal structure of discrimination*" [1], donde se describe una técnica para el reconocimiento de discriminación basada en la inferencia bayesiana y la teoría de Patrik Suppes sobre la causalidad probabilística.

A partir de ahí se ha adaptado y mejorado su algoritmo, el cual nos permite extraer modelos gráficos probabilísticos que, despúes de sufrir un entrenamiento por *likelihood-fit*, se transforman en *Suppes-Bayes Causal Networks* (SBCN). Nuestra SBCN generada es explotada con algoritmos estocásticos de recorrido de grafos que nos permiten clasificar y valorar la gravedad de la discriminación en caso de estar presente.

Finalmente, se ha programado y diseñado una aplicación web que permite expressar los resultados del algoritmo de manera visual e intuitiva. Esta página web se encuentra de manera totalmente *open-source* en una nube de AWS para su uso y desarrollo de futuras extensiones.

Experimentos con datos reales sobre juicios y demandas realizadas con este algoritmo y en esta plataforma demuestran el poder del enfoque causal para el reconocimiento de la discriminación.

Abstract - English

Machine learning or Artificial intelligence is now making some critical life-changing decisions, such as who gets the parole or what is the insurance costs for each individual of a family. We use these Machine Learning techniques thinking that they remove the human bias, forgetting that they just act like our mirror as they only learn based on what data we give them. Hence, blindly trusting a machine learning model for critical decisions is a big problem. This problem is becoming an important research area in Machine Learning now.

In this project, we try to handle this problem by using the techniques given in the paper *"Exposing the probabilistic causal structure of discrimination"* [1]. This paper describes a discrimination detection technique based on Bayesian inference and Suppes' probabilistic causation theory.

Starting from this point, we have adapted and upgraded his algorithm. That allows us to extract graphical probabilistic models that after being trained by *likelihood-fit* they transform into a *Suppes-Bayes Causal Networks* (SBCN). This generated SBCN can be exploited with random walk methods that provide us information about the presence of discrimination and its level, if any.

Finally, we have also designed a user-friendly web interface that extracts the information about this algorithm and shows it visually and intuitively. We deployed the application in AWS cloud for use and open-sourced the code base for future extensions. Multiple experiments with real-world datasets done with this algorithm and in this platform show the power of the causal approach of discrimination analysis.

Agraïments

En primer lloc, mostrar la meva gratitud cap a l'Oriol Pujol, tutor d'aquest treball, per posar-me en contacte amb Eurecat i fer possible aquesta feina. Gràcies per trobar el temps i dedicació per guiar i corregir-me en el procés.

En segon lloc, donar les gràcies a Eurecat, al seu departament de *Big Data & Data Science*, en concret a en Francesco Bonchi i Rohit Kumar, els quals sempre s'han mostrat disposats a ajudar-me.

Finalment, agrair la paciència de mon pare amb tots els dubtes matemàtics i els consells de distribució de l'espai i color de la dissenyadora Aranzazu Mozos.

Índex

1 Introducció **10**

1.1 Origen del projecte 10

1.2 Què entenem per discriminació? 11

1.3 Tipus de discriminació 11

1.4 Relacions espúries, correlació i causalitat 12

1.5 Proposta 13

1.6 Full de ruta 14

1.7 Objectius 15

2 Conceptes previs i estat de l'art **16**

2.1 Teoria causal de Suppes 16

2.1.1 Definició 16

2.2 "Altres" mètodes 18

2.3 Mesures de la discriminació 19

2.3.1 Puntuació de la discriminació per *random walks* 19

2.3.2 Mitjana de passos 20

2.3.3 Puntuació inconclusiva 21

2.3.4 Variable neutra 22

2.3.5 Discriminació explicable o condicionada 22

2.3.6 Puntuació de la discriminació amb *Page Rank* 23

3 Suppes-Bayes Causal Network **26**

3.1 Prioritat Temporal 26

3.2 *Probability Rising* 27

3.3 Aprenentatge de la SBCN 27

3.3.1 *Bayesian Information Criteria* 27

3.3.2 *Hill Climbing* 29

3.3.3 Pes de les arestes 32

4 Implementació **33**

4.1 Arquitectura 33
4.2 *Pipeline* de l'algoritme 33
4.3 Entrada i neteja de dades 34
4.4 Càlcul de probabilitats 35
4.5 Compliment de les condicions de la teoria de Suppes 37
4.5.1 Comprovació de la prioritat temporal 37
4.5.2 Comprovació de la *probability rising* 37
4.6 Entrenament i estimació de la xarxa 38
4.7 Reconstrucció de la SBCN 39
4.8 Puntuació de la discriminació 40
4.8.1 Puntuació de la discriminació per *random walk* 41
4.8.2 Puntuació de la discriminació per *Page Rank* 44
4.9 Estructura de l'aplicació web 45
4.9.1 *Deploy* 48
4.9.2 Política de privacitat de les dades 48

5 Experiments, Resultats i Discussió **49**
5.1 Datasets testejats 49
5.1.1 *COMPAS Scores* 57
5.1.2 *Huston College of Medicine Dataset* 63
5.2 Rendiment 68

6 Conclusions **70**

7 Futures Millores **71**

A *Appèndix*: Captures de la pàgina web **75**

Índex de figures

1 Exemple d'un petit graf resultant de despeses mèdiques als Estats Units 14
2 Arbre de decisió a seguir a l'hora d'establir un veredicte 22
3 Exemple de graf on aplicar Page Rank 23
4 Accions possibles en l'algoritme *Hill Climbing* [2] 30
5 Figura mostrant la mitjana de puntuacions positiva i negativa per cada fitxer, amb número d'iteracions 10, 100, 500 i 1000 51
6 DAG resultant del dataset *Berkeley Admission Data* 52
7 Puntuacions de *Page Rank* negatives pel subgrup de raça negre en el dataset *Adult* 53
8 Presència absoluta de classificació inconclusiva en el dataset *German* . . 54
9 En vermell, puntuacions de *Page Rank* positives pel subgrup de clients nous en el dataset *German* 54
10 A l'esquerra, puntuacions positives i a la dreta, puntuacions negatives, diferenciades per gènere. Dataset *Medical Costs US* 55
11 Les persones de més edat tenen menys puntuació positiva i més puntuació negativa. Dataset *Medical Costs US* 56
12 Boxplots del subgrup Overtime. Per ambdós tipus de puntuació, no observem diferència de valors. Dataset *IBM Employee Attrition* 57
13 DAG resultant de la SBCN pel dataset *COMPAS Scores* 58
14 Classificació de les variables del dataset *COMPAS Scores* 59
15 A la dreta, puntuacions negatives més grans per a la gent amb molt antecedents. A l'esquerra, puntuacions positives més baixes pel mateix subgrup. Dataset *COMPAS Scores* 60
16 En vermell, individus amb nombre alt d'antecedents. Puntuacions negatives més altes i positives més baixes. Dataset *COMPAS Scores* 61
17 Similitud de puntuacions tan positives com negatives per als individus de raça caucàsica i els que no. Dataset *COMPAS Scores* 62
18 Part del DAG resultant del dataset *Huston College of Medicine*. No s'observa cap aresta entre *publication rate low* i *salary increase high* 65

19 Part del DAG resultant del dataset *Huston College of Medicine*. Observem l'existencia d'una aresta entre *publication rate high* i *salary increase low* . . 66
20 Lleugera diferència de puntuacions positives entre els individus amb un alt nombre d'articles publicats i els indivius que no. Dataset *Huston College of Medicine* 67
21 Inexistència de discriminació per gènere. Dataset *Huston College of Medicine* 68
22 Temps d'execució (segons) en relació a la mida del fitxer (MB) 69
23 Càrrega d'un fitxer a la web 76
24 Llistat de fitxers 77
25 Primers resultats del anàlisi 78
26 Plana dels resultats de Page Rank 79

Índex de taules

1 Taula de resultats del dataset *Adult*, ordenada per puntuació negativa . . 52
2 Taula de resultats del dataset COMPAS Scores, ordenada per puntuació negativa 59
3 Taula de resultats del dataset Huston College of Medicine, ordenada per puntuació negativa 64

1 Introducció

El terme Big Data avui en dia està força arrelat a la nostra societat. El seu primer ús data del 1990 [3], però no es va popularitzar fins a principis de 2013. Tot i que no tothom coneix el seu verdader significat, ningú ho associa a quelcom negatiu. Malauradament, hi ha hagut males pràctiques del Big Data, i en força d'elles, ningú era coneixedor del seu mal ús.

Aquesta desconeguda part del Big Data, anomenada també "The Dark Side of Big Data" [4], fa referència a l'ús de models que determinen l'accés a un crèdit bancari, a quant paguem per una assegurança mèdica o quin tipus d'anuncis se'ns mostren al navegar per internet. Aquests algoritmes de decisió estan sent usats avui en dia de manera totalment opaca i sense cap tipus de regulació, afavorint així qualsevol tipus de discriminació que puguin causar.

1.1 Origen del projecte

Aquest treball de fi de grau és una col·laboració entre la Facultat de Matemàtiques i Informàtica i Eurecat Centre Tecnològic de Catalunya. Dins de l'àrea de *Big Data & Data Science* del centre, es vol desenvolupar una eina pública on l'usuari pugui pujar la seva base de dades i obtingui un informe sobre la presència de qualsevol tipus de discriminació en ella.

L'idea original sorgeix de dos dels experts d'aquests centre, en Francesco Bonchi i en Rohit Kumar. L'algoritme en el que es basa aquest treball està inspirat en el projecte *Exposing the probabilistic causal structure of discrimination* [1], d'en mateix Francesco Bonchi i tres col·laboradors més.

1.2 Què entenem per discriminació?

Quan parlem de discriminació ens referim a qualsevol distinció, exclusió, restricció o preferència a una persona o grup pel seu color de pell, sexe, orientació sexual, edat, religió, classe social, nacionalitat, etc. Amb altres paraules, podríem dir que és l'acte de "tractar de diferent manera a algú en comparació a com tractaries a la resta de persones" [5] .

1.3 Tipus de discriminació

Tot i que la definició principal de discriminació queda clara, hi ha una sèrie d'actes o decisions discriminatòries que sovint se'ns passen per alt. En conseqüència, a continuació veurem totes les maneres en les que podem dividir el concepte de discriminació i cobrir totes les formes en les que es pot manifestar.

- **Discriminació directa:** aquella que es manifesta directament per qualsevol característica del grup o individu, com per exemple, excloure d'una beca a les persones negres o limitar una posició vacant a persones heterosexuals.

- **Discriminació indirecta:** aquella que succeeix quan una norma o política és aplicada a tothom però afecta de pitjor manera a una persona o col·lectiu. Imaginem que a una feina hi ha la directiva de treballar mínim 2 dissabtes al mes. Els empleats jueus, els quals la seva religió dicta que el dissabte és un dia especial, estarien sent discriminats.

- **Favoritisme:** es dona quan una persona es tractada millor que d'altres sense una justificació aparent, quan es posa per sobre "el favor sobre el mèrit" [6].

- **Discriminació explicable o condicionada:** és aquella discriminació que pot ser justificada per un o més atributs relacionats amb la persona o grup. Donat un col·lectiu determinat que aplica per a una feina, resulta que la gran majoria de persones de sexe masculí tenen un nivell baix d'anglès. En conseqüència, el percentatge d'homes que aconsegueixen la feina és baix. No podrem dir que l'empresa discrimina pel gènere, sinó que ho fa pel nivell d'anglès [1].

1.4 Relacions espúries, correlació i causalitat

Hi ha un blog força popular d'un estudiant de Harvard anomenat Tyler Vigen que recull una sèrie de correlacions altes entre dues variables completament diferents [7]. Segons aquest blog, la quantitat de diners que es gasta els Estats Units en ciència i tecnologia és directament proporcional al número de suïcidis per sufocació i estrangulació. Vol dir això que com més diners es gastin els Estats Units més suïcidis hi haurà a la població? No, aquest tipus de comparació és el que anomenen una relació espúria: dos esdeveniments sense connexió lògica aparent poden ser causa-efecte degut a un tercer factor encara no considerat, anomenat també "variable amagada".

Un exemple senzill podria ser el cas de la mida de la sabata dels alumnes d'una escola de primària amb la seva facilitat per llegir. Podríem afirmar que com més gran és la sabata, millor llegeixen. En veritat, aquí hi ha una variable amagada: l'edat. A mesura que compleixen anys, més els hi creix el peu, i, a la vegada, més experiència lectora van adquirint.

Aquestes relacions espúries ens obren un nou problema en la detecció de discriminació. Per poder evitar aquests tipus de fals positius i ser capaços de detectar qualsevol dels tipus de discriminació anteriorment citats, es proposa una mètode basat en la teoria de la causalitat probabilística de Patrick Suppes [8].

[1] la discriminació condicionada o explicable no té per què ser vista 'de bona manera': pot ser igual de perjudicial que les anteriors.

1.5 Proposta

En aquest treball es vol adaptar l'algoritme desenvolupat en l'article "*Exposing the probabilistic causal structure of discrimination*" [1]. Aquest algoritme està en R, ja que els autors de l'article volien una aproximació senzilla de la teoria que es presentava. El nostre objectiu serà adaptar-lo a Python, intentant replicar els seus resultats i introduint nous conceptes per a la valoració de discriminació. En segon lloc, es farà accessible aquest algoritme en una plataforma web.

Principalment, l'algoritme extreu d'una sèrie de registres de decisions les relacions causals entre elles. Aquesta extracció serà una *Suppes-Bayes Causal Network* (SBCN), un graf acíclic dirigit on els nodes són variables de Bernoulli.

Entenem per graf aquell objecte abstracte de la matemàtica discreta que representa un conjunt d'objectes (nodes) interconnectats per enllaços (arestes). En un graf dirigit, aquestes arestes estan connectades de manera que cada aresta només té un sol sentit, són orientades. En un graf acíclic, és impossible començar en un punt del gràfic i recorre'l fins retornar al punt d'origen. Per tant, en un graf acíclic dirigit (*DAG*), és impossible començar en un punt del gràfic i recorre'l tot.

Una variable de Bernouilli només pot prendre dos valors. En aquest context, aquests dos valors seran "què ha succeït" o "què no ha succeït". Totes i cada una d'aquestes variables o nodes estaran formats per una estructura del tipus "atribut-valor", com pot ser el cas de <*sexe_masculí* >.

D'aquesta manera, en la nostra SBCN, cada aresta implica una relació de causalitat entre els dos nodes, i anirà acompanyada d'una puntuació que indicarà com de forta és aquesta relació. Arribats aquí, es plantegen dues filosofies que s'apliquen successivament per avaluar aquest graf:

- **Mètode basat en restriccions:** En la creació d'aquesta SBCN se li imposaran dues restriccions: les anomenades *Suppes' constraints* [8]. En són dues, *temporal priority* (tota causa ha d'anar precedida del seu efecte) i *priority rising* (tota causa ha d'augmentar la probabilitat de que ocorri l'efecte).

- **Mètode basat en puntuacions:** Finalment, es farà ús del *Bayesian Information Criterion* o BIC per obtenir el model que millor explica la nostra *data* i que afavoreixi les relacions causals més disperses, evitant així l'aparició de falsos positius o relacions espúries.

Aquesta SBCN resultant pot ser explotada amb diversos algoritmes del tipus *random-walk* o *Page Rank* per obtenir puntuacions sòlides sobre l'aparició de discriminació o no.

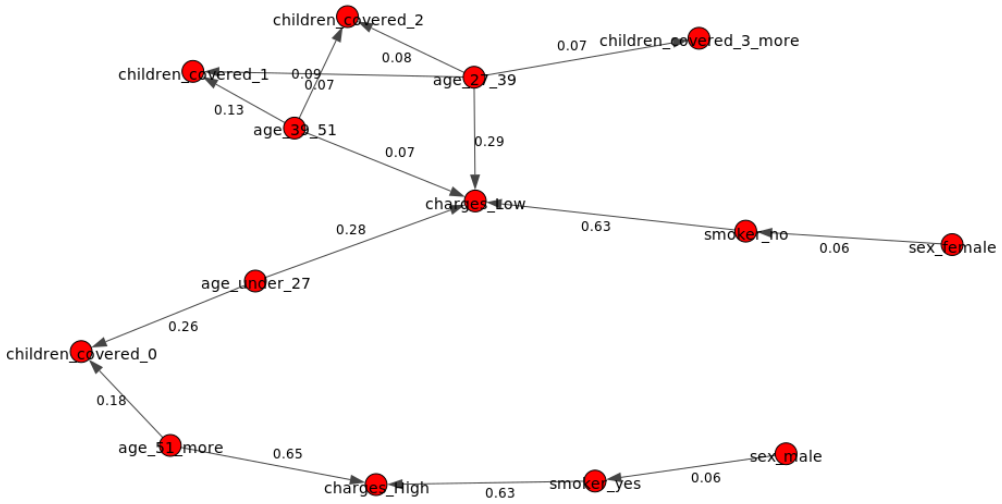


Figura 1: Exemple d'un petit graf resultant de despeses mèdiques als Estats Units

1.6 Full de ruta

Aquest treball està estructurat de la següent manera:

- **Objectius:** una sintetització d'on es vol arribar amb la implementació d'aquest treball.

- **Conceptes previs i estat de l'art:** es descriu el fonament teòric del treball, l'estat de l'art i què entenem com a mesura de discriminació.
- **Suppes-Bayes Causal Network:** s'expliquen les propietats d'aquesta xarxa i com es produeix el seu entrenament.
- **Implementació:** *pipeline* complet de com s'ha construït l'algoritme plantejat.
- **Experiments, Resultats i Discussió:** introducció dels datasets testejats, resultats obtinguts i la seva valoració.
- **Conclusions i futures millores:** valoració de la feina feta i com es podria optimitzar i continuar.

1.7 Objectius

En aquest treball de recerca és vol construir una infraestructura de programari que, mitjançant una base de dades categòrica amb atributs personals, sigui capaç de detectar qualsevol tipus de discriminació en ella.

Per aconseguir-ho, s'haurien d'assolir 3 fets

- L'entesa i clara definició de què és i com es pot manifestar la discriminació. Només d'aquesta manera sabrem com podem detectar-la i avaluar-la.
- L'entesa de l'algoritme de les SBCN des de l'arrel, per poder adaptar-lo al llenguatge de programació Python i saber com el podem perfeccionar i ampliar.
- Finalment, volem crear una plataforma en format d'aplicació web que sigui capaç d'executar aquest algoritme, recollir la seva informació i mostrar-la de manera comprensible en un temps raonable.

2 Conceptes previs i estat de l'art

En aquesta secció s'explica la teoria causal de Suppes i com s'adapta per a la detecció de la discriminació. A continuació, ens posem en context sobre altres projectes que hagin pogut usar un mètode científic semblant i finalment s'explica què entenem per "mesura de discriminació" i quines variants té. Abans, un petit glossari:

- **Nodes o variables de decisió:** són dues variables oposades presents en tota SBCN. Representen una decisió de "sí o no", com <*admision_denied*> i <*admision_rejected*>. Al voltant d'aquestes dues es fa el còmput de la valoració de la discriminació

- **Causalitat probabilística:** concepte usat per multitud de teories basat en que les "causes" canvien les probabilitats dels seus efectes [9].

2.1 Teoria causal de Suppes

Patrick Colonel Suppes (1922-2014) va ser un dels primers filòsofs de la ciència americà que va reconèixer la importància de la causalitat per a les ciències socials. Suppes deixa enrere les interpretacions de freqüència, propensió o correlació a l'hora d'estudiar la teoria de la probabilitat causal i ens dona una visió diferent basada en dos principis.

2.1.1 Definició

Donats dos esdeveniments A i B, els quals ocorren respectivament a instants de temps t i t' determinats, direm que A és *prima facie* causadora de l'esdeveniment B si:

$$0 < P(A_t), P(B_{t'}) < 1$$

$$t < t' \tag{1}$$

$$P(B_{t'}|A_t) > P(B_{t'}|\neg A_t) \tag{2}$$

Que es llegeix com "*la probabilitat de que succeeixi l'esdeveniment B si sabem que abans ha passat l'esdeveniment A ha de ser més gran que la probabilitat de que succeeixi l'esdeveniment*

B si no ha passat A". També és pot expressar com "*la probabilitat de que succeeixi B havent passat A ha de ser més gran que la probabilitat de que succeeixi B per si sola*":

$$P(B_{t'}|A_t) > P(B_{t'}) \tag{3}$$

Podem demostrar que 2 i 3 són equivalents?

- Gràcies a la formula de les probabilitats totals sabem quan val $P(B)$:

$$P(B) = P(B|A)P(A) + P(B|\neg A)P(\neg A) \tag{4}$$

- Aleshores, substituïm...:

$$4 \Leftrightarrow P(B|A) > P(B|A)P(A) + P(B|\neg A)P(\neg A) \tag{5}$$

- Traiem factor comú:

$$5 \Leftrightarrow P(B|A)(1 - P(A)) > P(B|\neg A)P(\neg A) \tag{6}$$

- Però (*1* - $P(A)$) = $P(\neg A)$, per tant:

$$6 \Leftrightarrow P(B|A)P(\neg A) > P(B|\neg A)P(\neg A) \tag{7}$$

- Sabent que $P(\neg A) > 0$, la eliminem dels dos costats, obtenint

$$7 \Leftrightarrow P(B|A) > P(B|\neg A)(2) \tag{8}$$

Les restriccions 1 i 2 són els dos pilars de la teoria de Suppes. La primera rep el nom de "prioritat temporal" o *temporal priority* mentre que la segona l'anomenem "incrementador de probabilitat" o *probability rising*. Al llarg d'aquest treball podrem veure com s'interpreten aquestes dues condicions en la xarxa causal que volem construir.

2.2 "Altres" mètodes

Dintre del global de les teories de causalitat no podem passar per alt el treball de Judea Pearl, contingut en major part en el llibre *Causality: Models, Reasoning and Inference* [10]. En ell, per exemple, s'explica l'algoritme *recovery* d'aprenentatge d'estructures, basat en les tres possibles relacions causals que hi poden haver en un DAG:

$$X \rightarrow Y \rightarrow Z$$

$$X \leftarrow Y \rightarrow Z$$

$$X \rightarrow Y \leftarrow Z$$

No obstant, nosaltres ens basem en la teoria de Patrick Suppes, el qual substitueix les tècniques habituals d'inferència basades en models gaussians (intervals de confiança, correlació...) per mètodes bayesians, basats en la probabilitat condicionada. Aquest enfoc és computacionalment més tractable. A continuació en podem veure alguns exemples.

Lu Zhang et al.,[11] per exemple, proposen una tècnica basada en el descobriment d'individus de característiques similars, els quals agrupen en tuples mitjançant xarxes causals bayesianes. Un cop formats tots els grups d'individus, s'examinen les diferències en una decisió concreta, definint així el grau de discriminació.

Niki Kilbertus et al., [12] es focalitza en evitar l'aparició de discriminació en algoritmes de decisió mitjançant un enfocament causal d'atributs protegits, com la raça, el sexe o la religió. Defineix quin ha de ser el criteri adequat d'equitat i a partir d'aquí desenvolupa algoritmes que el satisfan.

Yongkai Wu et al., [13] abandona el concepte d'atributs protegits i mitjançant un model log-lineal és capaç de detectar i mesurar discriminació entre multitud d'atributs de decisió. A partir d'aquest model també extreuen un graf d'independència condicional que examinat amb la Propietat de Màrkov és capaç de dictar patrons de discriminació.

Zliobaite et. al., [14] es centren amb el concepte de discriminació condicionada o explicable. Específicament, introdueixen un nou disseny en els models de detecció i eliminació de discriminació per a què siguin capaços de detectar també una discriminació condicionada.

Finalment, Binh Thanh Luong et al., [15] argumenta que la mineria de dades per al descobriment de discriminació en datasets de decisions històriques no mira un possible rerefons social ni un per què de l'estructura jeràrquica interna de la base de dades. En conseqüència, proposa un marc per al descobriment de discriminació basat en ontologies. [2]

Si es vol ampliar els coneixements de l'estat de l'art actual es pot consultar la secció *related work* de l'article "*Exposing the probabilistic causal structure of discrimination*" [1] sobre el que ja hem comentat que està basat aquest treball.

2.3 Mesures de la discriminació

Gràcies a la secció 1.2 coneixem el concepte de discriminació i de quina manera es pot manifestar. En aquesta apartat, veurem aquests mateixos conceptes aplicats a l'algoritme de la *Suppes Bayes Causal Network*. En concret, s'han implementat dos mètodes per la detecció i valoració de tots els tipus de discriminació citats.

2.3.1 Puntuació de la discriminació per *random walks*

L'algoritme de *random walk* en un graf de tipus DAG és ben senzill: es comença a caminar en un node i no es para fins arribar o bé a un node fulla (node sense arestes sortints) o bé al límit de passos establerts. A cada node per on passem, ens apuntem el seu nom, obtenint així una llista de nodes visitats o "camí".

Si en múltiples iteracions de l'algoritme en que es comença en una variable concreta s'acaba sovint en el node que representa la decisió negativa, possiblement estarem davant

[2] L'ontologia és l'estudi de l'ésser o de l'existència. Un ontologia per tant pot descriure les categories i relacions bàsiques entre éssers.

d'un cas de discriminació negativa pels individus presents en la mostra d'aquella variable. Si, en canvi, acabem sovint en la variable de decisió positiva, estarem davant d'un cas de favoritisme. Finalment, si sovint acabem en un node irrellevant o node fulla, ens trobem davant d'una situació inconclusiva.

Per tant, per a cada variable present es realitza un puntuatge de discriminació positiva, negativa i inconclusiva. Es defineix aquesta puntuació com:

$$\eta_{a \to b} = \frac{W_{a \to b}}{N} \tag{9}$$

On η pot ser puntuació positiva (+), negativa (-) o inconclusiva ($\varnothing$), N és el número de vegades que s'executa l'algoritme i $W\, a \to b$ és el número de vegades que hem acabat en b (el node de decisió positiva o negativa) començant des del node a. Per a la puntuació inconclusiva, b és qualsevol variable menys les de decisió positiva o negativa. En conseqüència:

$$\eta^{+} + \eta^{-} + \eta^{\varnothing} = 1 \tag{10}$$

El número de passos d'aquest algoritme és el diàmetre del graf. Definim el diàmetre d'un graf com la màxima excentricitat de qualsevol dels seus nodes, és a dir, el màxim dels mínims camins que poden haver-hi entre qualsevol parell de nodes. D'aquesta manera, ens assegurem que mai pararem per falta de passos.

2.3.2 Mitjana de passos

Cada vegada que l'algoritme de *random walk* arriba a un node de decisió, ho fa amb un nombre determinat de passos. Aquest nombre indica com de lluny està una variable d'una decisió negativa o positiva.

Gràcies a ell, podem determinar la mitjana de passos requerits o μ, un altre factor que donarà valor a la puntuació de discriminació ja que una variable serà més discriminada que una altre si a part de tenir un valor d'η alt té una mitjana de passos petita.

El càlcul d'aquestes mitjanes es realitza comptant quants passos han fet falta per arribar a un node de decisió positiva o negativa. Si en 100 crides a l'algoritme *random walk* 3 han acabat en un node de posició negativa amb 4, 5 i 6 passos per arribar-hi cadascuna, direm que:

$$\mu^{-}_{a\to b} = \frac{5+6+4}{3} = 5 \tag{11}$$

2.3.3 Puntuació inconclusiva

Les puntuacions negativa i positiva són força clares, però el concepte de inconclusiva és nou. La puntuació inconclusiva és aquella que valora quants *random walks* acaben en un node diferent dels de decisió positiva o negativa. Una puntuació inconclusiva es pot manifestar de 2 maneres:

1. Inconclusiva total: direm que tenim una variable inconclusiva total quan la seva puntuació inconclusiva estigui per sobre de 0,55. Per aquella variable, més de la meitat de *random walks* acaben en un node irrellevant, i, per tant, direm que el nostre mètode no pot mesurar l'aparició de discriminació o no en aquesta variable.

2. Inconclusiva parcial: què ocorre si part dels *random walks* acaben en un node irrellevant, l'altre part en una decisió positiva i casi mai en una decisió negativa? No podem parlar d'un favoritisme clar, ni tampoc d'una variable neutra ja que poc sovint arribem a una decisió negativa. Direm que ens trobem davant d'una inconclusivitat parcial si:

$$\eta^{-} \simeq 0$$

$$|\eta^{+} - \eta^{\varnothing}| \leq 0,25 \Leftrightarrow \text{Aparentment favoritisme} \tag{12}$$

O bé

$$\eta^{+} \simeq 0$$

$$|\eta^{-} - \eta^{\varnothing}| \leq 0,25 \Leftrightarrow \text{Aparentment discriminació negativa} \tag{13}$$

2.3.4 Variable neutra

Només ens queda un tipus de combinació de puntuacions possible: la neutra. Direm que una variable és neutra si i només si:

$$|\eta^+ - \eta^-| \leq 0.06 \Leftrightarrow \text{Neutralisme} \tag{14}$$

És a dir, pràcticament poques vegades arribo mai a una decisió neutra i la resta de camins em porten amb mitjanament igual probabilitat a una decisió tant positiva com negativa.

La figura 2 ens ajuda a entendre el procés de tria de veredicte.

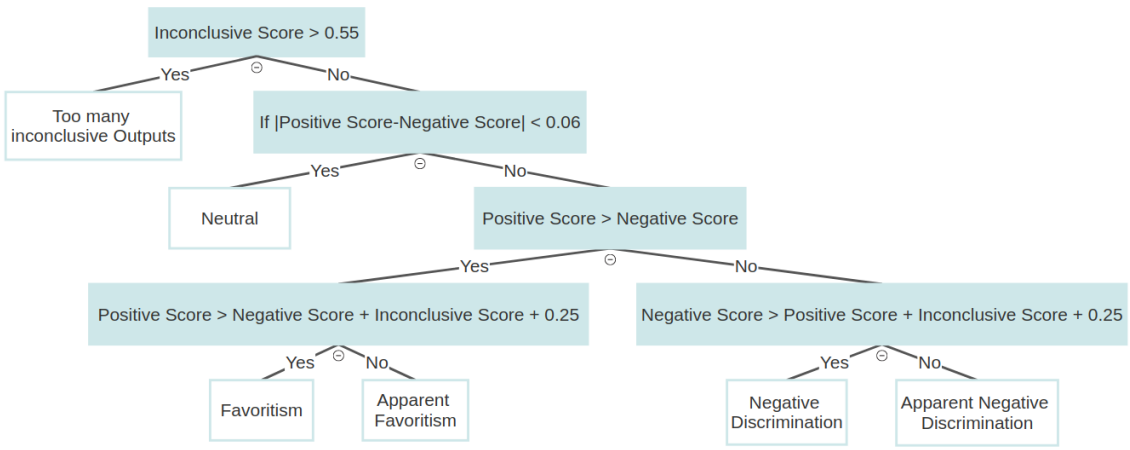


Figura 2: Arbre de decisió a seguir a l'hora d'establir un veredicte

2.3.5 Discriminació explicable o condicionada

Donat l'exemple plantejat en la secció 1.3, on coneixem el gènere i nivell d'anglès dels candidats a varis llocs de treball d'una empresa, realitzem 1000 iteracions de l'algoritme de *random walk* començant sempre des del node a = <*sexe_masculí*>. De tots els camins, la gran majoria acaben en una decisió negativa b = <*candidat_rebutjat*>. Sense valoració de discriminació explicable o condicionada, diríem que els homes estan sent discriminats. Si, en canvi, plantegem:

$$\eta^{cond} a \rightarrow b = \frac{W_{a \rightarrow c \rightarrow b}}{W_{a \rightarrow b}} \tag{15}$$

On c és el node que representa la variable <anglès_baix>, $W\ a \rightarrow c \rightarrow b$ el número de camins que impliquen passar abans pel node c abans d'arribar a la decisió negativa b i $W_a \rightarrow b$ el número de camins que no passen per c però arriben igualment a b, podrem valorar com d'explicable o condicionada està aquesta discriminació. Aquest node c l'anomenarem "node entremig".

2.3.6 Puntuació de la discriminació amb *Page Rank*

Larry Page, co-fundador de Google va ser la persona responsable de co-crear i anomenar l'algoritme *Page Rank*, un sistema per classificar per importància pàgines webs en el seu cercador. Per a *Page Rank*, la importància d'una pàgina web es basa en el número de links entrants i sortints [16].

Des de llavors, milers de persones han utilitzar la filosofia d'aquest algoritme per multitud d'enfocaments diferents, com mesurar el rendiment de jugadors de futbol americà [17] o predir quantes persones passejaran per un carrer [18].

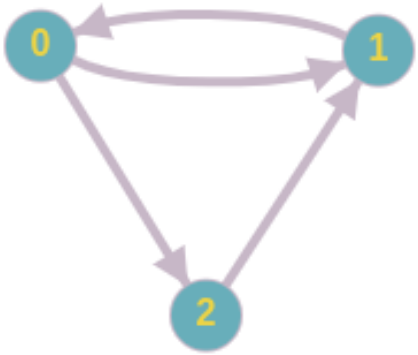


Figura 3: Exemple de graf on aplicar Page Rank

Donada la xarxa de la figura 3, l'algoritme de *Page Rank* ens donarà la probabilitat que una persona que va clicant de manera aleatòria els enllaços de les pàgines web de la

xarxa acabi en una d'altre concreta. Així doncs, definirem la puntuació de *Page Rank* per a qualsevol de les seves pàgines com:

$$PR(u) = \sum_{v \in K_u} \frac{PR(v)}{L(v)} \tag{16}$$

On u és qualsevol de les pàgines de la xarxa, K_u el conjunt de pàgines que referencien a u, $PR(v)$ la puntuació de *Page Rank* de la pàgina v i $L(v)$ el número de links de sortida de la pàgina v. Per tant:

$$\begin{aligned} PR(x_0) &= \frac{1}{1} \\ PR(x_1) &= \frac{1}{2} + \frac{1}{1} \\ PR(x_2) &= \frac{1}{2} \end{aligned}$$

Aquest sistema es pot resoldre iniciant les puntuacions a $1/N$ on N és el total de nodes o amb el mètode de la potència per intentar trobar el vector propi de valor propi 1:

$$\begin{aligned} \lambda &= 1 \\ \lambda x &= Gx \end{aligned} \tag{17}$$

On G és la representació en forma de matriu del sistema que acabem de veure. Tots dos mètodes necessiten un cert nombre d'iteracions fins assolir els valors més òptims.

La persona que va navegant, anomenada *random surfer*, eventualment es cansa de clicar. La probabilitat, a cada pas que realitza, de que es cansi i escrigui directament la *url* de la pàgina on vol anar s'anomena *1 - damping factor*. Si anomenem d a aquest *damping factor*:

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in K_u} \frac{PR(v)}{L(v)} \tag{18}$$

On N és el número total de pàgines. Dins d'N, hi ha pàgines sense enllaços o nodes fulla. La seva puntuació de *Page Rank*, aleshores, és directament *1-d*. Per tant, quan el *random surfer* arriba a una pàgina sense enllaços, saltarà cap a una altre pàgina amb una probabilitat uniforme, o el que és el mateix: una pàgina sense enllaços de sortida és com si tingues enllaços a totes les pàgines de la xarxa.

Però, en adaptar aquest algoritme en el nostre projecte, pateix dues personalitzacions claus:

- El *random surfer* no triarà el següent node (enllaç) de manera aleatòria, sinó que ho farà tenint en compte els pesos de cada aresta

- En topar amb un node fulla, la probabilitat de saltar no serà uniforme per a tots els nodes, sinó que estarà donada pel vector (fila) del dataset corresponent a l'individu que estiguem valorant. Específicament, li donem a l'algoritme una probabilitat concreta necessària per saltar a cada node del dataset. Agafem un d'aquests nodes o variables de Bernoulli. Si l'individu és una mostra d'aquesta (hi ha un 1 en la columna de la variable), li adjudicarem una probabilitat de $1/n$, on n és el total de variables. Si no forma part de la mostra, la probabilitat serà zero. Amb altres paraules, només podrem saltar als nodes de característiques que tingui l'individu.

Així doncs, *Page Rank* ens donarà per a cada individu un conjunt de puntuacions, tantes com variables hi hagin en el nostre DAG. La suma d'aquestes puntuacions és 1. D'aquestes, només ens interessen dues: les de decisió positiva i negativa.

3 Suppes-Bayes Causal Network

En aquest apartat s'introdueix un mètode per generar un tipus de xarxa Bayesiana causal perfilada per satisfer les condicions de la teoria de la causalitat de Patrick Suppes2.1, anomenada xarxa causal de Suppes-Bayes (SBCN).

La SBCN parteix d'una taula relacional D que ens proporciona l'usuari. En aquesta taula, cada columna serà una variable de Bernoulli. Cada fila present en D representa un individu, el qual es mostreja per a cada columna amb un 0 si posseeix la característica de la variable o amb un 0 si no. D serà vàlida sempre i quant hi hagi més de 2 variables amb més d'una mostra.

La raó per a la qual hi han d'haver mínim 3 variables és perquè en D sempre hi haurà obligatòriament dues que representaran una decisió negativa o positiva, com per exemple, <*loan_denied*> (negativa) o <*loan_accepted*> (positiva) [3].

Els següents apartats descriuen com es crea la SBCN per a què compleixi les dues condicions de la teoria de Suppes.

3.1 Prioritat Temporal

Tota causa ha d'anar precedida del seu efecte, és a dir, no poden haver-hi causalitats del tipus <*notaExamen_suspes*> a causa de <*paisNatal_Mèxic*> ja que el país d'origen d'un individu és determinat prèviament a la realització del examen.

El nostre model de dades no té cap tipus d'informació temporal de quan s'ha produït un esdeveniment, així que el que fem és proveir-nos d'una segona taula r que conté l'ordre temporal per a tots i cadascun dels atributs categòrics de D. Aquest ordre es representa amb un número natural.

Partint d'aquestes dues taules ja podem realitzar un graf $G = (V,E)$ on cada node de $|V|$ serà cadascuna de les m variables de Bernouilli i cada aresta $(u,v) \in E$ serà dibuixada

[3] S'ha de dir també, que en el cas de que hi hagi 3 variables la xarxa quedarà desconnectada

sempre i quan $r(u) \leq r(v)$. D'aquesta manera ens assegurem de tota relació de causalitat complirà el principi de la prioritat temporal.

3.2 *Probability Rising*

Tota causa ha d'augmentar la probabilitat que ocorri l'efecte respecte la possibilitat que ocorri l'efecte si no ha passat la causa, així que d'aquest mateix graf $G = (V,E)$ eliminem cada aresta (u,v) que compleixi $P(v \mid u) \leq P(v \mid \neg u)$, obtenint així un *prima facie* graf.

3.3 Aprenentatge de la SBCN

Fins ara hem seguit una filosofia basada en restriccions. Addicionalment, aquesta anomenada SBCN sofreix un "entrenament" per eliminar les possibles relacions invàlides (falsos positius o relacions espúries) basat en el mètode de puntuació anomenat "BIC".

En el context d'entrenament de grafs acíclics dirigits, un mètode de puntuatge intenta ajustar múltiples DAGs a un criteri comú i en selecciona el millor (el de més puntuació). En el mateix context, *Bayesian Information Criterion* (BIC) o també *Schwarz information criterion* (*SIC*) és precisament un mètode per avaluar com de bo és un DAG. Aquests DAGs són generats mitjançant l'algoritme *Hill Climbing*.

3.3.1 *Bayesian Information Criteria*

A l'hora d'ajustar models sempre correm el risc de cometre *overfitting*, és a dir, intentar acostar-nos massa a les dades "reals" reduint poder predictiu. En moltes ocasions, l'*overfitting* ocorre quan parametritzem les dades en excés. BIC, en conseqüència, penalitza l'adició de paràmetres: un model tindrà major puntuació que un altre si el primer té menys paràmetres.

Aquest "acostament" a les dades esmentat anteriorment es tradueix en el BIC com la *Likelihood function* o funció de versemblança, la qual permet deduir, a partir d'un conjunt d'observacions i paràmetres, un valor que mesura com de bé s'adapten a un model probabilístic donat:

$$L(b|A) = P(A|B = b) \tag{19}$$

On A i B són variables aleatòries que modelen uns paràmetres desconeguts i b el valor que volem inferir. La probabilitat de que ocorri B si sabem que ha passat A es dedueix gràcies al teorema de Bayes, el qual dicta:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{20}$$

Per millorar l'eficiència del comput d'aquesta funció, sovint s'utilitza la seva versió logarítmica, la qual ens aporta simplificació en els càlculs (productes convertits en sumes, divisions en restes i exponencials en multiplicacions...) i un escurçament considerable en el rang de dades possibles [19].

Així doncs, BIC es tradueix en:

$$BIC = -2L_{log}(x) + k\log(n) \tag{21}$$

On $L_{log}()$ és la funció de versemblança logarítmica, x representa les dades del nostre model, k el número de paràmetres i n el número de representacions de x. Per tant, el terme penalitzador $\log(\mathrm{n})$ es veu regulat per k, on aquest cas és la dimensió del graf.[4]. El model que triarem serà aquell amb la puntuació BIC més baixa.

[4]La dimensió d'un graf és l'enter més petit n amb el que podem representar el graf en un espai euclidià de dimensió n de manera que totes les arestes tinguin a la mateixa longitud

3.3.2 *Hill Climbing*

Hill Climbing és un algoritme d'optimització que intenta millorar una possible solució a un problema mitjançant un únic canvi incremental a cada iteració. Aquest canvi serà acceptat si i només si ofereix una millora a la solució anterior. Aquest procediment acaba quan (i) no hi ha cap millora des de l'últim canvi perfeccionat o (ii) s'han completat un número llarg d'iteracions.

En el millor dels casos, per tant, *Hill Climbing* pot trobar la solució més adequada al problema, però també pot trobar una solució secundària vàlida.

En el context de la nostra SBCN, apliquem *Hill Climbing* a un DAG. Es comença amb un graf buit, fins que es troba un node inicial en el qual la seva avaluació impliqui una maximització de la funció objectiu. En aquest cas, aquesta funció a maximitzar és la de versemblança, i la avaluació a la que es sotmet una solució és la del *Bayesian Information Criteria*.

A partir d'aquest node inicial, l'algoritme intentarà sempre realitzar 3 tipus de canvis, els quals es poden relacionar mínimament amb l'algoritme *recovery* de Judea Pearl que hem vist en la secció 2.2:

- **Inserció:** significa l'adició d'una nova aresta, i, per tant, de l'aparició d'un nou node al graf.
- **Reversió:** significa realitzar un canvi de sentit en alguna de les arestes de l'últim node afegit.
- **Supressió:** significa treure una aresta de l'últim node afegit, i, per tant, de la desaparició del graf del node de l'altre extrem.

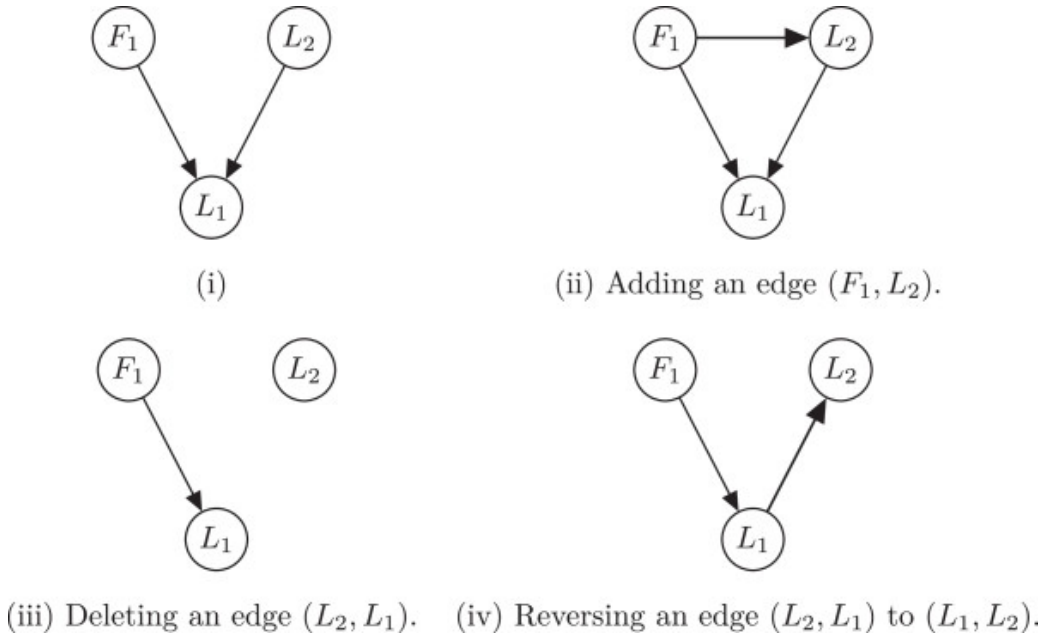


Figura 4: Accions possibles en l'algoritme *Hill Climbing* [2]

En resum, l'algoritme de *Hill Climbing* ens proporciona un seguit de models (o subgrafs) derivats del DAG inicial. Per escollir el millor d'ells, es puntuen amb la funció de versemblança, regularitzada amb el mètode de puntuació BIC.

En termes de pseudocodi, on "ADDNODE()" és una funció que permet afegir un node i les seves propietats al graf i "EVAL()" la funció de realitzar la puntuació d'un graf mitjançant la funció de proximitat i un mètode de puntuació, tenim:

Algorithm 1 Hill Climbing

Inputs: G(V,E), un DAG amb pesos. nIter: un màxim d'iteracions.

Output: G*(V*,E*), possible millor solució.

```
i ← 0
improvement ← true
currNode ← randomNode(G)
G* ← DAG()
while i < nIter and improvement do
    ADDNODE(G*, currNode)
    improvement ← false
    for all possible actions do
        tempScore, nextNode ← EVAL(G, currNode, action)
        if tempScore > score then
            score ← tempScore
            currNode ← nextNode
            improvement ← true
        end if
    end for
    i ← i + 1
end while
return G*
```

3.3.3 Pes de les arestes

El nostre model final ens dona la millor aproximació de totes les relacions causals que existeixen en el dataset inicial. Per a cada aresta present en aquest DAG reconstruït se l'hi ha d'assignar una puntuació, un pes que ens representi la intensitat de la relació entre els dos nodes que uneix. Així doncs, per a cada aresta que uneix dos nodes v i u del DAG realitzem:

$$S(v, u) = P(u|v) - P(u|\neg v) \tag{22}$$

És a dir, una resta de probabilitats condicionades. Com que complim amb les restriccions de la teoria de Suppes, $P(u \mid v) \gg P(u \mid \neg v)$, i, per tant, $S(v,u) \in (0,1]$.

4 Implementació

En aquest apartat es descriu tot el procés d'adaptació de l'algoritme, com s'han programat els algoritmes de recorregut de grafs i finalment s'explica l'esquelet de l'aplicació web i el seu procés de *deploy*.

4.1 Arquitectura

El projecte inicialment es volia implementar amb Java per facilitar més tard el seu traspàs a pàgina web. Python, però, és un llenguatge de programació més afí pel que fa a la gestió d'estructures de dades, amb biblioteques de codi obert molt eficaces i eines d'anàlisi de dades de tot tipus.

Es fa ús principalment de cinc llibreries:

- **Pandas:** per les estructures de dades d'*input* i d'*output*, és a dir, bases de dades relacionals. És de codi obert.
- **NumPy:** per operar amb vectors i matrius. És de codi obert.
- **Rpy2:** interfície per fer ús del llenguatge de programació R des de Python. És de codi obert.
- **bnlearn:** paquet d'R per l'aprenentatge d'estructures Bayesianes. És de codi obert.
- **iGraph:** per la creació i visualització de grafs, així com la realització d'algoritmes en ells. És de codi obert.

4.2 *Pipeline* de l'algoritme

La implementació d'aquest algoritme es pot dividir en 6 processos:

- **Entrada i neteja de dades:** s'explica com han de ser les dades d'entrada i per quins processos es sotmeten fins a ser aptes per la computació de l'algoritme
- **Càlcul de probabilitats:** s'explica el fonament teòric del càlcul de probabilitats i

la seva utilitat en l'algoritme.

- **Compliment de les condicions de la teoria de Suppes:** es realitzen les comprovacions necessàries per assegurar que seguim la teoria de Suppes.
- **Entrenament i estimació de la xarxa causal:** s'explica per què i com es realitza un entrenament en la nostra xarxa i què s'espera després.
- **Reconstrucció de la SBCN:** descripció del procés de construcció del model entrenat.
- **Puntuació de la discriminació:** s'expliquen quins mètodes de puntuació s'han usat i quines són les millors mètriques per mesurar la discriminació.
- ***Deploy* de l'aplicació web:** explica el procés de creació i desplegament de la plataforma.

4.3 Entrada i neteja de dades

L'algoritme necessita una entrada de dades en forma de taula relacional, on cada columna hi trobem una variable de Bernoulli (recordem, del tipus *atribut - valor*) representada de forma categòrica. En aquest cas, aquestes variables només poden prendre dos valors, '1', hem observat el fenomen, o '0', no hem observat el fenomen. Així doncs, cada fila d'aquesta taula correspon a un individu on es descriuen les seves propietats amb uns i zeros.

Per tant, en aquest procés del *pipeline* es realitzen dues accions claus en cada cel·la de la taula. La primera, la comprovació de que en cadascuna d'elles hi ha algun valor (és a dir, no hi ha cap cel·la buida), i la segona, que aquest valor sigui, obligatòriament, o un zero, o un u.

Finalment, es fa una comprovació més: que el número de columnes sigui igual o superior a tres i que el número de files sigui superior o igual a dos, ja que es necessita mínim dues variables de decisió i una "d'atribut" i mínim dues o més persones. Aquesta comprovació es repeteix una vegada més en el transcurs de l'algoritme ja que en cert

moment és possible que s'eliminin esdeveniments.

4.4 Càlcul de probabilitats

Un cop sabem que les dades d'entrada són vàlides i aptes per a l'explotació, prosseguim amb el càlcul de les probabilitats a través de freqüències relatives de cara parell de variables presents (les anomenarem X i Y en la taula, per després construir el nostre graf. En són 3:

- **Probabilitat Marginal d'X:** és aquella computada fent cas omís d'Y:

$$P(X) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{23}$$

  És a dir, sumar tota la columna i dividir pel nombre d'individus (n).

- **Probabilitat Conjunta d'X i Y:** indica la probabilitat de que succeeixin els dos esdeveniments a la vegada. Per exemple, donades les variables X='*sexe_dona*' i Y='*raça_llatina*, direm que:

$$P(X \cap Y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} \tag{24}$$

  És a dir, multiplicar les columnes d'X i Y obtenint el nombre de de persones de sexe femení i raça llatina i després dividint pel nombre d'individus totals.

- **Probabilitat condicionada d'X donada Y:** és aquella que indica la probabilitat de que ocorri X havent succeït Y. Per exemple, donades les variables X='*sexe_dona*' i Y='*raça_llatina*, direm que:

$$P(X|Y) = \frac{P(X \cap Y)}{P(X)} \tag{25}$$

És a dir, la probabilitat conjunta d'X havent succeït Y dividida per la probabilitat marginal d'X, obtenint la probabilitat de que una persona sigui de sexe femení si sabem que és de raça llatina. De la mateixa manera, podem deduir la probabilitat de que una persona sigui de sexe femení sabent que no es de raça llatina com:

$$P(X|\neg Y) = \frac{P(X) - P(X \cap Y)}{1 - P(Y)} \tag{26}$$

Les primeres probabilitats que calculem són les marginals i conjuntes. Per calcular la condicionada només hem de dividir aquestes dues. Durant aquest procés, és possible que ens trobem una divisió que doni de resultat u. Això no és incorrecte, simplement vol dir que els dos fenòmens de les dues variables en qüestió "van de la mà", no es poden distingir. Així doncs, el que fem és unir-les, o més ben dit, eliminar una de les dos. En conseqüència, es torna a comprovar que el número de columnes sigui major o igual a tres.

També és possible que la taula sigui errònia i hi hagi una columna plena de uns o plena de zeros. En aquest cas, s'elimina la columna en qüestió també.

El comput de probabilitats condicionades es fa en una matriu n x n on n és igual al número de variables (columnes) del dataset D. En conseqüència, a la diagonal d'aquesta hi col·locarem un u, ja que es tracta de la mateixa variable. Aquest conjunt de probabilitats li direm "Model Prima Facie", ja que en un principi assumim que es tracta d'esdeveniments causals fins que no es demostri el contrari.

De la mateixa manera, definirem el model "Prima Facie Not Happened" com el conjunt de probabilitats condicionades del tipus $P(X \mid \neg Y)$, calculades també en un espai format per una matriu n x n on n és igual al número de variables (columnes) del dataset.

4.5 Compliment de les condicions de la teoria de Suppes

L'objectiu d'aquest apartat és extreure una matriu d'adjacència A que compleixi amb les condicions de prioritat temporal i *probability raising* a partir dels models definits anteriorment. En un principi, aquesta matriu d'adjacència esta plena d'uns, excepte la seva diagonal, en la qual hi trobem zeros.

Definirem el graf d'aquesta matriu com $G = (V,E)$, on $|V|$ són totes les variables de Bernoulli presents en el dataset. Pel que fa al conjunt d'arestes, $|E|$, en un principi establim que totes les variables estan lligades entre si (per això la matriu en un primer moment està plena d'uns). A mesura que explorem les condicions de la teoria de Suppes, algunes d'aquestes arestes aniran desapareixent.

4.5.1 Comprovació de la prioritat temporal

És en aquest apartat on entra en acció per primera vegada la segona taula relacional que ens ha de proporcionar l'usuari, l'anomenada taula d'ordre temporal o r. Aquesta té dos columnes, "atribut" i "ordre". En la columna atribut es llisten totes les variables presents en el dataset. En la segona, es llista un número enter natural que defineix l'ordre de cadascuna d'aquestes variables.

Així doncs, per a cada aresta $(u,v) \in E$ de G, comprovarem si $r(u) \leq r(v)$. De no complir-se la condició, canviarem el valor de la matriu d'adjacència en aquella posició de u a zero. És possible que diferents variables posseeixin el mateix valor d'ordre temporal. En aquest cas, no es canvia cap valor.

4.5.2 Comprovació de la *probability rising*

Gràcies al nostre model "Prima Facie" i "Prima Facie Not Happened", posseïm per a cada parell de variables les seves probabilitats condicionals. Així doncs, per a cada aresta $(u,v) \in E$ de G, comprovarem si $P(v \mid \neg u) \leq P(v \mid u)$. De no complir-se la condició, en cas de que hi hagués un 1 en la matriu d'adjacència en aquella posició, canviaríem el seu valor per un zero.

4.6 Entrenament i estimació de la xarxa

En aquest moment ens trobem amb una matriu que ens defineix totes les possibles relacions causals entre les variables del nostre dataset. Nogensmenys, d'entre totes elles, només unes quantes en són vàlides. Tot i ja n'hem destriat forces fent ús de les condicions de la teoria de Suppes, ens disposem a aplicar un mètode d'entrenament que ens acabi de perfilar una xarxa causal molt pròxima a la realitat.

Tal i com hem explicat en la secció 3.3, procedim a entrenar el nostre model mitjançant *Hill Climbing*, una funció de versemblança i *Bayesian Information Criteria* com a criteri. Abans de procedir a executar l'algoritme, hi ha dues accions importants a realitzar:

- Preparar el dataset D: la llibreria que realitzarà el procés d'entrenament no pot processar un *input* numèric, encara que sigui de l'estil binari. Així doncs, iterem D canviant els uns per un *string* com "*hit*" i els zeros com un altre *string* com "*miss*".

- Llistar les relacions invàlides: gràcies a la matriu d'adjacència construïda sabem quines relacions entre dues variables estan "prohibides" per la teoria de Suppes. La manera d'indicar al nostre procés d'aprenentatge que no miri aquest conjunt de relacions o arestes és creant un dataset amb dues columnes: la primera és l'origen de la aresta (el nom de la variable o node) i la segona el destí de la aresta. Podríem pensar en aquest dataset com una espècie de llista negre o *blacklist*.

La llibreria que ens proporcionarà els recursos necessaris per realitzar l'entrenament s'anomena *bnlearn*. Concretament, volem funcionalitat de la seva funció *hc*, la qual resulta idònia pel tipus de necessitats que tenim. Específicament, és la única funció que s'ha pogut trobat que entén com a *input* una *blacklist*, una llista d'arestes les quals ha d'ignorar.

A l'inici d'aquest projecte bnlearn només estava disponible per a R. Així doncs, es va fer servir una altre llibreria anomenada Rpy2. Rpy2 és una interfície cap al llenguatge de programació R i els seus paquets. Funciona de manera que els objectes R es definei-

xen com instàncies de classes implementades amb Python, amb les funcions d'R com a mètodes enllaçats a aquests objectes.

Així doncs, l'entrenament es realitza a través d'aquesta interfície. Per a cridar la funció *hc* a través d'ella, s'han hagut de realitzar les següents mesures:

- Definir el nostre dataset com una matriu de NumPy de tipus "categòrica", ja que sinó bnlearn detecta de manera errònea que el dataset conté NaN's o valors nuls.
- Instal·lar l'objecte que conté la instància de la llibreria Rpy2 el paquet "utils", sense el qual no es podrien instal·lar més paquets.
- Instal·lar el paquet bnlearn l'objecte que conté la instància de la lliberia Rpy2.
- Passar el nostre dataset i la nostre *blacklist* a datasets d'R, amb la funció de Rpy2 anomenada *py2ri*.

Un cop tenim tots els paràmetres a punt, es crida la funció *hc* amb mètode de puntuació BIC i es recullen les arestes del model trobat. A continuació, es construeix la matriu d'adjacència d'aquest model, la qual ens permetrà realitzar el DAG amb pesos finals.

4.7 Reconstrucció de la SBCN

Per fi s'ha trobat el model més pròxim a la realitat possible i que compleix amb les condicions de la teoria de Suppes. Tot i així, es desconeix com de "forta" és una relació entre dos nodes adjacents: falta establir una puntuació o "pes" a cada aresta.

Tal i com s'ha definit a la secció 3.3.3, procedim a realitzar una resta de probabilitats condicionades per a cada aresta en el graf (per cada "1" de la matriu d'adjacència). Per fer-ho, recuperem les probabilitats marginals i conjuntes que s'han calculat en la secció 4.4.

Específicament, es construeixen 4 llistes de llistes. Aquestes, per a cada u de cada variable, apunten:

- La primera, apunta el nom de la variable.
- La segona, la probabilitat condicionada $P(i,j)$
- La tercera, la probabilitat condicionada de "no haver observat l'esdeveniment": $P(i,\neg j)$
- La quarta, la resta de la segona amb la tercera llista, és a dir, el pes de cada aresta o $S(u,v)$.

De manera que, si una variable qualsevol té 3 vèrtexs de sortida, tindrem en un índex concret de cada llista de llistes una llista de longitud 3. [5].

Finalment, amb la llista de noms i de restes es crea un dataset amb tres columnes:

- ***From*:**, la qual conté el node d'on neix l'aresta
- ***Target*:** la qual conté el node on va a parar l'aresta
- ***Weight*:** la qual conté el pes d'aquesta aresta

La creació d'aquest dataset es pot usar com a *input* en mètodes per crear un graf "des de zero" de multitud de llibreries.

Un cop tenim el graf creat, podem veure la desaparició d'algunes variables, és a dir, nodes que han quedat aïllats. Quan una variable queda aïllada es diu que és irrellevant per a la detecció de la discriminació, és a dir, que no aporta cap tipus d'informació. Aquest conjunt de variables és mostrada l'usuari en forma de taula.

4.8 Puntuació de la discriminació

Res del que hem implementat anteriorment serveix de res si no li donem un valor, un significat que ens proporcioni el resultat que volíem descobrir. En conseqüència, s'han

[5]Per facilitar el codi que computa el graf final, la llista de llistes de noms conté en segons quins índex el mateix nom repetit

implementat dos sistemes de puntuació de la discriminació.

El primer calcula per a cada variable la seva puntuació de discriminació positiva o favoritisme, la seva puntuació de discriminació negativa i la seva puntuació inconclusiva. Aquest sistema l'anomenem *puntuació de la discriminació per random walk*.

El segon també està basat (i usa) l'algoritme de *random walk*, però mesura la importància d'un subconjunt de variables respecte al total de variables presents en el DAG, és a dir, implementa l'algoritme de *Page Rank*.

4.8.1 Puntuació de la discriminació per *random walk*

En la secció 2.3.1, hem definit l'algoritme de *random walk* i l'ús que en fem de la informació que ens proporciona. En la secció 4.1 hem explicat l'ús de *iGraph* com a llibreria pel tractat de grafs en l'algoritme. Bé, *iGraph* no té cap mètode de *random walk*, així que es va decidir crear-lo de zero.

Algorithm 2 Personalized Weighted Random Walk

Inputs: G(V,E), start, end

Output: llista de nodes visitats

1: $visited \leftarrow list()$
2: *append start to visited*
3: **for** *i = 0 to diameter*(G) **do**
4: $nodes \leftarrow list()$
5: $probs \leftarrow list()$
6: $currNode \leftarrow visited[-1]$
7: **if** $currNode = end$ **then**
8: **return** $visited$
9: **end if**
10: **for** *every neighbor of currNode* **do**
11: *append neighbor name to names*
12: *append neighbor edge probability to probs*
13: **end for**
14: **if** *names is empty* **then**
15: **return** visited
16: **end if**
17: *make all probs sum to 1*
18: $nextNode \leftarrow CHOICE(names, probs)$
19: *append nextNode to visited*
20: **end for**
21: **return** $visited$

Tal i com podem veure, es segueix la filosofia tradicional del *random walk* però es modifica a la vegada per personalitzar-lo amb les nostres necessitats, que són:

- Modificar la tria del següent node a visitar segons el pes de l'aresta (línia 18). D'aquesta manera, les arestes amb més pes són les que tenen més probabilitats a ser triades. Recordem que hem definit el pes de l'aresta com $S(v,u) = P(u \mid v)$ -

$P(u \mid \neg v)$. Aquesta tria s'implementa amb el mètode *choice*(*elements*, *probabilitats*) de NumPy.

- L'adició d'un paràmetre *end* que permet parar "la caminada" si arribem a un node de nom (linies 7 i 8) *end*. Aquest paràmetre, per tant, serà el nom de la variable o node de decisió positiva quan vulguem puntuar la discriminació per favoritisme, i el nom de la variable o node de decisió negativa quan vulguem puntuar la discriminació negativa. Per la puntuació inconclusiva o neutra, no tindrà valor.

Aquest algoritme és sòlid i eficaç i ens proporciona tota la informació necessària per puntuar la discriminació de cada variable i per tant formular un veredicte final. Concretament, per a cada iteració de l'algoritme un mètode agafa els seus resultats i:

- Mira si l'últim element de la llista de visitats és un node de decisió. De ser així, incrementa en 1 el comptador de cada variable de decisió, apunta els passos que han estat necessaris per arribar-hi i apunta el nom del penúltim node visitat (node entremig)
- Si l'últim element de la llista no és un node de decisió, incrementa en 1 el comptador de variable possiblement inconclusiva.

Finalment, un tercer mètode realitza els càlculs finals de puntuació per a cada variable i defineix el veredicte que li pertoca. Amb aquesta informació construeix una taula de resultats amb el següent ordre de columnes:

- Nom
- Puntuació de discriminació positiva
- Mitjana de passos cap a una discriminació positiva
- Puntuació de discriminació negativa
- Mitjana de passos cap a una discriminació negativa
- Node entremig

- Puntuació inconclusiva
- Veredicte

4.8.2 Puntuació de la discriminació per *Page Rank*

Seguint la teoria plantejada a la secció 2.3.6, la nostra plataforma és capaç de computar les puntuacions de *Page Rank* per a cada individu en D. Per fer-ho, fem ús del mètode *personalized_pagerank*() de la llibreria *iGraph*.

Aquest mètode es crida per a cada fila del dataset. En cada crida, no es guarda tot el conjunt de puntuacions, sinó aquelles que corresponen als nodes de decisió. Els paràmetres passats al mètode són:

- ***directed*:** un booleà per indicar si es realitza *Page Rank* en un graf dirigit o no. En aquest cas, el paràmetre tindrà valor *True*.
- ***weights*:** un *string* per indicar si el graf conté arestes amb pes. En cas afirmatiu, el paràmetre s'ha de dir igual que l'atribut de l'aresta que representa el seu pes (en *iGraph* les arestes poden tenir diferents atributs).
- ***reset*:** el vector amb les probabilitats de saltar a un node aleatori quan el *random surfer* arriba a un node fulla. En el nostre cas, aquest vector és la fila de l'individu que estem puntuant, on cada valor està dividit pel nombre total de variables.
- ***implementation*:** un *string* per indicar la llibreria que es vol usar per resoldre el problema dels valors i vectors propis de *Page Rank*. Les més populars són *prpack* i *arpack*. En aquest cas usem *prpack* ja que és la més eficient.

El mètode retorna un vector indexat de puntuacions, així que abans de començar a iterar hem de saber l'índex dels nodes de decisions, de manera que crida rere crida ens guardem els valors desitjats.

Amb aquestes dues columnes valors i una tercera que representa la mostra d'una variable determinada (la que l'usuari ha triat per veure la diferenciació de puntuacions) es

crea un dataset que s'usa per crear dos gràfics els quals ajuden a entendre la distribució de les puntuacions [6].

4.9 Estructura de l'aplicació web

Un cop es va aconseguir tenir l'algoritme programat, perfilat i funcionant correctament, es va decidir implementar una plataforma web que agafés la informació de l'algoritme i la transformés en quelcom visual, fàcil d'entendre amb un simple cop d'ull, independentment dels grau de coneixement de la persona.

Per a desenvolupar-la es disposava de poc menys d'un mes. Es va optar per fer ús del *framework* Django 2.2 i Python 3.5.

De cara al *frontend* no es van usar cap tipus de plantilla feta, però si s'usen llibreries de CSS amb Bootstrap. Específicament, ens va triar per MDBootstrap [7], un *framework* amb aires de Material Design. En segon lloc, es va usar JavaScript per a la visualització de gràfics i taules.

Pel que respecte al *backend*, la única forma de informació a guardar és la següent:

- Fitxer "dataset": un fitxer ".csv" o ".txt" amb els requisits anteriorment esmentats 3.
- Fitxer d'ordre temporal: un fitxer ".csv" o ".txt" on es descriu amb un número enter l'ordre de cadascuna de les variables del dataset 4.5.1.
- Nom de la columna de decisió positiva
- Nom de la columna de decisió negativa

Així doncs, es va optar per usar una estructura Multipart/Form Data [8] guardada en una base de dades relacional gestionada per SQLite.

[6] Exemples d'aquests gràfics es poden veure a la secció d'experiments i resultats.
[7] https://mdbootstrap.com/
[8] http://fileformats.archiveteam.org/wiki/Multipart/Form-Data

L'estructura del codi de la pàgina web, per tant, segueix el següent esquelet:

- *Media*: carpeta que emmagatzema els fitxers que ha pujat l'usuari.
- *mysite*: carpeta que conté tots els arxius i codis propis de la pàgina web.
    1. *migrations*: carpeta que conté la informació de canvis produïts en els models interns de la pàgina.
    2. *SBNC*: carpeta que conté tot el codi de l'algoritme en si.
    3. *static*: carpeta amb tots els recursos d'estil de la pàgina web, com CSS, *JavaScript*, tipus de lletra, imatges, icones...
    4. *views*: arxiu de Python que coordina les crides web per retornar-ne una resposta.
    5. *validators*: arxiu de Python que usem per comprovar si el fitxer que l'usuari ens intenta pujar al sistema compleix les característiques necessàries.
    6. *models*: arxiu de Python on es descriuen els models de la web. En aquest cas, només un.
    7. *forms*: arxiu de Python on es defineixen els formularis presents en la web. En aquest cas, només un.
    8. *templates*: carpeta que conté el codi *HTML* de totes les planes de la web.
    9. *settings*: arxiu de Python amb les diferents configuracions de la instal·lació de Django i com funcionen.
    10. *urls*: arxiu de Python que rel·laciona cada URL de la pàgina amb la funció que la recull i coordina.
- *db.sqlite3*: base de dades SQLite de la pàgina.
- *manage*: fitxer de Python per on es coordinen les crides de comanda de l'aplicació,

com la inicialització del servidor o produir una migració.

La web permet la pujada i visualització de fitxers, pels quals tindrem dues opcions: eliminar-los o "analitzar-los", és a dir, executar l'algoritme i extreure'n els resultats. Els resultats són mostrats de la mateixa manera per tots els tipus de fitxers, i són els següents:

- Gràfic circular amb la classificació de les variables per tipus de discriminació.
- Taula amb la puntuació de discriminació de cada variable i el seu veredicte. Cada variable és un enllaç a una plana web amb l'anàlisi de les seves puntuacions de *Page Rank*:
    1. Gràfics de distribució: són dos, un on es mostra la distribució de la puntuació positiva de cada individu, separadament segons si forma part de la mostra de la variable concreta (un 1) o si no (un zero) i l'altre amb la distribució negativa organitzada de la mateixa manera.
    2. Gràfic de punts: amb la puntuació positiva a l'eix de les "x" i la negativa a l'eix de les "y". Cada punt és un individu, i prendrà un color o un altre segons si forma part de la mostra de la variable que s'estigui analitzant o no.
    3. Boxplot: són dos també, un amb el conjunt de les puntuacions positives i l'altre amb el de les negatives. Cada gràfic té dos caixes: una pels valors dels individus que formen part de la mostra de la variable que s'estigui analitzant i l'altre pels que no.
- Imatge del graf generat per la SBCN.
- Taula on es llisten les variables irrellevants per a l'anàlisi de la discriminació (nodes desconnectats).

4.9.1 *Deploy*

La pàgina web va ser posada en marxa a un servidor d'Amazon Web Services (AWS) que va proporcionar Eurecat. Es va optar per un servidor amb 2 nuclis i 8 GB de RAM, corrent amb sistema operatiu Linux de distribució Ubuntu 18.04. Aquest servidor té el llenguatge R instal·lat i tots els paquets de Python necessaris per executar la pàgina web. L'adreça d'aquest servidor (i de la pàgina web) és aquesta. No es recomana obrir-la des d'un *smartphone* o dispositiu similar.

Si es vol provar i testejar de manera més extensa l'aplicació web, en l'entrega del codi d'aquest treball es dona la possibilitat d'executar una versió local d'aquesta mateixa pàgina.

Aquesta pàgina web de moment té caràcter de "demo" i durant la correcció i presentació d'aquest treball estarà disponible temporalment per a provar l'execució de l'algoritme i visualitzar-ne els resultats. No està oberta, per tant, al públic, ni posseeix d'un gran ventall de solucions contra possibles *exploits* o atacs derivats.

4.9.2 Política de privacitat de les dades

Tal i com s'ha esmentat, la web de moment té caràcter de "demo" experimental i funcionarà durant el període de presentació d'aquest treball per a què un grup reduït de persones la puguin provar.

Així doncs, ni Eurecat ni la Universitat de Barcelona ni l'autor d'aquest treball es fa responsable del contingut i privacitat de les dades que s'hi pugen. Sempre que s'hi vulgui penjar un fitxer és farà sota la responsabilitat de qui ho fa.

D'igual manera, no es reclamarà propietat sobre cap contingut, text, dades, informació o fitxers que s'hi pengin, ni s'usaran en campanyes de *marketing*.

5 Experiments, Resultats i Discussió

Per a examinar el treball fet i posar a prova l'algoritme s'han fet multitud de tests amb datasets del món real. L'objectiu de testejar l'aplicació amb dades reals és poder contrastar el seu resultat, és a dir, veure sí l' *output* obtingut és correcte o concorda amb la realitat.

5.1 Datasets testejats

A continuació es presenten tots els datasets amb els que hem posat a prova l'aplicació. De tots aquests, ens centrarem amb els dos que millor ajuden a entendre la teoria i la implementació explicada. Alguns d'aquests són força populars mentre que d'altres segurament és la primera vegada que s'han adaptat per a aquests tipus d'anàlisis.

- ***Berkeley Admission Data*:** a la tardor de 1973 la Universitat de Berkeley va ser demandada per discriminar per gènere els candidats a entrar a la seva universitat. A la demanda s'hi adjuntava un total de 4.526 aplicacions amb 3 atributs: admissió (sí o no), gènere i departament (6 en total). L'ordre temporal és 1 pel gènere, 2 pel departament i 3 per les variables de decisió.

- ***Adult Dataset* ó *Census Income*:** popular dataset extret del *UCI Reporsitory of machine learning databases* [9], format per 48.842 registres. Les dades provenen de la Oficina del Cens dels Estats Units als anys 1994-1995. Per cada un d'ells es descriuen 10 atributs personals, com edat, raça, gènere, educació... El sou forma les dues variables de decisió.

- ***German Credit Data*:** 1.000 clients d'un banc els quals apliquen per un crèdit bancari. Les variables de decisió són si hi ha risc o no que el client pugui retornar l'import. De cada persona coneixem 21 variables, algunes personals (edat, gènere, feina, propòsit del crèdit...) i d'altres relacionades amb les seves característiques bancàries, com l'historial de crèdits o la seva nòmina. Prové també del

[9] http://archive.ics.uci.edu/ml.

UCI Reporsitory of machine learning databases.

- ***COMPAS Scores*:** algoritme usat per jutges i examinadors de llibertat condicional que serveix per mesurar el risc que una persona que ha delinquit ho torni a fer en un futur. El formen 6.173 individus que es descriuen amb atributs com el número de delictes, edat, raça i gènere.

- ***Huston College of Medicine Dataset*:** demanda contra el *Huston College of Medicine* on s'argumentava una discriminació de gènere contra les dones a l'hora de decidir a qui es promocionava o dictar salaris. Consta de 262 individus de la facultat dels quals sabem el seu gènere, especialitat, grau de responsabilitat, número de publicacions i salari, entre d'altres.

- ***Medical Cost in the US*:** aquest dataset prové d'un llibre que tracta sobre l'ensenyança de pràctiques de *Machine Learning* amb el llenguatge R [20]. Conté 1339 individus d'una asseguradora mèdica. Per a cada un d'ells coneixem atributs personals com l'edat, gènere, índex de massa corporal, si són fumadors o si tenen fills també inclosos en l'assegurança. El dataset és de domini públic, el podem trobar tant a Kaggle [10] com a GitHub [11].

- ***IBM Employee Attrition*:** dataset popular de Machine Learning on es simulen 1.471 treballadors de l'empresa IBM. D'ells sabem l'edat, el gènere, el departament on treballen, els anys que fa que no reben una promoció, els anys que porten treballant del mateix, si estan casats o no, si viatgen sovint, etc. La variable de decisió és si han marxat de la empresa o no.

D'aquests, ens enfocarem especialment amb el de COMPAS Scores i Huston College of Medicine Dataset. En primer lloc, han resultat funcionar bé davant la filosofia de l'algoritme (pocs resultats inconclusius). En segon lloc, es divideixen les seves variables en diversos tipus de discriminació, estan més repartides que en d'altres datasets. Final-

[10] https://www.kaggle.com/mirichoi0218/insurance
[11] https://github.com/stedy/Machine-Learning-with-R-datasets

ment, el graf generat és fàcil d'entendre, no és excessivament gran i les relacions entre els nodes s'intueixen de manera senzilla. De totes maneres, també esmentarem petites curiositats dels resultats dels altres datasets.

Totes les proves realitzades s'han produït amb 1000 iteracions de l'algoritme *random walk*, un número ni força alt ni força baix i que no compromet a gran escala el temps d'execució de l'algoritme. El número d'iteracions no produeix grans diferències en els resultats, demostrant així la solidesa de l'algoritme:

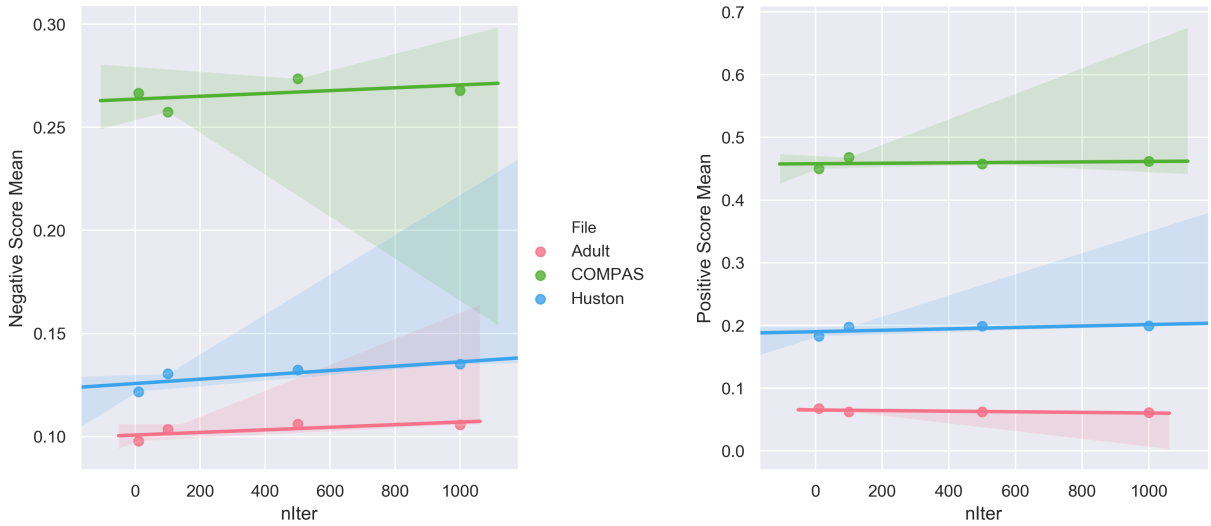


Figura 5: Figura mostrant la mitjana de puntuacions positiva i negativa per cada fitxer, amb número d'iteracions 10, 100, 500 i 1000

Els paràmetres per establir un veredicte són els esmentats en la secció 2.3.3 i 2.3.4, és a dir, 0,55 per la puntuació inconclusiva, 0,25 per als veredictes aparents i 0,06 per als veredictes neutrals.

- **Berkeley Admission Data:** aquesta demanda s'usa molt sovint per explicar el fenomen de la paradoxa de Simpson, el qual succeeix quan les dades "juntes" et diuen una cosa diferent de quan estan separades. En aquest cas, quan es mira el número d'admissions de les dones en els departaments, es veu molt més baix que el dels homes (d'aquí la raó de la demanda). Però quan es mira cada departament desglossat i desenvolupes el rati d'admissió desagregat per sexe, les dades

et diuen tot el contrari: les dones tenien lleugerament millors ratis d'admissió!

El nostre algoritme no cau en paradoxes de Simpson, i tal i com podem veure en la figura 6, no hi ha cap aresta que connecti el node de gènere femení amb la decisió negativa *Admit_Rejected*. Senzillament, les dones tendeixen més a aplicar a departaments amb rati d'entrada més alt (departaments C, D, E i F), mentre que els homes als de ratis més baixos (A i B).

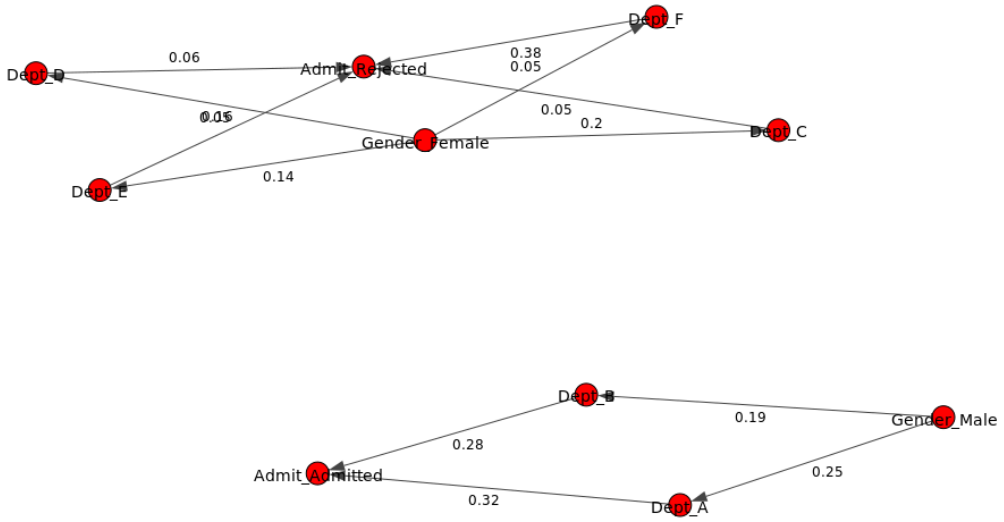


Figura 6: DAG resultant del dataset *Berkeley Admission Data*

- ***Adult Dataset*:** en analitzar aquest dataset veiem que les persones que menys diners cobren són les de menys edat i les de menys educació. De totes maneres, de 90 variables 79 són inconclusives, així que tampoc és el dataset més adequat per usar el nostre algoritme. A la taula 1 podem veure 10 primeres variables ordenades per puntuació negativa.

Name	Positive Score	Avg. Positive Steps	Negative Score	Avg. Negative Steps	Intermediate Node	Inconclusive Score	Veredict
occupation_Other_service	0.000	0.00	1.000	1.00	None	0.000	Negative Discrimination
marital_status_Never_married	0.004	2.25	0.987	1.21	None	0.009	Negative Discrimination
relationship_Own_child	0.006	3.67	0.936	2.23	marital_status_Never_married: 76.86% of times	0.058	Explainable/Conditional discrimination because of marital_status_Never_married
age_young	0.003	4.33	0.750	2.43	marital_status_Never_married: 62.68% of times	0.247	Explainable/Conditional discrimination because of marital_status_Never_married
education_11th	0.005	4.20	0.740	2.78	marital_status_Never_married: 50.20% of times	0.255	Explainable/Conditional discrimination because of marital_status_Never_married
relationship_Other_relative	0.005	3.60	0.736	2.21	marital_status_Never_married: 70.31% of times	0.259	Explainable/Conditional discrimination because of marital_status_Never_married
relationship_Not_in_family	0.050	2.42	0.567	1.76	marital_status_Never_married: 44.25% of times	0.383	Apparent Negative Discrimination
education_10th	0.000	0.00	0.519	2.00	occupation_Other_service: 100% of times	0.481	Apparent Negative Discrimination
race_Black	0.003	4.00	0.464	2.46	marital_status_Never_married: 57.17% of times	0.533	Apparent Negative Discrimination
education_Some_college	0.001	4.00	0.446	3.19	marital_status_Never_married: 80.76% of times	0.553	Too many inconclusive outputs

Taula 1: Taula de resultats del dataset *Adult*, ordenada per puntuació negativa

L'algoritme de *random walk* ens dona sovint una puntuació negativa de 0,47 per als individus de raça negre, i una puntuació inconclusiva de 0,52. En aquest cas, les puntuacions de *Page Rank* ens treuen de dubtes, ja que tal i com podem observar en la figura 7, les puntuacions no són dispars.

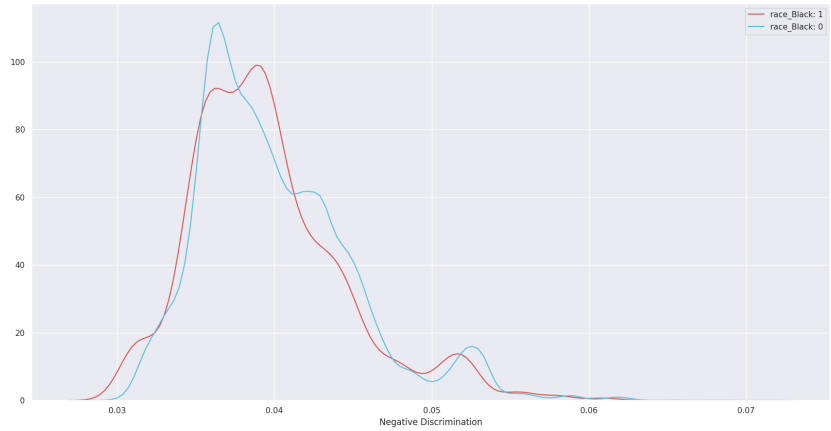


Figura 7: Puntuacions de *Page Rank* negatives pel subgrup de raça negre en el dataset *Adult*

- ***German Credit Data*:** aquest dataset no ha resultat molt adient al nostre mètode de detecció de discriminació ja que de 71 variables, 69 són inconclusives. Ho podem veure representat en la figura 8.

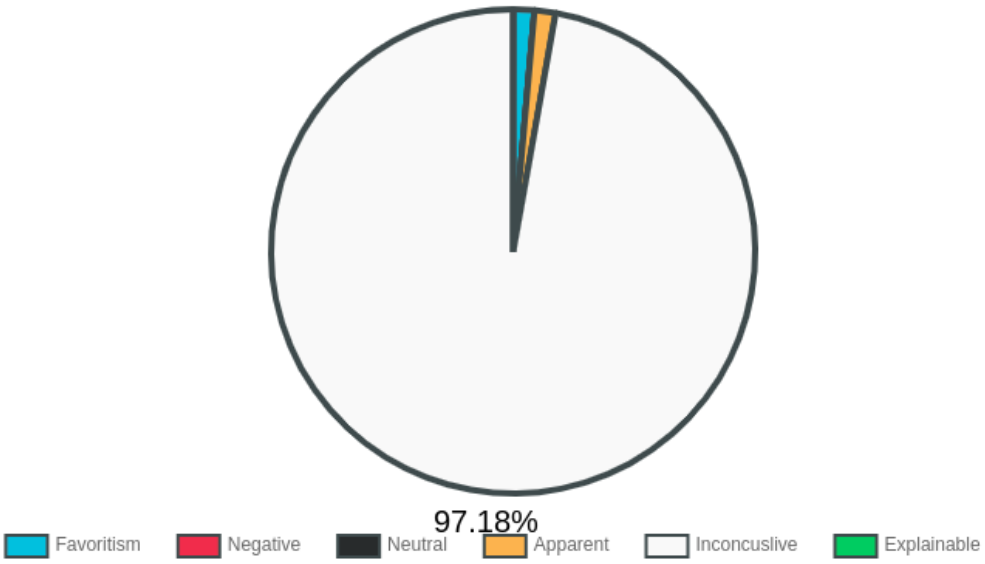


Figura 8: Presència absoluta de classificació inconclusiva en el dataset *German*

Només hi ha un únic subgrup de persones les quals tenen favoritisme aparent a l'hora d'obtenir un crèdit, i són aquelles persones les quals no tenen cap compte obert al banc (nous clients). Amb aquesta variable el 62% de les vegades arribem a una decisió positiva i un 37% a un node fulla. La figura 9 mostra els individus d'aquest subgrup i les seves puntuacions de *Page Rank*.

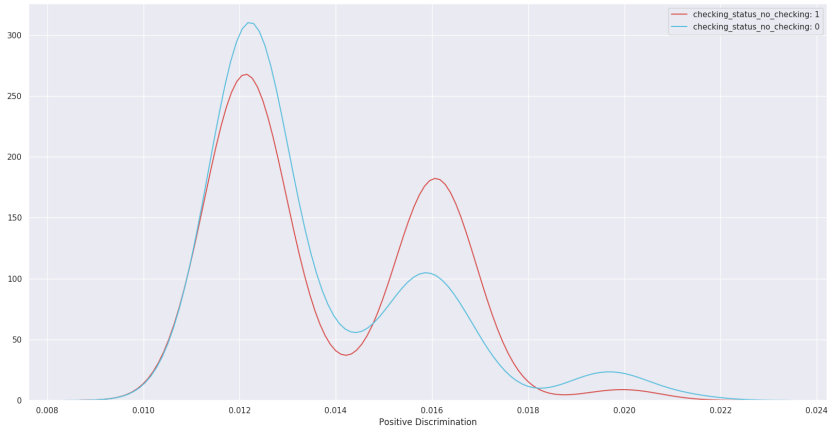


Figura 9: En vermell, puntuacions de *Page Rank* positives pel subgrup de clients nous en el dataset *German*

Així doncs, direm que el banc du a terme una pràctica de favoritisme cap als seus nous clients a l'hora de concedir un crèdit, el que segurament sigui una estratègia de *marketing*.

- ***Medical Cost in the US*:** dataset format, citant textualment el llibre, per "*un conjunt de dades simulat que conté despeses mèdiques per a pacients als Estats Units* [...] *Aquestes dades es van crear per aquest llibre usant estadístiques demogràfiques de l'Oficina de Cens dels Estats Units, i, per tant, reflecteixen aproximadament les condicions del món real*". El dataset és de domini públic, el podem trobar tant a Kaggle [12] com a GitHub [13].

  Aquest dataset és força útil per explicar la discriminació condicionada o explicable, ja que s'examina que els grups més discriminats són els homes i els fumadors. Quan examinem el node entremig ens donem compte que els homes no estan sent discriminats, sinó que la gran majoria d'homes fumen! En la figura 10 podem observar puntuacions similars de *Page Rank* entre ambdós generes.

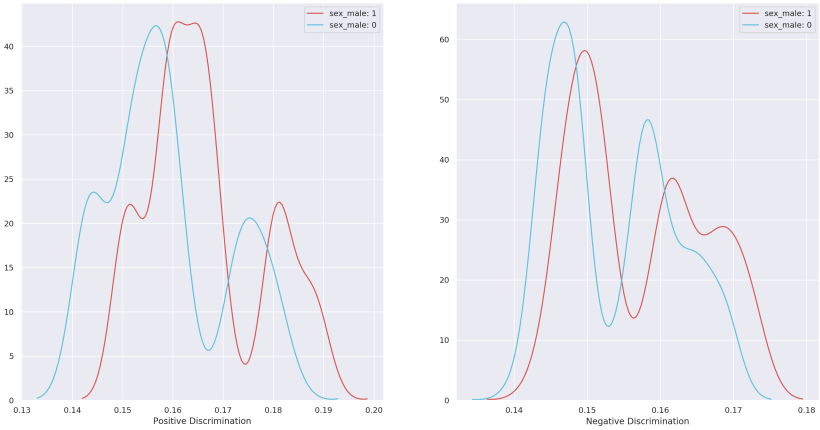


Figura 10: A l'esquerra, puntuacions positives i a la dreta, puntuacions negatives, diferenciades per gènere. Dataset *Medical Costs US*

En resum, les despeses són més grans per als clients fumadors. També s'observen

[12] https://www.kaggle.com/mirichoi0218/insurance
[13] https://github.com/stedy/Machine-Learning-with-R-datasets

valors alts de puntuació negativa a la gent gran, fet que seria explicable degut a què com més grans s'és, més necessitat mèdica hi ha. Ho comprovem amb les puntuacions de *Page Rank* a la figura 11.

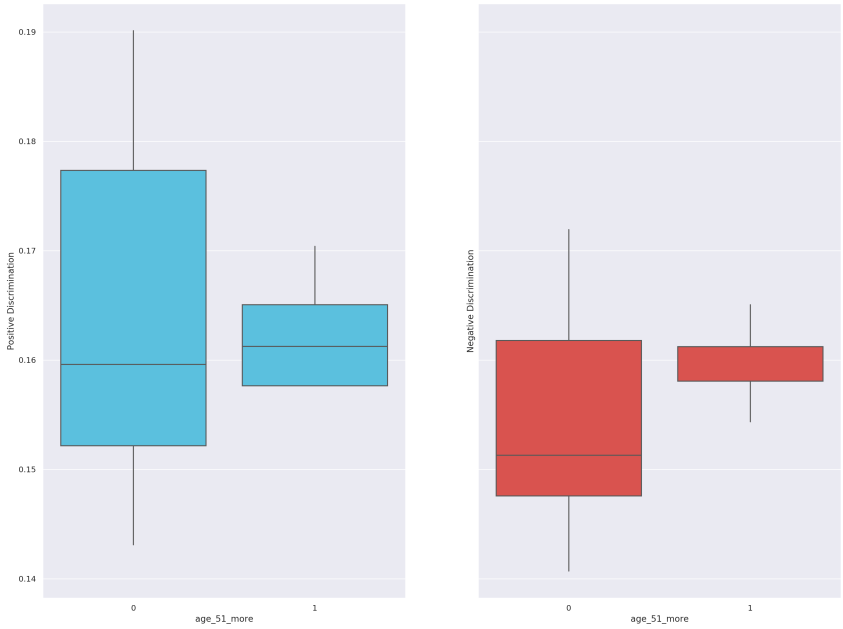


Figura 11: Les persones de més edat tenen menys puntuació positiva i més puntuació negativa. Dataset *Medical Costs US*

- ***IBM Employee Attrition*:** en aquest dataset s'observa també moltes variables inconclusives, però de negativa en tenim una de clara: *OverTime_Yes*. Pel que sembla, el personal que fa hores extres és la mes propensa a anar-se'n de l'empresa. De totes maneres, mirant la figura 12, amb el mètode de puntuació *Page Rank* no observem discriminació.

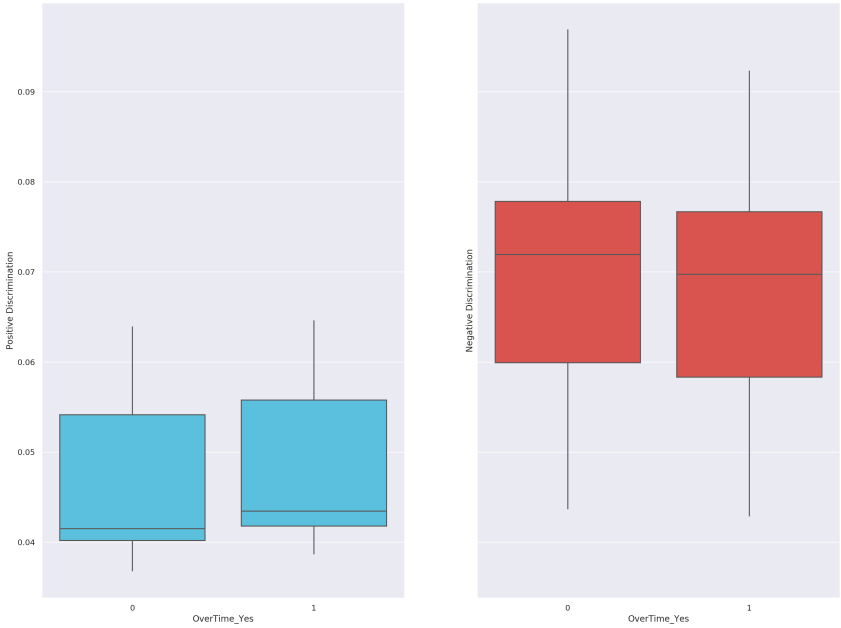


Figura 12: Boxplots del subgrup Overtime. Per ambdós tipus de puntuació, no observem diferència de valors. Dataset *IBM Employee Attrition*

5.1.1 *COMPAS Scores*

Aquest dataset mereix una atenció especial ja que és un clar exemple d'allò que es parla a la introducció: avui en dia hi ha algoritmes que dicten el nostre destí i sovint no estan controlats per cap organisme ni llei. COMPAS, o més ben dit, *Correctional Offender Management Profiling for Alternative Sanctions* avui en dia és usat als estats de Nova York, Wisconsin i California.

ProRepublica, una organització sense lucre de Nova York dedicada al periodisme transparent i neutra, va publicar al 2016 un extens estudi sobre COMPAS [14] [15] on comprovava que "les persones de raça negra tenien gairebé el doble de probabilitat de ser etiquetades amb un risc elevat respecte a les persones de raça blanca" i que "només el 20% de

[14]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[15]https://github.com/propublica/compas-analysis

les persones per a les quals predeia que cometrien un delicte violent el van cometre".

En aquest treball s'han agafat les dades d'aquest estudi, les quals contenen puntuacions recollides durant els anys 2013 i 2014. Són un total de 6.173 puntuacions, les quals venen acompanyades amb tot tipus d'informació que ens és irrellevant, com el delicte comés, el número de sentència, la data d'arrest, etc.

Per realitzar el nostre propi estudi, s'han seleccionat les variables de número d'antecedents (alt, mitjà i baix), edat (menys de 25 anys, entre 25 i 45 i major de 45), raça (asiàtica, afroamericana, llatina, caucàsica i "americans nadius"), gènere i gravetat del delicte, el qual pot ser lleu o greu.

Els atributs de edat, raça i gènere tenen nivell 1 a la taula d'ordre temporal. El número de d'antecedents i el tipus de delicte, tenen ordre 2. Finalment, les variables de decisió tenen ordre 3.

Després de computar les dades obtenim el següent graf:

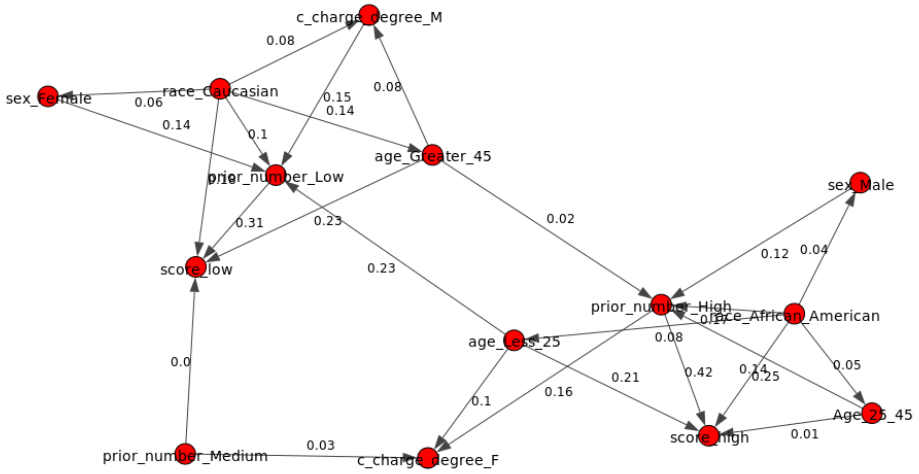


Figura 13: DAG resultant de la SBCN pel dataset *COMPAS Scores*

Que després de ser explotat amb *random walk* ens proporciona la següent taula:

Name	Positive Score	Avg. Positive Steps	Negative Score	Avg. Negative Steps	Intermediate Node	Inconclusive Score	Veredict
race_African_American	0.044	3.00	0.791	1.57	None	0.165	Negative Discrimination
Age_25_45	0.000	0.00	0.748	1.92	prior_number_High: 91.84% of times	0.252	Explainable/Conditional discrimination because of prior_number_High
prior_number_High	0.000	0.00	0.739	1.00	None	0.261	Negative Discrimination
sex_Male	0.000	0.00	0.728	2.00	prior_number_High: 100% of times	0.272	Explainable/Conditional discrimination because of prior_number_High
age_Less_25	0.412	2.00	0.208	1.00	prior_number_Low: 66.45% of times	0.380	Apparent Favoritism
age_Greater_45	0.943	1.57	0.002	2.00	None	0.055	Favoritism
race_Caucasian	0.917	1.92	0.001	3.00	prior_number_Low: 47.28% of times	0.082	Favoritism
prior_number_Low	1.000	1.00	0.000	0.00	None	0.000	Favoritism
prior_number_Medium	0.153	1.00	0.000	0.00	None	0.847	Too many inconclusive outputs
c_charge_degree_F	0.000	0.00	0.000	0.00	-	1.000	Too many inconclusive outputs

Taula 2: Taula de resultats del dataset COMPAS Scores, ordenada per puntuació negativa

On podem veure en la figura 14 que hi ha gairebé tots els possibles tipus de discriminació.

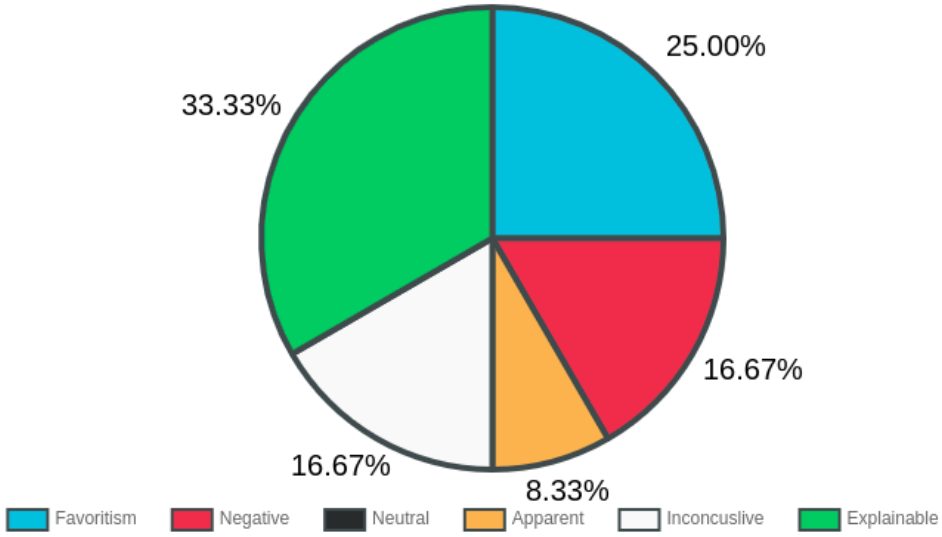


Figura 14: Classificació de les variables del dataset *COMPAS Scores*

A la taula no es veuen dues variables, <*prior_number_Medium*> i <*charge_degree_F*>, les quals tenen una puntuació inconclusiva alta. De la resta, en podem dir que hi ha una alta discriminació cap a les persones de raça afroamericana, ja que un 74,3% dels camins acaben en decisió negativa, l'altre 20,3% en un node fulla i finalment un 5,4% en una decisió positiva.

Un altre indicador absolut de quan es tindrà una puntuació negativa alta és el número d'antecedents. Si és alt, COMPAS dicta que hi ha una alta probabilitat que es torni a

delinquir. Ho veiem també amb les puntuacions de *Page Rank* en la figura 15 i 16, tan en forma de *boxplot* com en forma de distribució de les puntuacions.

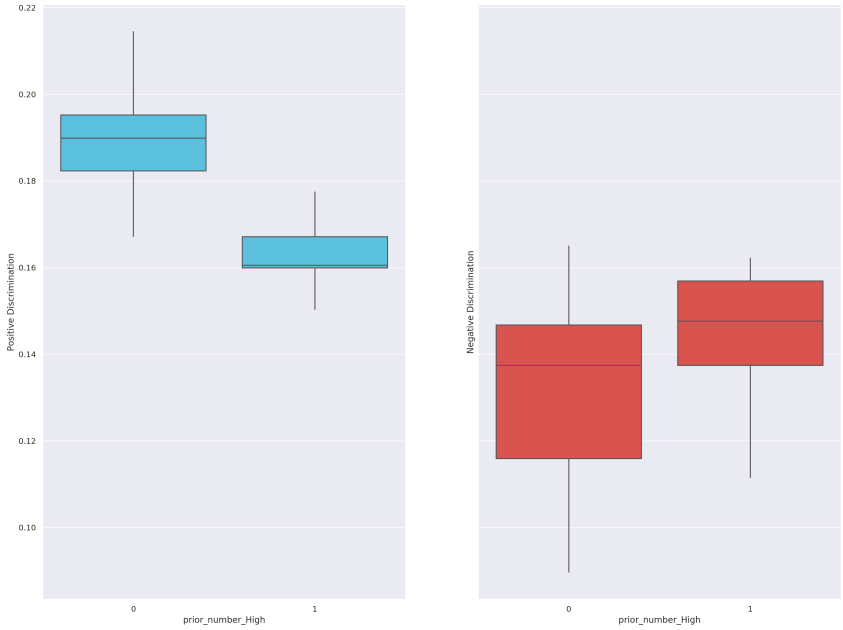


Figura 15: A la dreta, puntuacions negatives més grans per a la gent amb molt antecedents. A l'esquerra, puntuacions positives més baixes pel mateix subgrup. Dataset *COMPAS Scores*

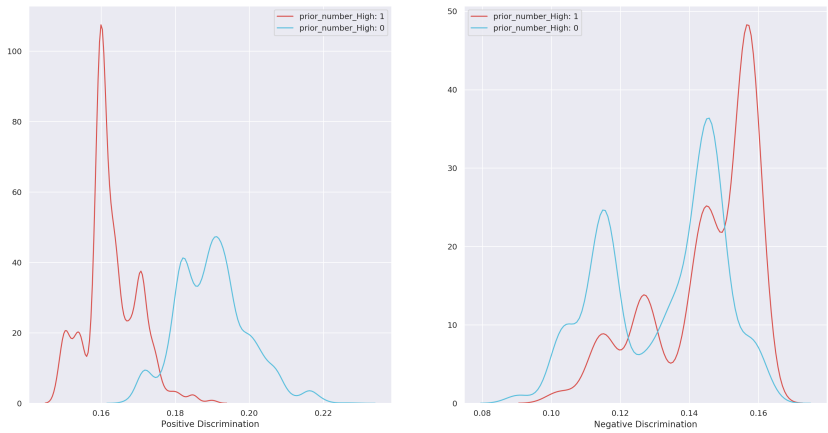


Figura 16: En vermell, individus amb nombre alt d'antecedents. Puntuacions negatives més altes i positives més baixes. Dataset *COMPAS Scores*

També veiem una paradoxa de Simpson amb el gènere masculí, ja que la gran majoria d'homes tendeix a tenir un número elevat d'antecedents, i, per tant, els homes tendeixen a tenir puntuacions altes. A l'inrevés succeeix amb les dones.

Ara bé, seguint amb les races, s'hi observa que el 92% dels homes de raça caucasiana obté puntuacions baixes, és a dir, que estan gaudint de favoritisme. Tot i això, tenim un indicador que ens fa sospitar que potser no és realment així: el de node entremig. Específicament, tenim que un 47% de les vegades passem pel node de pocs antecedents. Per sortir de dubtes, mirem les puntuacions de *Page Rank* en la figura 17.

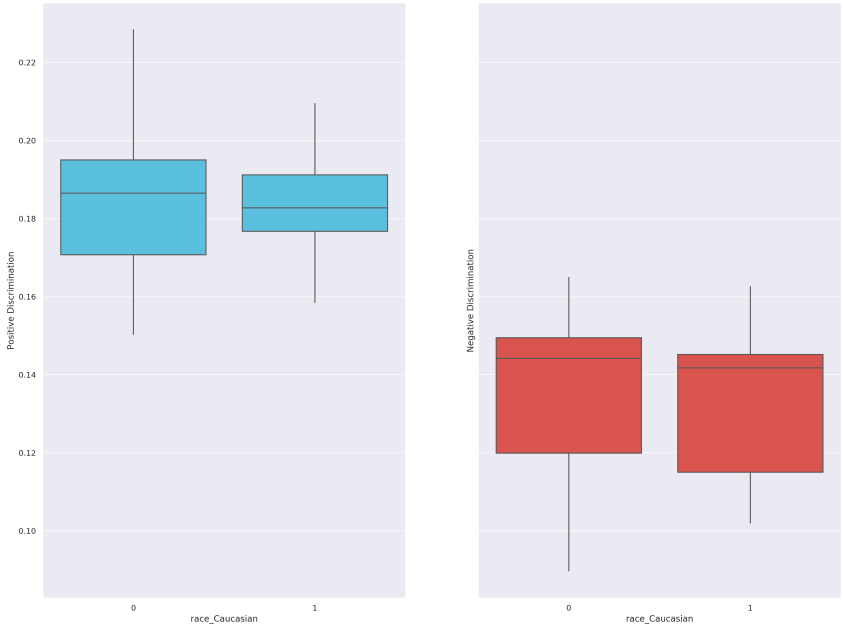


Figura 17: Similitud de puntuacions tan positives com negatives per als individus de raça caucàsica i els que no. Dataset *COMPAS Scores*

On observem que les puntuacions positives i negatives estan relativament igualades. D'on prové aquesta marcada diferència entre ambdues puntuacions? Bé, tot i que ambdós mètodes es basen en la mateixa filosofia, a *Page Rank* es produeixen salts aleatoris produïts pel *random surfer* però que no succeeixen al *random walk*. Així doncs, el favoritisme existeix (hi ha una aresta que connecta <*race_caucassian*> amb <*score_low*>) però no és tan fort com podia semblar.

5.1.2 *Huston College of Medicine Dataset*

La idea d'usar aquestes dades es va treure del projecte *Gender Discrimination* de la pàgina web de la fundació XPRIZE [16], una organització sense ànim de lucre que crea i dissenya concursos públics que fomenten algun desenvolupament tecnològic que pugui beneficiar la humanitat.

En la descripció del projecte s'esmenta que els demandants argumentaven que "*el col·legi s'havia dedicat a una pauta i pràctica de discriminació de gènere en contra de les dones a l'hora d'establir sous i donar promocions*". Per demostrar-ho, els demandants havien adjuntat el dataset que analitzarem a continuació.

- Departament: categoritzat amb els noms de Bioquímica, Medicina, Cirurgia, Pediatria, Psicologia i Genètica. Degut a la manca d'informació, no sabem quina diferencia hi ha entre la gent dels departaments Medicina, Cirurgia i Pediatria.
- Rang: categoritzat amb els noms de *Assistant*, *Associate* i *Full Professor*.
- Gènere: categoritzat segon home o done.
- Èmfasis primari: el qual pot ser "Clínic" o de "Recerca".
- Certificat: el tipus de llicència mèdica de la persona en qüestió. Si es posseeix és perquè s'ha superat un examen i la junta [17] et reconeix la llicència. Els que no, poden practicar la medicina en un grau menor.
- Nivell d'experiència: anys que han passat des de l'obtenció del títol. S'ha dividit en "baix", "mitjà", "alt" i "molt alt". La mostra de la variable original té una mediana de 9 anys, el primer quartil (25%) es troba en 6 anys i el tercer quartil (75%) es troba en 14 anys.
- Rati d'articles publicats: la mitjana d'articles publicats per any. La variable té

[16] https://data.xprize.org/selinamyers4people/5d2380b78b034012184d2948
[17] https://www.abim.org/

una mediana de 4,4 d'articles, un primer quartil en 3,2 articles i finalment un tercer quartil en 6,9 articles. Per tant, s'ha categoritzat la variable com rati "baix", "mitjà", "alt" i "molt alt".

- Diferència de salari: la resta entre el salari a l'any 1995 i el salari a l'any 1994, ja que tots els individus presenten un augment de salari. La mediana de l'augment de salari està en 12.955$, així que la divisió per a què es converteixi en dues variables categòriques de decisió es fa amb aquest nombre: <*augment_baix*> i <*augment_alt*>.

El gènere és l'única variable amb ordre 1 a la taula d'ordre temporal. La resta, tenen ordre dos, excepte els nodes de decisió que tenen ordre 3. Així doncs, la classificació de les variables és la següent:

Name	Positive Score	Avg. Positive Steps	Negative Score	Avg. Negative Steps	Intermediate Node	Inconclusive Score	Veredict
Research_emphasis	0.000	0.0	1.000	1.00	None	0.000	Negative Discrimination
Board_certified_no	0.000	0.0	0.796	1.00	None	0.204	Negative Discrimination
PublicationRate_High	0.000	0.0	0.725	1.43	None	0.275	Negative Discrimination
PublicationRate_Very_High	0.000	0.0	0.561	2.00	Research_emphasis: 75.58% of times	0.439	Apparent Negative Discrimination
PublicationRate_Low	0.836	2.0	0.000	0.00	Dept_Surgery: 62.20% of times	0.164	Explainable/Conditional favoritism because of Dept_Surgery
PublicationRate_Medium	0.697	2.0	0.000	0.00	Dept_Medicine: 63.41% of times	0.303	Explainable/Conditional favoritism because of Dept_Medicine
years_since_MD_Low	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs
years_since_MD_Medium	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs
years_since_MD_High	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs
years_since_MD_Very_High	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs

Taula 3: Taula de resultats del dataset Huston College of Medicine, ordenada per puntuació negativa

On s'hi pot observar que els grups més discriminats són els que es dediquen a la recerca i els que no estan certificats per la junta. En canvi, els més beneficiats són els individus que treballen al departament de medicina i cirurgia. Aquest primer anàlisi sembla correcte:

- Es probable que als Estats Units ser cirurgià es pagui molt millor que dedicar-se a la recerca en disciplines bàsiques.

- Tenir menys qualificació (no posseir totes les llicències) implica tenir menys sou.

Ara bé, els valors de favoritisme relativament alts en aquella gent que té un baix rati d'articles publicats com s'explica? Bé, ho podem veure a través del graf de la figura 18,

on no hi ha cap aresta que connecti la variable de rati de publicació baix o mitjà amb la variable de decisió positiva.

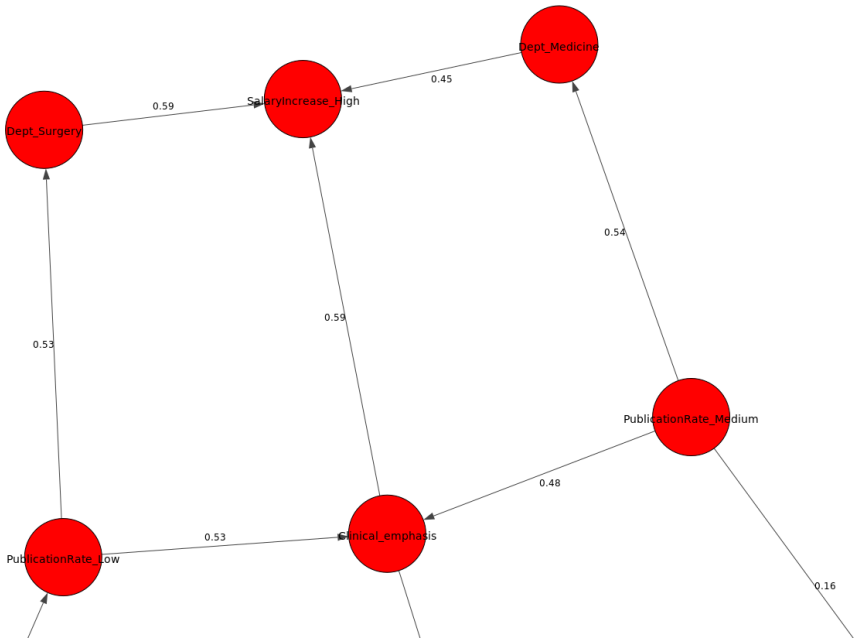


Figura 18: Part del DAG resultant del dataset *Huston College of Medicine*. No s'observa cap aresta entre *publication rate low* i *salary increase high*

És el cas d'una altre paradoxa de Simpson: la majoria de persones amb baix rati de publicacions són del departament de cirurgia o medicina, la qual cobra més diners. Al revés hi pot haver una mica més de discussió, ja que observant la figura 18 veiem que sí que hi ha una aresta que connecta el node de publicacions alt (no "molt alt") amb el node de decisió negativa.

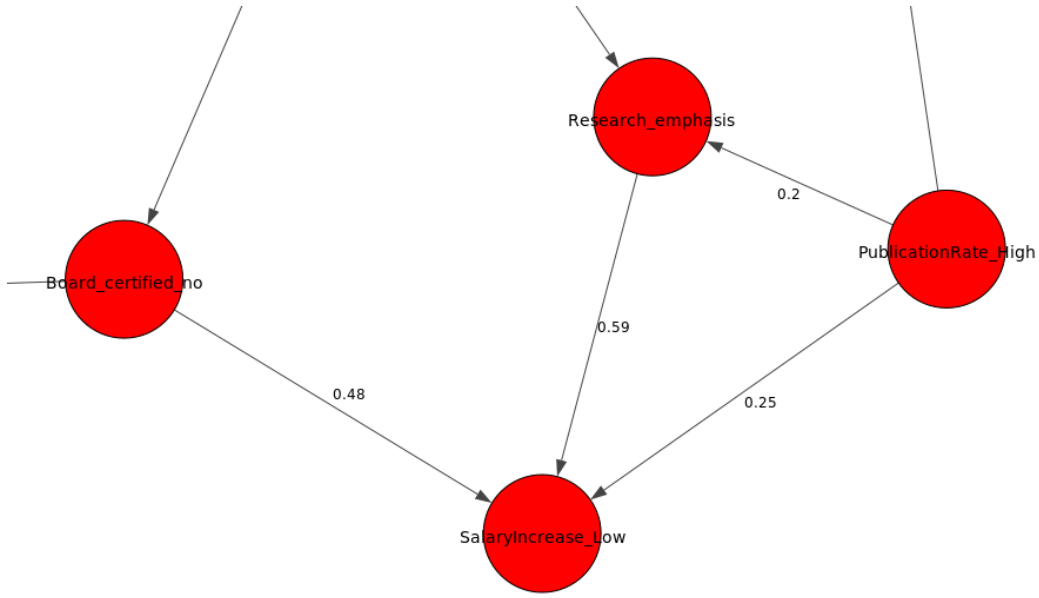


Figura 19: Part del DAG resultant del dataset *Huston College of Medicine*. Observem l'existencia d'una aresta entre *publication rate high* i *salary increase low*

Així que mirem les puntuacions de *Page Rank* en la figura 20 i observem que la diferència no és tan gran.

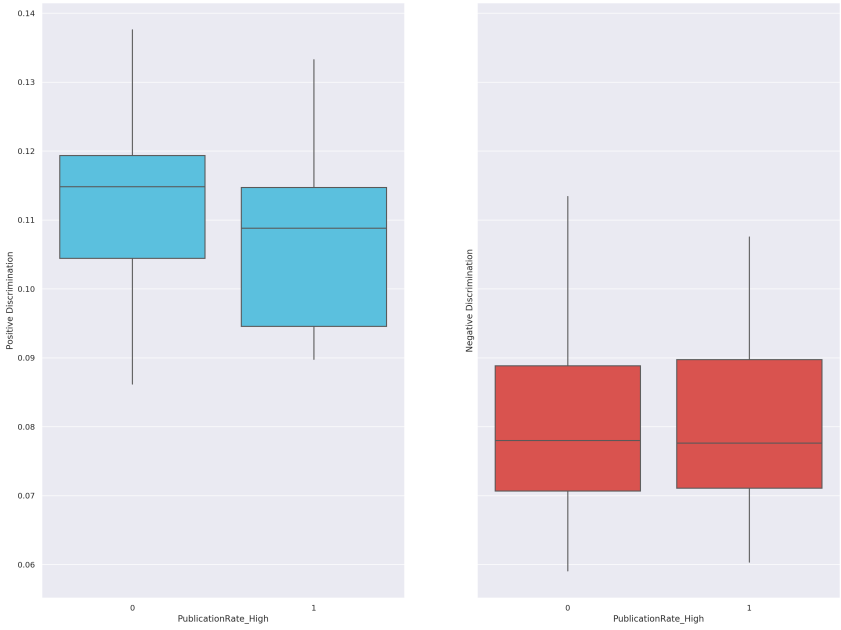


Figura 20: Lleugera diferència de puntuacions positives entre els individus amb un alt nombre d'articles publicats i els indivius que no. Dataset *Huston College of Medicine*

Simplement, la gent amb alt rati d'articles publicats tendeix a tenir èmfasis de recerca, variable que ja hem observat que dictamina cobrar menys.

Sigui com sigui, no hi ha cap indici que ens faci creure que la Universitat estigui discriminant a les dones. De fet, mirant la figura 21, podríem pensar que succeeix al inrevés.

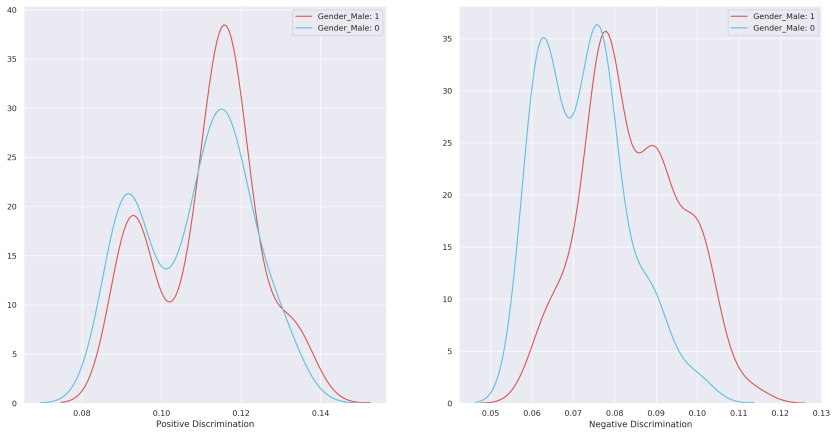


Figura 21: Inexistència de discriminació per gènere. Dataset *Huston College of Medicine*

5.2 Rendiment

Per valorar el rendiment de l'algoritme hem decidit usar la suma de les mides del fitxer dataset i ordre temporal i el temps que triga en computar-lo. Aquest temps es comença a comptar des de la lectura dels datasets fins el càlcul de la valoració de la discriminació per *random walk*.

Mesurar el rendiment de l'algoritme és un afer complicat ja que en la seva execució hi ha un procés d'aprenentatge parcialment aleatori, l'anomenat *Hill Climbing*. Tal i com s'ha descrit en la secció 3.3.2, l'obtenció del millor model es pot donar en un màxim local o al arribar al màxim d'iteracions permeses. En conseqüència, independentment de la mida del fitxer, aquest procés pot trigar més o menys.

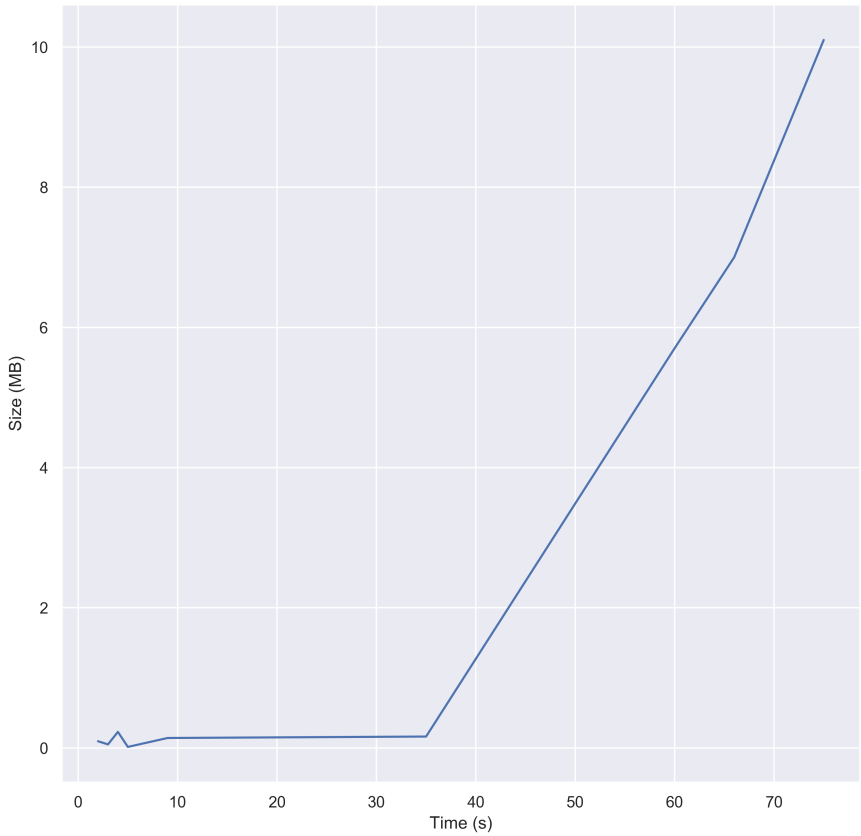


Figura 22: Temps d'execució (segons) en relació a la mida del fitxer (MB)

Un altre problema d'aquesta mètrica és que no es posseeix de datasets amb pesos variats, tal i com es pot veure en la figura 22. Exceptuant els datasets d'Adult i German, els datasets esmentats tenen poques files i columnes, i, per tant, el seu pes és baix.

Concretament, la seva mitjana de pes és de 0,1 MB, mentre que Adult i German tenen 5,13 MB. a seva mitjana de temps també canvia dràsticament: 4,6 segons pels datasets petits i 55 segons per Adult i German.

6 Conclusions

En aquest treball s'ha aconseguit adaptar al llenguatge de programació Python l'algoritme per a la creació de Xarxes Causals de Suppes-Bayes de l'article "*Exposing the probabilistic causal structure of discrimination*" [1] el qual estava una mica desordenat i fragmentat.

En segon lloc, després de realitzar una anàlisi sobre els tipus de discriminació i la manera en què es manifesta, s'ha pogut ampliar la informació extreta d'aquest algoritme per a la puntuació de la discriminació. Específicament, s'introdueix i es codifica el concepte de puntuació inconclusiva que ajuda a definir el grau de discriminació.

Finalment, no només s'ha exportat l'algoritme a Python sinó que s'ha aconseguit la seva execució en un servidor que a través d'una aplicació web programada des de zero ens mostra els seus resultats. Concretament, podem visualitzar la informació extreta amb elements gràfics i didàctics que ajuden a entendre-la independentment del nivell de coneixement sobre la matèria el tipus i grau de discriminació.

Ara bé:

- Part del codi adaptat a Python requereix encara d'una biblioteca existent únicament per R, fet que provoca un endarreriment en l'execució global de l'algoritme.
- La introducció del concepte de puntuació inconclusiva quan sovint s'arriba a un node fulla obre un nou debat sobre si el mètode per puntuar la discriminació és ineficaç o no.
- No s'ha pogut aconseguir una correlació absoluta entre els resultats de la valoració de discriminació per *random walk* i *Page Rank*.
- Tot l'algoritme gira al voltant de variables de decisió, les quals no sempre són dues, no sempre són fàcils de trobar o inclús és possible que no hi siguin en tots els entorns.

7 Futures Millores

Aquest projecte s'ha dut a terme amb la col·laboració d'Eurecat i per part d'aquesta institució seguirà endavant, millorant la plana web i l'eficàcia de l'algoritme. De fet, s'ha posat sobre la taula la possibilitat de realitzar una "demo" més completa per presentar-la a la edició del 2020 de l'*European Conference on Machine Learning Principles and Practice of Knowloedge Discovery in Databases* o PDKK, la qual tindrà lloc a Bèlgica. Així doncs, diverses propostes de millora s'han comentat:

- La eliminació de la taula d'ordre temporal. En l'article original ja ho van aconseguir, és a dir, s'hi defineix una funció que és capaç d'intuir quin és l'ordre temporal de les variables. Significaria una càrrega menys per l'usuari però s'haurien de realitzar multitud de tests per assegurar que les valoracions són similars.

- La possibilitat d'ajuntar dues variables i re-calcular la discriminació. És a dir, unir <*sexe_masculí*> amb <*fumador_si*>, creant <*sexe_masculí_i_fumador*>. A la vegada, provocaria també la creació del seu oposat, <*sexe_no_masculí_o_no_fumador*>. Amb aquesta característica, es podria obtenir una millor descripció de la valoració de la discriminació amb nous resultats més determinants.

- El dos de Gener de 2020 va sortir la primera *release* de la lliberia bnlearn per a Python [18], la qual en aquest treball s'executa cridant codi d'R des de Python. No obstant, aquestes primeres versions de la llibreria per a Python son un pèl diferents a la original, canvien paràmetres i sub-llibreries usades per fer el còmputs. Es va provar de fer el canvi però el rendiment no millorava i no hi havia temps de jugar amb els paràmetres. En un futur, de ben segur que es podrà replicar els resultats amb molta exactitud i millorant el rendiment pel fet de ser tot executat des de Python.

- La detecció automàtica de llindars a l'hora de valorar la discriminació. A dia d'a-

[18] https://github.com/erdogant/bnlearn

vui s'han de canviar expressament, ja el sistema usa uns valors per defecte. Amb certs càlculs, i dins d'un rang, es podria determinar el llindar més adequat en funció de les dades.

- L'ús d'una llibreria diferent a iGraph pel tractat de grafs. Té algun que altre problema per instal·lar-se en segons quin sistema operatiu i els seus mètodes de dibuix de graf no són especialment visuals o bonics. Una alternativa podria ser NetworkX[19].

- De cara a l'aplicació web, i tenint en compte que ara mateix és una petita "demo" funcional, es podrien plantejar diverses millores, com l'ús d'una altre llibreria diferent a Pandas que sigui *thread-safe* per la visualització de les puntuacions de *Page Rank*, l'addició d'una barra de progrés en el procés d'anàlisi, adaptar la pàgina web per a *smartphones* i similars o executar l'anàlisi d'errors en els fitxers durant el procés de pujada i no quan es prem el botó d'analitzar.

[19] https://networkx.github.io/

Bibliografia

[1] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21, Feb 2017.

[2] Ramanuja Simha and Hagit Shatkay. Protein (multi-)location prediction: Using location inter-dependencies in a probabilistic framework. In Aaron Darling and Jens Stoye, editors, *Algorithms in Bioinformatics*, pages 3–17, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[3] John R. Mashey. In *Big Data and the Next Wave of InfraStress*. in 1999 USENIX Technical Conference, Monterey, 1998.

[4] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016.

[5] Cambridge Dictionaries Online. Discrimination, definition. Cambridge University.

[6] Gran diccionari de la llengua catalana. Favoritsime. Enciclopèdia.cat.

[7] Tyler Vigen. Spurious correlations. *consultat el 21 de Novembre de 2019*.

[8] Patrick Suppes. *A probabilistic theory of causality*. North-Holland, 1970.

[9] Christopher Hitchcock. *Probabilistic Causation*. The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), 2018.

[10] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[11] Lu Zhang, Yongkai Wu, and Xintao Wu. Situation testing-based discrimination discovery: A causal inference approach. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 2718–2724, 2016.

[12] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[13] Y. Wu and X. Wu. Using loglinear model for discrimination discovery and prevention. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 110–119, Oct 2016.

[14] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining*, pages 992–1001, 2011.

[15] B. T. Luong, S. Ruggieri, and F. Turini. Classification rule mining supported by ontology for discrimination discovery. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 868–875, 2016.

[16] Leise T. Bryan, K. The 25,000,000,000 dollar eigenvector: The linear algebra behind google. *International Journal of Geographical Information Science*, (3):569–581, 2006.

[17] Ron; Ball Sarah Zack, Laurie; Lamb. An application of google's pagerank to nfl rankings. *Involve, A Journal of Mathematics*, (4):463–471, 2012.

[18] B. Jiang. Ranking spaces for predicting human movement in an urban environment. *International Journal of Geographical Information Science*, (7):823–837, 2006.

[19] Marco Taboga. In *Lectures on Probability Theory and Mathematical Statistics*. CreateSpace Independent Publishing Platform, 2017.

[20] B. Lantz. *Machine Learning with R: Expert Techniques for Predictive Modeling, 3rd Edition*. Packt Publishing, 2019.

A *Appèndix*: Captures de la pàgina web

A continuació s'exposen figures amb les captures de pantalla de les planes de la pàgina web més importants.

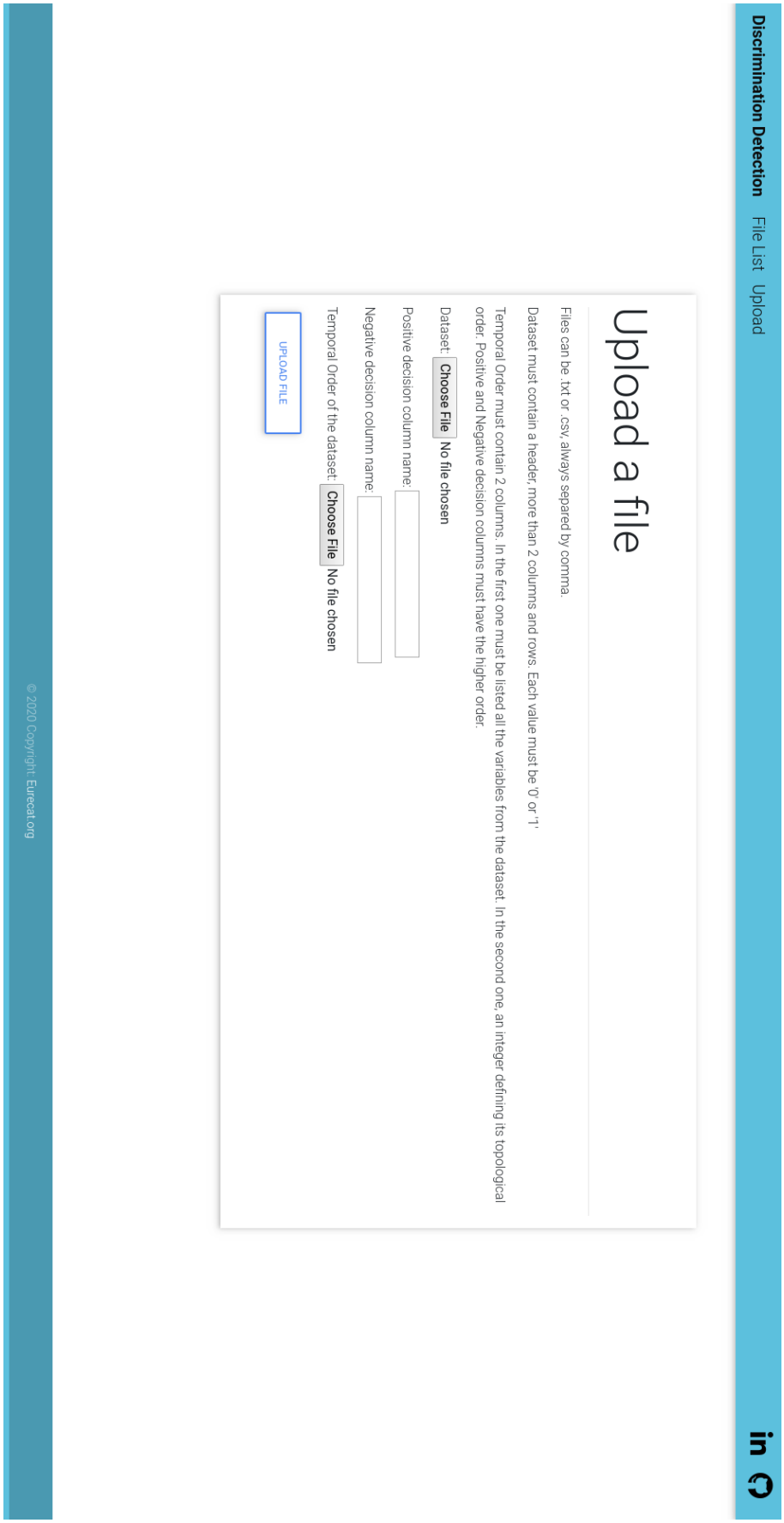


Figura 23: Càrrega d'un fitxer a la web

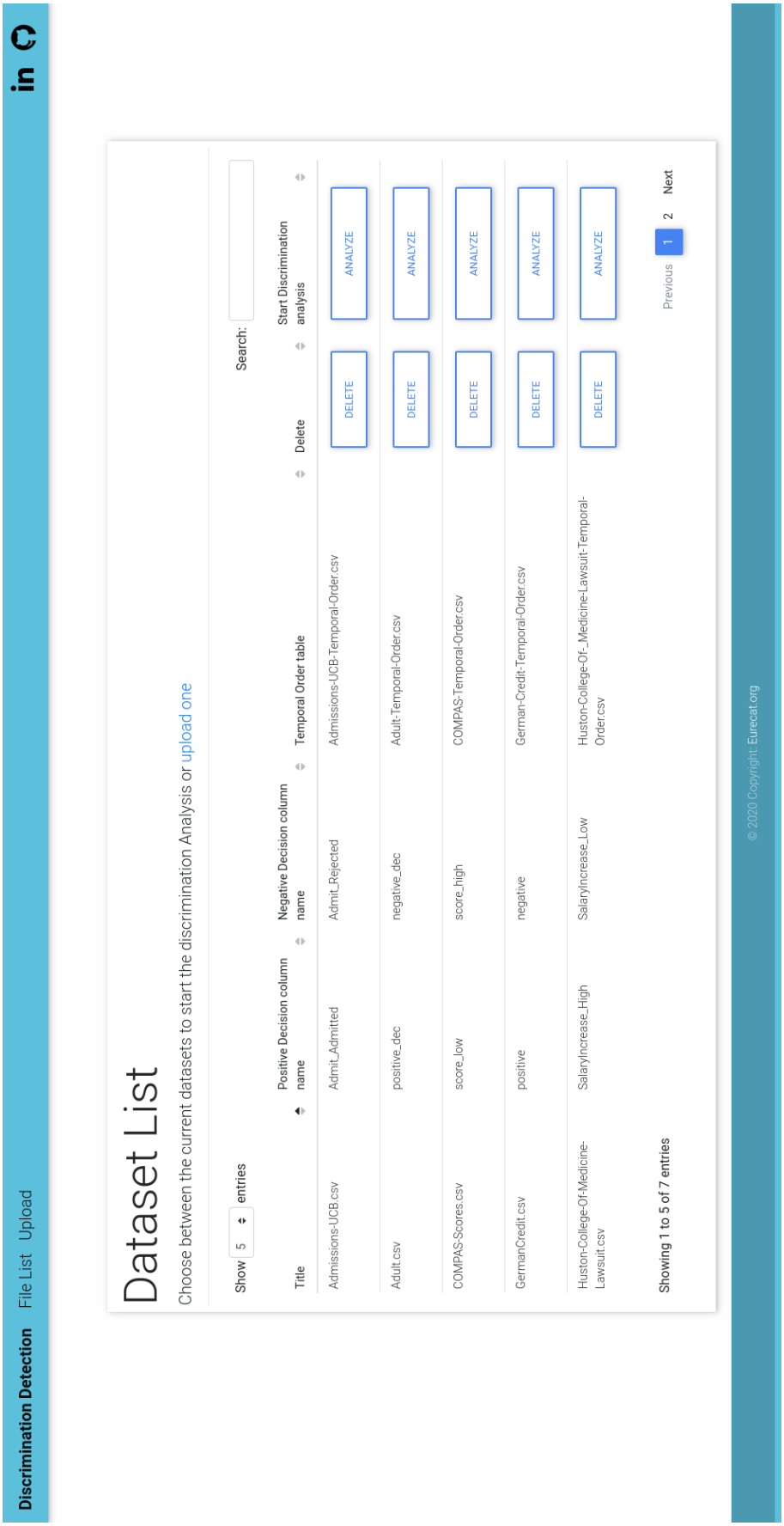
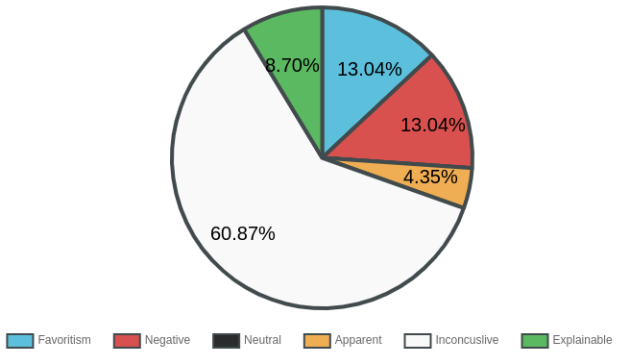


Figura 24: Llistat de fitxers

Discrimination Detection File List Upload

Discrimination Classification of

Huston-College-Of-Medicine-Lawsuit.csv



Show 10 entries Search:

Name	Positive Score	Avg. Positive Steps	Negative Score	Avg. Negative Steps	Intermediate Node	Inconclusive Score	Veredict
Research_emphasis	0.000	0.0	1.000	1.00	None	0.000	Negative Discrimination
Board_certified_no	0.000	0.0	0.796	1.00	None	0.204	Negative Discrimination
PublicationRate_High	0.000	0.0	0.725	1.43	None	0.275	Negative Discrimination
PublicationRate_Very_High	0.000	0.0	0.561	2.00	Research_emphasis: 75.58% of times	0.439	Apparent Negative Discrimination
PublicationRate_Low	0.836	2.0	0.000	0.00	Dept_Surgery: 62.20% of times	0.164	Explainable/Conditional favoritism because of Dept_Surgery
PublicationRate_Medium	0.697	2.0	0.000	0.00	Dept_Medicine: 63.41% of times	0.303	Explainable/Conditional favoritism because of Dept_Medicine
years_since_MD_Low	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs
years_since_MD_Medium	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs
years_since_MD_High	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs
years_since_MD_Very_High	0.000	0.0	0.000	0.00	-	1.000	Too many inconclusive outputs

Figura 25: Primers resultats del anàlisi

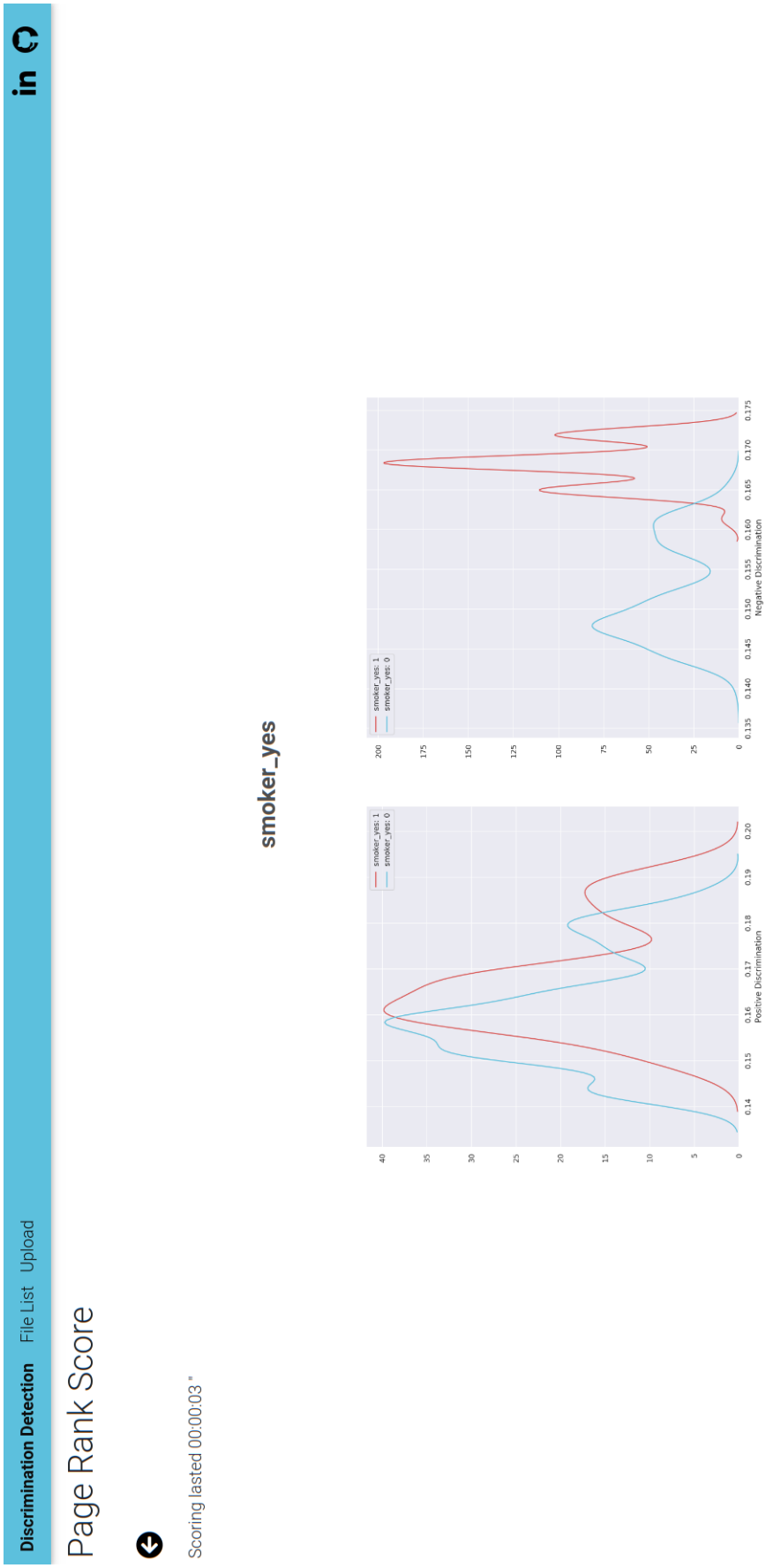


Figura 26: Plana dels resultats de Page Rank