Ekaia, ale berezia 2025, 185-201
https://doi.org/10.1387/ekaia.26327

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Bootstrap ez-parametrikoa algoritmoen errendimenduaren konparaketarako estatistika bayestarraren alternatiba gisa

(Non-parametric bootstrap for algorithm performance comparison as an alternative to Bayesian statistics)

Izei Múgica, Usue Mori, Borja Calvo*

Intelligent Systems Group, Informatika Fakultatea (UPV/EHU)

LABURPENA: Adimen artifiziala eremu ia gehienetan txertatzen ari den garai honetan, ezinbestekoa da erabiltzen ditugun algoritmoen errendimenduen konparaketak egitea. Sarritan, algoritmo baten bitartez lortutako emaitzak erabilitako datuen izaeraren eta algoritmoen estokastizitatearen arabera aldakorrak dira, eta horrek konparaketa zailtzen du. Testuinguru horretan, beraz, algoritmoen arteko konparaketa egitean, lortutako ziurgabetasunaren kuantifikazioa alderdi garrantzitsua da, baina gehienetan ez zaio nahikoa garrantzi ematen. Izan ere, ziurgabetasunaren kuantifikaziorako urteetan zehar erabili diren estatistika frekuentistako hipotesi-probek gabezia nabariak dituzte arlo honetan. Alternatiba gisa, estatistika bayestarra da soluziorik onartuena gaur egun, baina badira beste metodo batzuk oraindik esploratu ez direnak. Artikulu honetan, ziurgabetasuna azaltzeko, bootstrap ez-parametrikoaren erabilera aztertuko dugu, eta metodo horren portaera estatistika bayestarrarekin alderatu eta antzekotasunak eta desberdintasunak agerian jarriko ditugu.

HITZ GAKOAK: Ziurgabetasuna, estatistika bayestarra, birlaginketa metodoak, algoritmoen konparaketa.

ABSTRACT: *In times where artificial intelligence is being incorporated into almost all application domains, it is essential to make comparisons between the results of the algorithms we use. It is often the case that the results obtained from an algorithm are dependent on their stochasticity and the nature of the data being used. This in turn can result in having greater difficulty in comparing algorithms with each other. Therefore, in such scenario, when comparing algorithms with each other, the quantification of the uncertainty is an important aspect, but it is generally not given sufficient importance. In fact, in order to quantify such uncertainty, the frequentist statistical tests that have been used over the years have obvious shortcomings in this regard. As an alternative, Bayesian statistics is the most widely accepted solution today, but there are other methods that have not yet been explored in depth. In this article we will examine the use of non-parametric bootstrap to illustrate this uncertainty, and we will compare the behavior of this method with that of Bayesian statistics, underlining the similarities and differences.*

KEYWORDS: Uncertainty, Bayesian statistics, resampling methods, algorithm comparison.

*** *Harremanetan jartzeko* / *Corresponding author:*** Usue Mori. Intelligent Systems Group, Informatika Fakultatea (UPV/EHU). Manuel de Lardizabal, 1 (20018 Donostia-Espainia). – usue.mori@ehu.eus – https://orcid.org/

Nola aipatu* / *How to cite: Múgica, Izei; Mori, Usue; Calvo, Borja (2025). «Bootstrap ez-parametrikoa algoritmoen errendimenduaren konparaketarako estatistika bayestarraren alternatiba gisa». *Ekaia*, ale berezia 2025, 185-201. (https://doi.org/10.1387/ekaia.26327).

Jasotze-data: 2024, maiatzak 20; Onartze-data: 2024, uztailak 24.

ISSN 0214-9001 - eISSN 2444-3255 / © UPV/EHU Press 2025

Lan hau *Creative Commons Aitortu-EzKomertziala-PartekatuBerdin 4.0 Nazioartekoa* lizentzia baten mende dago

1. **SARRERA**

Adimen artifiziala gure eguneroko bizitzako arlo eta ingurune guztietara hedatu den garai honetan, argi dago sistema horien erabilerak zenbait erronka eta arazo sortzen dituela. Haien artean, azken urteetan asko hitz egin da ereduen «kutxa beltza» izaerari buruz, eta baita adimen artifizialak sortzen dituen alborapenei buruz ere. Kontu horiei erantzunez, komunitate zientifikoa buru-belarri jarri da lanean eredu gardenagoak eta justuagoak eraikitzeko asmoarekin [1, 2].

Hori hala izanik ere, badago hainbesteko garrantzia eman ez zaion arlo bat, baina izugarrizko eragina duena ereduen erabileran eta haiengan jartzen dugun konfiantzan. Arlo hori algoritmo eta ereduen ebaluazioa da, eta zehatzago esanda, lortutako emaitzen ziurgabetasunaren azterketa.

Kontu horien inguruan hausnarketa egiteko adibiderik sinpleenean oinarrituko gara. Demagun bi algoritmo ditugula eta populazio baten gainean duten jarduna konparatu nahi dugula (adibide gisa, demagun n tamainako TSP problema ebazteko bi algoritmo estokastiko ditugula eta bietatik onena zein den jakin nahi dugula). Onartuko dugu badugula eraren bat algoritmoen exekuziotik lortutako emaitzen kalitatea neurtzeko. Kontuan izan dezagun aztertu nahi dugun fenomeno honetan hainbat iturritatik etor daitekeela ziurgabetasuna; besteak beste, algoritmoen ausazko izaeratik edo populaziotik egindako lagin-hautaketatik. Beraz, bi algoritmok ematen dituzten emaitzen kalitateen arteko diferentzia X zorizko aldagaiaren bitartez adieraziko dugu, zeinak probabilitate-banaketa ezezaguna baitu. Helburua zorizko aldagai hori aztertzea izaten da, edo, zehazkiago, aldagai horren batezbestekoa aztertzea: μ.

Kontuan izan, populazio osoa elementu infinituz osatuta egonda, ezinezkoa dela denetan algoritmoa probatzea. Hartara, ohiko prozedura esperimentazio bat egitea da, non zenbait datu multzotan edo problema-instantziatan (sarritan, optimizazioaren alorrean) bi algoritmoak aplikatzen diren eta lortutako emaitzen diferentziak jasotzen diren: $x_1, x_2, ..., x_n$. Beste era batean esanda, X aldagaiaren laginketa bat dugu. Ohartu gaitezen esperimentazio honetatik lortutako emaitzak ez direla deterministak eta ziurgabetasuna hainbat iturritatik datorrela. Hasteko, hautatutako lagineko ale bakoitzean emaitza ezberdinak lortuko dira. Bestalde, hautatutako laginak berak ere badu eragina. Azkenik, algoritmoak berak estokastikoak izanik, emaitza ezberdinak eman ditzakete exekuzio bakoitzean.

$x_1, x_2, ..., x_n$ diferentzia balioen batezbestekoa egiten badugu, μ diferentzien batezbestekoaren estimazio puntual bakar bat dugu, zenbaki bat. Baina zein da estimazio horren ziurgabetasuna? Zenbateraino fidatu gaitezke emaitza horretaz? Ba al dago ezberdintasun esanguratsurik algoritmoen emaitzen artean?

Galdera horri erantzuna emateko, emaitzen azterketa estatistikoa egitea da ohiko prozedura. Hasiera batean, eta urte askoan zehar, komunitateak aho batez onartzen zituen estatistika frekuentistak proposatutako hipotesi-probak eta konfiantza-tarteak [3]. Azkenaldian, ordea, ezagun egin da prozedura horiek gabezia handiak dituztela [4, 5, 6], eta, ondorioz, estatistika bayestarraren erabilera hobetsi da [7]. Beste abantaila batzuen artean, estatistika bayestarrarekin lortutako emaitzak interpretagarriagoak dira, eta emaitzen ziurgabetasuna era naturalean neurtzea ahalbidetzen dute.

Alabaina, estatistika bayestarrak baditu bere gabeziak eta zailtasunak. Hasteko, matematikoki konplexua da eta urteetan zehar erabili dugun estatistika frekuentistatik kontzeptualki guztiz ezberdina. Bestalde, alderdi praktikoetan zentratuz, *a priori* banaketa batzuen aukeraketa ez da tribiala eta horrek eragina izan dezake emaitzetan.

Estatistika frekuentistaren barruan, badago algoritmoen konparaketa egiteko hain erabilia ez den metodo bat: bootstrap ez-parametrikoa. Hurbilketa hori oso intuitiboa eta kontzeptualki sinplea izanik ere, erabilitako estimatzaileen ziurgabetasuna kuantifikatzeko erabili daiteke. Hori horrela, artikulu honen helburua hau da: algoritmoen ebaluaziorako bootstrap ez-parametrikoaren erabilera estatistika bayestarrak eskaintzen duen alor teorikoarekin alderatzea eta uztartzea.

Horretarako, artikulua sei ataletan banatu dugu. 2. atalean, estatistika frekuentistaren eta bayestarraren sarrera bat egingo dugu, haietako bakoitzaren filosofia zein den azalduz eta ezberdintasunetan arreta ipiniz. 3. atalean, lan honetan konparatuko ditugun bi teknikak sakonago aztertuko ditugu: korrelaziodun t-test bayestarra eta bootstrap ez-parametrikoa. Ondoren, 4. atalean, egindako esperimentuen diseinua azalduko dugu, eta 5. atalean lortutako emaitzak. Azkenik, 6. atalean ondorioak laburbildu eta etorkizunerako ikerketa-ildo batzuk proposatuko ditugu.

2. ZIURGABETASUNAREKIN LAN EGITEKO BI HURBILKETA

Nahiz eta matematikan adituak ez izan, badaude gure eguneroko bizitzan etengabe erabiltzen ditugun termino matematiko batzuk. Arruntenetariko bat «probabilitate» terminoa da, edozer gauzaren inguruan zalantzaren bat edo ziurtasun falta dugula adierazteko erabiltzen duguna. Noski, egunerokoan termino horri ematen diogun erabilera eta haren inguruan egiten dugun interpretazioa ez dator beti bat matematikan esleitzen zaion interpretazioarekin. Izan ere, ez da kontzeptu tribiala, eta estatistikaren arloan badaude bi interpretazio probabilitate kontzeptuarentzat, bi pentsamendu-eskola (eta lan egiteko era) sortu dituztenak: interpretazio frekuentista eta interpretazio bayestarra.

Eskola frekuentistarentzat, probabilitateak zerbait objektiboa izan behar du, eta, izenak adierazten duen bezala, maiztasun terminoarekin lotzen da. Hau da, errepikakorra den ausazko esperimentu bat baldin badaukagu, gertaera jakin baten probabilitatea esperimentuaren hainbat errepikapenetan gertaera horren maiztasun erlatiboa izango da.

Pentsa dezakegu interpretazio frekuentista dela naturalena edo intuitiboena, gehienbat lehen aipatutako adibideak erabiltzen direlako matematikoki ilustratzeko. Alabaina, interpretazio hori ez dator beti bat intuizioarekin, batik bat «esperimentua» errepikakorra ez bada. Esate baterako, nola interpreta dezakegu egun zehatz batean euria egiteko probabilitatea? Lehen aipatu bezala, gure eguneroko bizitzan, gure usteak adierazteko erabiltzen dugu askotan probabilitate terminoa, eta, hortaz, argudia daiteke probabilitatearen interpretazioa hain justu hori izan behar dela. Hori da eskola bayestarraren funtsa, non probabilitateak «usteak» adierazteko erabiltzen diren edo, beste era batean esanda, fenomeno baten inguruan daukagun ziurgabetasuna.

Ezberdintasuna hobeto ulertzeko, adibide sinple bat jarriko dugu. Demagun txanpon bat dugula. Hura jaurtitzean, bi emaitza posible daude: aurpegia edo gurutzea. Jaurtiketa horren emaitza modelatzeko, Bernoulli banaketa bat erabil dezakegu.[1] Banaketa horrek parametro bakarra du, θ, bi balio posibleren artean baten probabilitatea adierazten duena (aurpegiarena, adibidez; bestearen probabilitatea, gurutzearena, $1 - \theta$ izango da). Helburua θ parametroaren estimazio bat egitea bada, txanpona behin eta berriro jaurti eta laginketa bat lortuko dugu.

Bai estatistika frekuentistan, bai bayestarrean, banaketa bera erabiltzen da esperimentua modelatzeko, baina parametroaren tratamendua ezberdina da.

Estatistika frekuentistan, onartzen da modelatu nahi dugun esperimentu horretan benetako balio bat (bakarra) dagoela parametroarentzat, nahiz eta balio hori ezezaguna izan. Estatistika-eskola horretan, aurpegia ateratzeko probabilitatea (θ) gertaeraren maiztasun erlatiboa erabiliz estimatzen da, zenbaki baten bidez, kontuan izanda jaurtiketa kopurua infiniturantz doanean estimazio hori benetako probabilitatera hurbilduko dela. Kontuan izan estimazio hori egiteko, laginketatik lortutako datuetan soilik oinarritzen garela.

Estatistika bayestarrean, ordea, probabilitateak gure ziurgabetasuna modelatzeko erabiltzen direnez, parametroa bera zorizko aldagai moduan modelatzen dugu. Hau da, θ parametroak edozein balio har dezake, balio

[1] Hainbat jaurtiketetan lortzen den emaitza modelatu nahi badugu binomial bat erabiliko genuke, baina egoera sinplifikatzeko jaurtiketa bakarra kontsideratuko dugu.

bakoitzak bere probabilitatea izanik (betiere bere domeinuan, hau da, 0 eta 1 artean), eta haren dentsitate-funtzioak adieraziko du nola banatzen den probabilitate-masa domeinu horretan. Estimazioaren helburua parametro horren *a posteriori* banaketa lortzea izango da, eta, horretarako, Bayesen teorema erabiltzen da. Txanponaren adibidean, θ parametroaren banaketa honela lor dezakegu:

$$p\left(\phi|D\right) = \frac{p(\phi)p\left(D|\phi\right)}{p(D)}$$

non p funtzioak kasu jarraituan dentsitate-funtzioa adierazten duen eta kasu diskretuan, aldiz, probabilitate-masa funtzioa. Gainera, $p(\theta)\theta$ parametroaren *a priori* banaketa da, hau da, banaketaren gainean dugun hasierako ustea; $p(D|\theta)$ laginketan lortutako datuen egiantza da (hau da, eskuragarri dugun lagina behatzeko probabilitatea θ balioa emanda); eta $p(D)$ konstante normalizatzailea. Kontuan izan estimazio mota horrek datuetatik ateratako informazio objektiboa gure usteekin nahasten duela. Hortik dator estatistika bayestarraren izaera «subjektiboa», urteetan zehar gabezia moduan ikusi dena baina abantaila asko eskaintzen dituena.

Parametroak eta lortutako estimazioak interpretatzeko moduan ezberdintasun horrek inplikazio sakonak ditu. Esate baterako, θ parametroaren interpretazio frekuentista egiten badugu, hau da, zenbaki bat dela onartzen badugu, $p(\theta > 0{,}5)$ kalkulatzeak ez dauka zentzurik. Aldiz, θ ikuspuntu bayestarretik ulertzen badugu $p(\theta > 0{,}5)$, probabilitatea kalkulatzeak zentzu handiagoa dauka. Hurrengo ataletan ikusiko dugun bezala, bi hurbilpenak erabil daitezke algoritmoen edota ereduen errendimenduak konparatzean ziurgabetasuna aztertzeko, baina bi hurbilpenetan lortutako emaitzak interpretatzean oso ondo ulertu behar da bakoitzaren atzean dagoen «filosofia».

3. ALGORITMOEN EMAITZEN KONPARAKETARAKO METODOAK

Honelako egoera batean, algoritmoen emaitzak konparatzeko eta lortutako emaitzen analisi estatistikoa egiteko modurik ohikoena estatistika frekuentistan definitutako hipotesi-proba bat egitea da [3], adibiderik oinarrizkoena t-test delakoa izanik. Hipotesi-probetan, bi hipotesi definitzen dira, eta, gure problema partikularrean (algoritmoen exekuzio-emaitzen diferentzien batezbestekoaren balioa aztertzea) honela definituko dira: $H_0 : \mu = 0$ (hipotesi nulua), $H_1 : \mu \neq 0$ (hipotesi alternatiboa). Hau da, diferentzien batezbestekoa 0 da, eta beraz, bi algoritmoen ontasun-maila berbera da, edo batezbestekoa 0 bestelakoa da, eta, beraz, portaera desberdina dute. Ja-

rraian, metodo horiek, hipotesi nulua egia balitz, esperimentuan lortutako $\bar{x}$ (algoritmoen exekuzio-emaitzen diferentzien batezbestekoa) estimazioa edo hori baino balio arraroagoa lortzeko probabilitatea kalkulatzen dute. Balio hori da, hain zuzen, denok ezagutzen dugun *p*-balio famatua. Ondoren, *p*-balioa txikia bada (normalean, 0,05 baino txikiagoa), ulertzen da lortutako balioa oso inprobablea dela hipotesi nulupean eta, beraz, hipotesi nulua errefusatzen da. Kontrako kasuan, aldiz, *p*-balioa handia denean, hipotesi nulua ezin dugu errefusatu, eta bi algoritmoen arteko diferentziak esanguratsuak direla esateko nahikoa ebidentziarik ez dugula ondoriozta dezakegu soilik.

Horrelako hurbilketek hainbat arazo dituzte [4, 6, 5], baina aipagarrienetako bat interpretazioa da. Testaren emaitzak zuzenean esaten digu, ezarritako konfiantzarekin, ea onar dezakegun batez-bestekoen artean diferentziarik dagoen ala ez, baina erantzun hori, Benavoliren hitzetan, erantzun «zuri-beltza» da [7]. Ziurgabetasuna testaren barne-kalkuluetan kontuan hartzen bada ere, testaren emaitzak ez digu informaziorik ematen horren inguruan.

Gabezia horiei aurre egiteko asmoz, artikulu honetan algoritmoen konparaketa egiteko bi metodo alternatibotara joko dugu. Hasteko, alternatiba nagusitzat hartzen den estatistika bayestarreko testak erabiliko ditugu [8]. Bestalde, soluzio berritzaile gisa, estatistika frekuentistan oinarria duen bootstrap ez-parametrikoa [9] erabiliko dugu. Bi metodoen arteko konparaketa egingo dugu batik bat ziurgabetasunaren kuantifikazioari begira bakoitzaren ahalmenak eta ahuldadeak azpi marratzeko asmoz.

3.1. **Estatistika bayestarreko korrelaziodun t-testa**

Aurreko ataletan aipatu dugun moduan, estatistika bayestarrean hurbilketa (eta interpretazioa) oso ezberdina da. Adibide sinple batekin hasteko, demagun bi algoritmoren emaitzen arteko diferentziak μ eta σ parametro ezezaguneko banaketa gausstarrari jarraitzen diola. Estatistika bayestarrean bi parametro horiek ere zorizko aldagaiak direla onartzen da. Gure kasuan μ parametroa interesatzen zaigu, hain justu diferentzien batezbestekoa adierazten duelako (hasieran ipinitako formulekin bat egiteko, pentsa μ parametroa θ parametroa dela eta haren inguruko analisia egin nahi dugula). Hortaz, *a priori* banaketa batetik abiatuz, hurbilketa bayestarrean esperimentazioan lortutako emaitzak erabiliko ditugu μ parametroaren *a posteriori* banaketa lortzeko, banaketa horrek algoritmoen diferentziaren batezbestekoari buruz espero dezakeguna adieraziko duelakoan. Kontuan izan probabilitate-banaketa bat emateak estimazio puntual bat lortzeko aukera ematen duela —itxaropena erabiliz, adibidez—, baina, gainera, ziurgabetasunari buruzko informazioa ere ematen digula —adibidez *a posteriori* banaketaren bariantzaren bitartez—.

Algoritmoen konparaketarako estatistika bayestarra erabiltzeko, zenbait erabaki hartu behar ditugu; besteak beste, estimatu nahi dugun parametroaren *a priori* banaketa, eta datuen egiantz-funtzioaren forma. Hainbat aukera egonik ere, gure testuinguruan korrelaziodun t-test bayestarra [10] erabiltzea izaten da ohikoena. Metodo horretan, honako erabaki hauek hartu ohi dira:

1. **Egiantz-funtzioa.** Metodoan datuak (hau da, esperimentuan lortutako emaitzen arteko diferentziak) honako banaketa honetatik sortu direla onartzen da:

$$\boldsymbol{x} = \mathbf{1}\mu + \boldsymbol{\upsilon},$$

non $\boldsymbol{x} = (x_1, x_2, ..., x)$ esperimentuan lortutako diferentzien bektorea den, $\mathbf{1}$ n tamainako batekoen bektorea den eta $\boldsymbol{\upsilon}$ n dimentsioko banaketa gausstar batetik ateratako zarata den, hots, $\boldsymbol{\upsilon} \sim MVN(\mathbf{0}, \Sigma_{n\times n})$. Azkenik, $\Sigma_{n\times n}$ kobariantza matrizearen balioak honela definitzen dira: $\Sigma_{ii} = \sigma^2$ eta $\Sigma_{ij} = \rho\sigma^2$, $i \neq j$ denean, non σ^2 bariantza eta ρ korrelazio- koefizientea - zeina [–1,1] tartean baitago (neurketak askeak ez direnean, haien arteko korrelazioa aintzat hartzen da; adibidez, balidazio gurutzatua erabili bada eta neurketak datu-base beraren gainean eginda badaude, baina, bestela, gure kasuan bezala, $\rho = 0$ hartzen da)- diren. Datuen gaineko eredu hori onartuz gero, egiantz-funtzioak honako itxura hau hartzen du:

$$p\left(\boldsymbol{x}\middle|\mu,\sum\right) = \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\mathbf{1}\mu\right)^T \sum^{-1}\left(\boldsymbol{x}-\mathbf{1}\mu\right)\right)}{(2\pi)^{n/2}\sqrt{\left|\sum\right|}}$$

2. ***A priori* banaketa.** Normal-Gamma banaketa bat aukeratzen da, $\{\mu_0 = 0, k_0 \to \infty, a = -1/2, b = 0\}$ parametroduna.

Aukeraketa horiek ez dira ausazkoak: hautatutako *a priori* banaketa definitu dugun egiantz-funtzioaren konjugatua da, eta ondorioz, *a posteriori* banaketaren kalkulua analitikoki egin daiteke (horrela ez balitz, zenbakizko metodoak, laginketa-metodoak eta abar erabili beharko genituzke *a posteriori* banaketa tratatzeko), emaitza $n - 1$ askatasun graduko Student-en t banaketa bat izanik:

$$p\left(\mu\middle|\,\boldsymbol{x},\mu_0,k_0,a,b\right) \equiv St\left(\mu;n-1,\bar{x},\left(\frac{1}{n}+\frac{\rho}{1-\rho}\right)\hat{\sigma}^2\right)$$

A priori banaketarentzat aukeratu ditugun parametroen balioekin, lortzen dugun *a posteriori* banaketa estatistika frekuentistako t-test korrelatuan *p*-balioa kalkulatzeko erabiltzen den banaketa bera da. Baina gogoan izan, alde batetik, estatistika frekuentistan banaketa bera ez dela itzultzen, hau da, *p*-balioa kalkulatzeko erabiltzen dela bakarrik, eta bestetik, bi kasuetan interpretazioa ezberdina dela. Alde horretatik, argi dago estatistika bayestarrak informazio gehiago eta interpretatzeko errazagoa dena ematen digula; interesatzen zaigun parametroaren (μ-ren) probabilitate-banaketa bat, alegia.

Alabaina, galdetu genezake ea zer punturaino adierazten duen ondo banaketa horrek zer ziurgabetasun dagoen (edo izan beharko genukeen) benetan' testuinguru zehatz batean. Ez hori bakarrik; izan ere, ba al dago eraren bat ikuspuntu frekuentistatik antzeko zerbait lortzeko? Bestalde, test bayestar horren izaera parametrikoak (bai *a priori* banaketari, bai egiantz-funtzioari dagokienez) ere baditu bere mugak; besteak beste, zer gertatzen da gure datuek ez badiote onartutako ereduari jarraitzen? Alternatiba ez-parametriko bayestarrak ere badiren arren, ez daude hain hedatuta. Hori guztia kontuan izanda, hurrengo ataletan alternatiba bat proposatuko dugu, bootstrap ez-parametrikoa tresna gisa erabiliz.

3.2. **Bootstrap ez-parametrikoa algoritmoen konparaketarako**

Behin baino gehiagotan aipatu dugun moduan, algoritmoak konparatzean funtsezkoa da egindako esperimentaziotik atera ditugun ondorioen ziurgabetasuna, nolabait, neurtzea. Egin dezakegun analisi estatistikorik sinpleena esperimentazioaren emaitzei (algoritmo horien «erabileraren» lagin bati) estimatzaileak aplikatzea da. Adibidez, artikulu honetan erabiltzen ari garen kasuan, algoritmoen arteko diferentziaren itxaropenaren estimatzailea, $\hat{\mu}$, laginaren batezbestekoa izango da estatistika frekuentistan: $\bar{X}$. Estimatzaile hori daukagun laginari aplikatuz estimazio puntual bakar bat lortuko dugu: $\bar{\boldsymbol{x}}$.

Adibide horretan argi ikusten da estimatzaileak ere zorizko aldagaiak direla; izan ere, beste lagin ezberdin batzuei aplikatuz gero (beste esperimentazio batzuei), beste estimazio batzuk lortuko ditugu. Kasu sinple horretan, nola jakin dezakegu zein den egin dugun estimazioaren ziurgabetasuna? Esperimentazioa hainbat aldiz errepikatuz gero, zer punturaino izango dira ezberdinak lortuko ditugun estimazioak?

Testuinguru kontrolatuetan, galdera horri teorikoki erantzun diezaiokegu, estimatzailearen probabilitate-banaketa analitikoki kalkula daitekeelako. Zoritxarrez, kasu orokorrean hori ez da posible. Estimatzailearen banaketaren hurbilpen bat lortzeko existitzen diren tekniken artean bat landuko dugu lan honetan: bootstrap ez-parametrikoa.

Bootstrap metodoa birlaginketa-teknika bat da [9] eta *plug-in* printzipioa dago haren oinarrian. Printzipio horren ideia sinplea da: zerbait ezagutzen ez baduzu, estimazio batekin hurbildu. Gure helburua estimatzailearen banaketa probabilistikoa ezagutzea da, baina analitikoki lortu ezin dugunez, haren espresioaren hurbilketa bat egingo dugu hura laginduz. Horretarako bi pauso eman beharko genituzke: i) jatorrizko populazioa lagindu, eta ii) lagin horri estimatzailea aplikatu. Alabaina, kasu honetan, ez dugu ezagutzen jatorrizko populazioa (algoritmoak aplikatzean lor daitezkeen emaitza posible guztiak), baina badugu haren hurbilketa bat: gure hasierako lagina. Beraz, prozesuaren lehenengo pausoan, populazioa lagindu beharrean, hasierako lagina (esperimentazioaren emaitzak) laginduko dugu, eta, ondoren, aztertu nahi dugun estimatzailea aplikatuko diogu birlaginketa horietako bakoitzari. Ikusten dugun moduan, etengabe *plug-in* printzipioa aplikatzen dugu, ezezaguna dena estimazio bidez hurbilduz.

Zehatzago esanda, hasierako lagina, $(x_1, ..., x_n)$, gure populazioa dela onartzen badugu eta x_i guztiak ekiprobableak izanik, bakoitzaren probabilitatea $1/n$ izango da. Beraz, lehenengo pausoan «populazio» hori laginduko dugu ausazko laginketa itzuleraduna erabiliz, n elementuko bootstrap lagin bat eraiki arte: $\boldsymbol{x}_1^*$. Lagin horri gure estimatzailea aplikatu, eta estimazio bat lortuko dugu: gure kasuan $\bar{\boldsymbol{x}}_1^*$. Prozedura hori behin eta berriro errepikatuko dugu, B bootstrap lagina, $\{\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, ..., \boldsymbol{x}_n^*\}$ eta dagozkien B estimazioak lortu arte. Prozedura horri bootstrap ez-parametrikoa deritzo eta lortutako estimatzaile multzoak B tamainako estimatzailearen bootstrap laginketa bat osatzen du, estimatzailearen banaketa hurbiltzeko erabiliko duguna.

4. **ESPERIMENTUEN DISEINUA**

Atal honetan, algoritmoen konparaketarako erabiltzen diren hurbilpenak gure proposamenarekin (hurbilpen frekuentistaren barruan kokatua) alderatuko ditugu, zenbait esperimenturen bitartez. Hurbilpen frekuentista eta bayestarra izaeraz desberdinak diren arren, bakoitzak bere indargune eta ahuleziak dituelarik, aztertuko dugu biak zer punturaino izan daitezkeen bateragarriak jarraian jorratuko ditugun esperimentuekin.

Hurbilpenak konparatu ahal izateko, interesgarria izan daiteke algoritmo-pare jakin baten exekuzioaren portaera era artifizialean simulatzea; izan ere, simulazio horren gainean hurbilpen frekuentista eta bayestarra aplikatuta, bien arteko desberdintasunak azaleratuko dira (desberdintasunik egonez gero, jakina). Zehatzago esanda, algoritmo-pare baten exekuzioen emaitza banaketa- pare jakin batzuekin adieraziz gero, banaketa horiek lagindu genitzake, diferentziak kalkulatu, birlaginketa-metodoa eta korrelaziodun t-test bayestarra aplikatu, eta abar. Gainera, jatorrizko banaketak

ezagutzen ditugunez, X diferentzia-aldagaiaren benetako banaketa ere kalkula dezakegu, bertatik itxaropena, desbideratze estandarra eta bestelakoak ateratzeko eta informazio hori bi hurbilpenek eskaintzen dituzten emaitzekin uztartzeko, 1. irudian ageri den moduan. Jakina, egoera erreal batean ez dugunez X aldagaiaren banaketaren inguruko informazio gehigarri hori izango, benetako egoera bat ere tratatuko dugu artikulu honetan, zeinaren inguruko xehetasunak 4.1. atalean deskribatuko diren.

Testuinguru horretan, bost esperimentu egin ditugu, guztiak oinarri beraren gainean. Hasteko, simulazio kasuetarako, algoritmoek lortutako emaitzak banaketa banarekin modelizatuko ditugu. Adibidez, onar genezake, A algoritmoak problema jakin bat ebaztean lortzen dituen emaitzek banaketa gausstar bati jarraitzen diotela. Jarraian, 50 tamainako lagin bana aterako dugu banaketa horietatik. Azken esperimenturako —kasu errealerako hain zuzen— jatorrizko banaketak ezezagunak direnez, 50 tamaina finkoko lagin batetik abiatuko gara zuzenean (berehala tratatuko dugu datu horien izaera). Hemendik aurrera, esperimentu guztietan prozedura berbera aplikatuko dugu, sekuentzialki gauzatu beharreko lau urratsez osatua:

1. **Laginen arteko diferentzia kalkulatu.** Demagun bi algoritmoren exekuzioen portaerak konparatzeko $\boldsymbol{x}^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}\right)$ eta $\boldsymbol{x}^{(2)} = \left(x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)}\right)$ laginak lortu ditugula, hurrenez hurren. Orduan, honako kalkulu hau egingo dugu:

$$\boldsymbol{x} = \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} = \left(x_1^{(1)} - x_1^{(2)}, x_2^{(1)} - x_2^{(2)}, ..., x_n^{(1)} - x_n^{(2)}\right)$$

2. **Bootstrap bidezko birlaginketa eta dentsitatearen estimazioa.** Bootstrap teknikaren bidez aurreko $\boldsymbol{x}$ birlaginduko dugu, $n = 50$ tamainako $B = 50$ lagin sortuko. Lagin bakoitzeko haren batezbestekoa kalkulatuko dugu, diferentzien banaketaren itxaropenaren estimazio- puntual gisa. Hortaz, $B = 50$ estimazio-puntualez osaturiko $\hat{\boldsymbol{\mu}}$ lagin bat lortuko dugu:

$$\hat{\mu}^* = \left(\overline{x}_1^*, \overline{x}_2^*, ..., \overline{x}_{50}^*\right)$$

   Ohartu lagin hori μ-ren estimatzailearen banaketaren bootstrap hurbilketa bat dela. Dena den, aurrerago azalduko dugun moduan, hurbilpen bayestarrarekin konparaketak egiteko, interesgarria izango da μ-ren estimatzailearen banaketa jarraitu bat lortzea, bi banaketak diskretuak edo jarraituak izatea eskatzen baitu erabiliko dugun konparaketa-neurriak. Hori dela eta, $\hat{\boldsymbol{\mu}}$ laginetik abiatuta, banaketa hori hurbilduko dugu, kernel dentsitatearen estimazio (KDE) teknika erabiliz [11]. Teknika hori, hitz gutxitan, banaketa enpirikoen ideia-

tik abiatzen da, *f* dentsitate-funtzio bat hari dagokion lagin baten bidez estimatuz, honako formula honen arabera:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

non *n* lagin-tamaina den, *K* estimaturiko dentsitate-funtzioa leuntzen duen kernel deritzon funtzio bat den eta *h* leuntasun-maila kontrolatzen duen banda-zabalera den. Gure kasuan, ohartu behar dugu halaber $x_i = \bar{\boldsymbol{x}}_i^*$ dela.

Argi dago erabiltzen diren kernel funtzioek eta banda-zabalerek eragina dutela estimazioan, eta ikuspegi horri jarraikiz, interesgarria litzateke parametro horien beste konfigurazio batzuetan emaitzak nola aldatzen diren ikustea. Edonola ere, limitearen teorema zentrala kontuan hartuz, μ-ren estimatzailearentzat banaketa gausstar bat espero daitekeenez, eta bestalde, *h*-ren doikuntzarako literaturan emaitza onak ematen dituen heuristiko bat erabili ohi denez [12, 13], $\hat{f}_h(x)$ kalkulatzeko, kernel gausstarra eta heuristiko hori erabiliko ditugu kasu guztietan. Hau da:

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2},$$

$$h = 0.9 \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) n^{-\frac{1}{5}},$$

non $\hat{\sigma}^2$ lagin-kuasibariantza den, IQR kuartil arteko heina eta *n* intereseko banaketa hurbiltzeko erabiliko dugun laginaren tamaina.

3. **Korrelaziodun t-test bayestarraren aplikazioa.** Hurbilpen bayestarra kontuan hartuz, korrelaziodun t-test bayestarra aplikatuko dugu $\boldsymbol{x}$ laginaren gainean, *a priori* banaketarentzat 3.1. atalean zehazturiko banaketa hartuz, dagozkion parametroekin eta $\rho = 0$ hartuz (suposatzen da behaketen artean korrelaziorik ez dagoela). Hartara, μ-ren *a posteriori* banaketa honako Student banaketa hau izango da:

$$St\left(\mu; n-1, \bar{\boldsymbol{x}}, \frac{\hat{\sigma}^2}{n}\right),$$

$\bar{\boldsymbol{x}}$ lagin-batezbestekoa izanik.

4. **Banaketen arteko diferentzia estatistikoaren kalkulua.** Azkenik, hurbilpen bakoitzarekin lorturiko banaketen arteko diferentzia kalkulatuko dugu. Horretarako, Jensen-Shannon dibergentzia proposatzen dugu neurri gisa, honako formula honen bidez emana datorrena:

$$D_{JS}\left(f \parallel g\right) = \frac{1}{2} D_{KL}\left(f \parallel q\right) + \frac{1}{2} D_{KL}\left(g \parallel q\right),$$

non f eta g konparatu beharreko banaketak diren, $q = \frac{1}{2}(f + g)$ den eta $D_{KL}\left(f \parallel q\right)$ eta $D_{KL}\left(g \parallel q\right)$, berriz, Kullback-Leibler dibergentziak. Azken horiei dagokionez, f eta g bi banaketa emanda, Kullback-Leibler dibergentzia formula honen bidez kalkulatzen da:

$$D_{KL}\left(f \parallel g\right) = \int_{-\infty}^{\infty} f(x)\, log\left(\frac{f(x)}{g(x)}\right) dx.$$

Banaketen arteko konparaketa egitean oso tipikoa da Kullback-Leibler dibergentzia erabiltzea, baina Jensen-Shannon dibergentzia erabiltzea proposatzen dugu, Kullback-Leibler-en zenbait ahulezia gainditzen dituelako. Hasteko, kontuan izanik edozein f eta g banaketa emanda, $D_{KL}(f \,||\, g) \geq 0$ dela, hau da, ez dagoela goitik bornatua, lortutako balioak interpretatzea ez da tribiala. Adibidez, $D_{KL}(f \,||\, g) = 100$ izateak ez du adierazten balio handia edo txikia denik. Are gehiago, Kullback-Leibler dibergentzia ez da simetrikoa, eta horrek interpretagarritasuna are gehiago zailtzen du. Jensen-Shannon dibergentziaren kasuan, berriz, alde batetik, edozein f eta g banaketa emanda, $0 \leq D_{JS}(f \,||\, g) \leq 1$ betetzen da; eta bestetik, simetrikoa da. Beraz, interpretagarriagoa da.

Arestian deskribaturiko lau urratsak eginez, hurbilpen frekuentistaren eta bayestarraren artean desberdintasunik baden aztertu ahalko dugu.

4.1. **Esperimentaziorako datuak**

Atalaren hasieran aipatu bezala, esperimentazioa aurrera eramateko, lau simulazio eta kasu erreal bat hartuko ditugu kontuan. Atal honetan, horren guztiaren inguruko xehetasunak eskainiko ditugu.

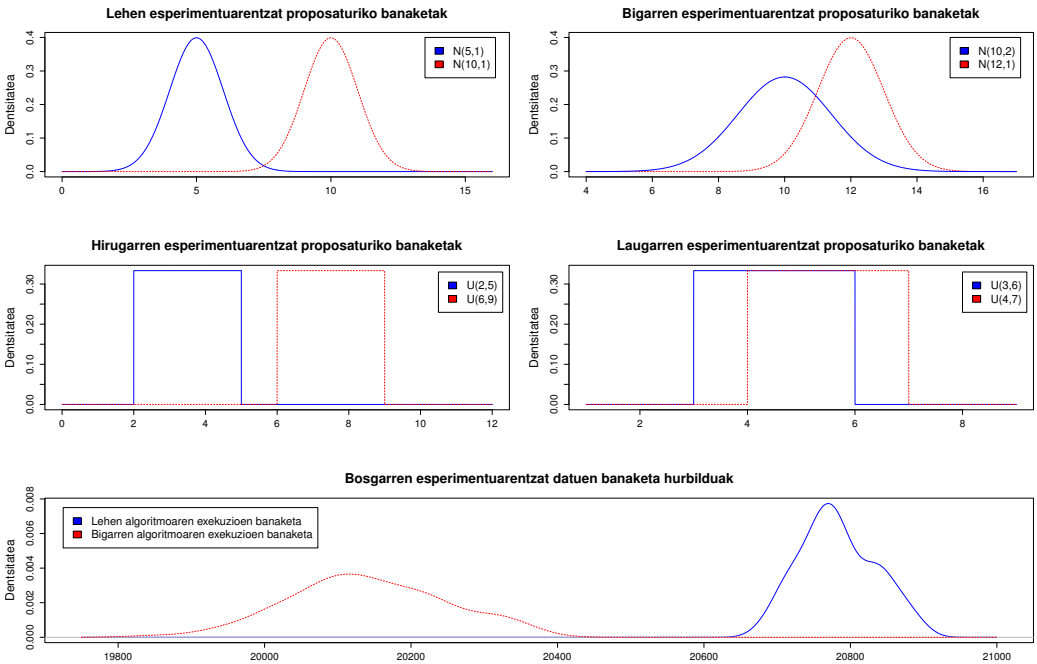
Simulazioekin hasteko, bi banaketa-familia proposatzen ditugu lehen lau esperimentuak egiteko:

— **Banaketa gausstarrak.** Bi egoera aztertuko ditugu. Lehenengoan, algoritmo baten emaitzak $\mathcal{N}(5,\ 1)$ banaketari jarraituko dio eta bes-

tearenak, berriz, $\mathcal{N}(10,\ 1)$ banaketari. Bigarrenean, algoritmoen emaitzek $\mathcal{N}(10,\ 2)$ eta $\mathcal{N}(12,\ 1)$ banaketei jarraituko diete, hurrenez hurren.

— **Banaketa uniformeak.** Kasu honetan ere bi egoera izango ditugu. Lehenengoan algoritmoen emaitzen banaketak $\mathcal{U}(2,\ 5)$ eta $\mathcal{U}(6,\ 9)$ izango dira, hurrenez hurren, eta bigarrenean, berriz, $\mathcal{U}(3,\ 6)$ eta $\mathcal{U}(4,\ 7)$, hurrenez hurren.

Esperimentuetan erabiliko ditugun datuen banaketak 1. irudian ageri dira.



1. irudia. Esperimentuetan erabilitako banaketak.

Azkenik, datu errealei dagokienez, Linear Ordering Problem (LOP) optimizazio-problemaren zenbait instantzia ebazteko aplikaturiko bi algoritmo estokastikoren emaitzak erabili ditugu. Zehatzago esanda, beste artikulu zientifiko bateko [14] esperimentaziorako sortutako 50 LOP instantzietan, bi optimizazio algoritmoren emaitzak erabili ditugu. Datu horiek sortzeko prozesua [14] artikuluan azalduta dago xeheki, eta, beraz, artikulu hau ez luzatzearren ez dugu hemen errepikatuko. Besterik gabe, komeni da aipatzea bi algoritmoen emaitzek jarraitzen dituzten banaketak ezezagunak direla.

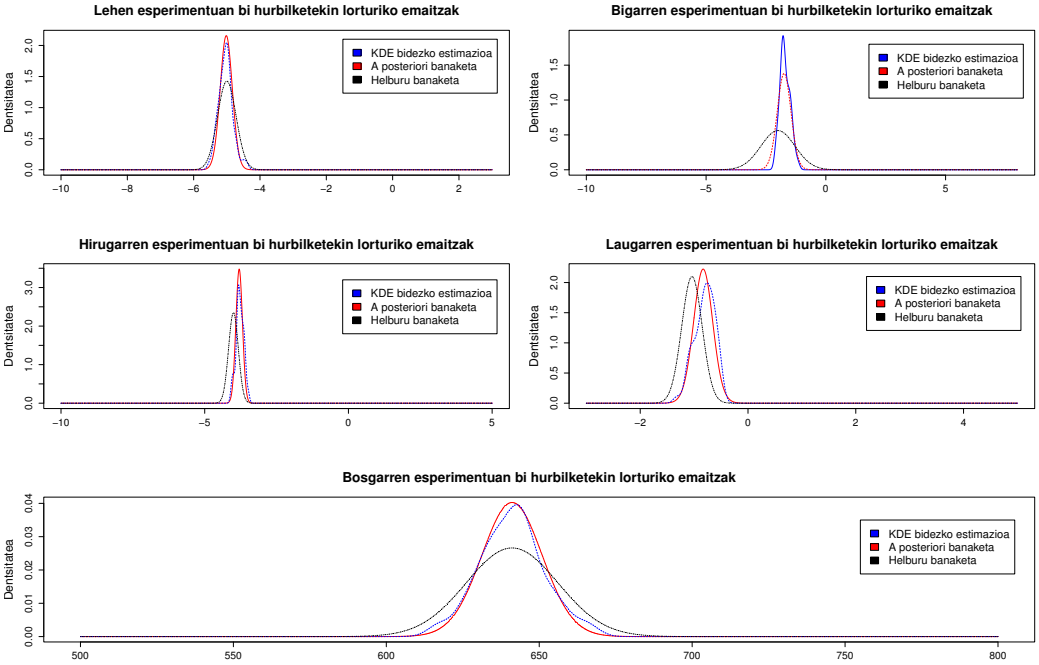
5. EMAITZAK ETA EZTABAIDA

Aurreko atalean deskribaturiko irizpideak eta metodologia kontuan hartuz, atal honetan esperimentu bakoitzean lorturiko emaitzak aurkeztuko ditugu. Zehazkiago esanda, hurbilpen bayestarraren eta frekuentistaren arteko desberdintasunak islatu ditzaketen Jensen-Shannon dibergentziak aurkeztuko ditugu esperimentu bakoitzeko. Emaitzak 1. taulan jasota daude.

1. taula. Bost esperimentuei dagozkien Jensen-Shannon dibergentziaren emaitzak.

		JS dibergentzia	
		Boot + KDE	Bayes
1. esperimentua (Gaussiarrak)	Boot + KDE	0	0.011
	Bayes	0.011	0
2. esperimentua (Gaussiarrak)	Boot + KDE	0	0.013
	Bayes	0.013	0
3. esperimentua (Uniformeak)	Boot + KDE	0	0.007
	Bayes	0.007	0
4. esperimentua (Uniformeak)	Boot + KDE	0	0.005
	Bayes	0.005	0
5. esperimentua (Errealak)	Boot + KDE	0	0.006
	Bayes	0.006	0

Taulak agerian uzten du hurbilpen frekuentistak eta bayestarrak eskaintzen dituzten emaitzak oso antzekoak direla, jatorrian paradigma desberdinak izan arren. Hori oso interesgarria da; izan ere, hein batean esan genezake bootstrap ez-parametrikoa estatistika bayestarraren alternatiba sendoa izan daitekeela, hau da, biek ondorio berera garamatzate, helburu desberdina badute ere. Lagungarri gisa, esan berri dugunaren ideia interesgarri bat eskaintzen du 2. irudiak. Itxaropenari zein bariantzari begira, oso antzekoak dira bi banaketak, eta beraz, μ estimatzeko itxaropena edo moda hartuta, adibidez, espero dugu pareko emaitzak lortzea. Are gehiago, oro har helburu-banaketaren antz handia dute, nahiko zentratuak egonik, bariantza txikiagoa duten arren.



2. irudia. Hurbilketa frekuentista eta bayestarra erabiliz egindako bost esperimentuetan lorturiko banaketak.

Horren guztiaren gainean, galdera logiko bat agertzen da: bi paradigmek ongi kuantifikatzen al dute ziurgabetasuna? Hau da, hurbilpen frekuentistaren bidez lorturiko estimatzailearen banaketak eta hurbilpen bayestarraren bidez lorturiko *a posteriori* banaketak ideia zentzuzko bat eskaintzen al dute benetako balioaren inguruan? Izan ere, 2. irudiko emaitzetan badirudi, oro har, hurbilpen frekuentistan zein bayestarrean oinarrituriko analisietatik informazio aberatsa erauz dezakegula, baina konfiantza-sentsazio hori benetakoa edo faltsua al da? Galdera horri erantzuna ematea ez da ataza batere tribiala, eta etorkizunerako lan gisa oso interesgarria izan daitekeelakoan gaude.

6. ONDORIOAK ETA ETORKIZUNEKO LANA

Lan honetan adimen artifizialean funtsezkoa den arazo bati heldu diogu; algoritmoen (edota ereduen) konparaketaren inguruan sortzen den ziurgabetasunaren kuantifikazioa. Azken urteetan, zenbait planteamendu berri egin dira algoritmoen emaitzen azterketa estatistikoaren inguruan, asko estatistika frekuentistak eskaintzen dituen hurbilketa klasikoek (test

estatistikoek) dituzten gabeziak gainditzeko asmoarekin. Ildo horretatik, estatistika bayestarra alternatiba aberasgarri moduan proposatu dugu, emaitza sinpleak eman beharrean (bai/ez motakoak), banaketa probabilistikoak eskaintzen dituelako intereseko parametroen inguruan.

Artikulu honetan, antzerako informazioa eskaintzeko gaitasuna duen beste metodo estatistiko bat ere esploratu dugu: bootstrap ez-parametrikoa. Ikuspuntu filosofiko batetik bi metodoek ideiak ezberdinak badituzte ere, birlaginketa-metodoek egindako estimazioen inguruan dagoen ziurgabetasunaren ideia ematen digute, estimatzaileen banaketa-hurbilketen bidez. Egindako esperimentazioan bi hurbilketek ematen dituzten banaketak alderatu ditugu, datu simulatuetan zein errealetan aplikatuz.

Ondorioak garbiak dira: hurbilketa bakoitzaren helburua desberdina izanagatik ere, toki berera garamatzate biek. Hurbilpen frekuentistaren bidez lorturiko estimatzailearen banaketa hurbildua eta hurbilpen bayestarraren bidez lorturiko *a posteriori* banaketa elkarren oso antzekoak dira. Are gehiago, aukera izan dugu helburu-banaketaren antz handia dutela ere ikusteko (bereziki kokapenari begira, bariantza apur bat txikiagoa duten arren). Jakina, lan honetan egindako esperimentazioaren emaitzak oso mugatuak eta atarikoak dira, baina joera mantenduko delakoan gaude.

Hortaz, horren gainean plantzeatzen den ikerketa-lerro naturalak ziurgabetasunaren kuantifikazioarekin du lotura, bi hurbilpenen bidez egiten diren analisiek ondorio egokietara ba ote garamatzaten jakiteko. Zer esan nahi du, baina, «egokiena» izateak? Ez da erantzun azkarreko galdera, inondik ere, baina, planteatzen den bidea astiro tratatuz gero, pentsatzen dugu jakituria interesgarria erauz daitekeela. Esan genezake birlaginketa-metodoen erabilera zabalagoa egiteak (beste estimatzaile konplexuagoak erabiltzea, adibidez) eta metodo horien onuren eta gabezien eragina algoritmoen emaitzen analisian aztertzeak ere lagundu diezagukeela.

ERREFERENTZIAK

[1] W. Saeed eta C. Omlin. 2023. «Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities», *Knowledge-Based Systems*, **263**, 110-273.

[2] S. Caton eta C. Haas. 2024. «Fairness in machine learning: A survey», *ACM Computing Surveys*, **56**(7), 166.

[3] J. de Mšar. 2006. «Statistical comparisons of classifiers over multiple data sets», *Journal of Machine Learning Research*, **7**, 1-30.

[4] R. L. Wasserstein eta N. A. Lazar. 2016. «The asa's statement on p-values: context, process, and purpose», *The American Statistician*, **70**(2), 129-133.

[5] S. GREENLAND, S. J. SENN, K. J. ROTHMAN, J. B. CARLIN, C. POOLE, S. N. GOODMAN eta D. G. ALTMAN. 2016. «Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations», *European journal of epidemiology*, **31**(4), 337-350.

[6] S. N. GOODMAN, 2008, «A dirty dozen: Twelve p-value misconceptions», *Seminars in Hematology*, **45**(3), 135-140.

[7] A. BENAVOLI, G. CORANI, J. DEMŠAR eta M. ZAFFALON. 2017. «Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis», *Journal of Machine Learning Research*, **18**(77), 1-36.

[8] J. ROJAS-DELGADO, J. CEBERIO, B. CALVO eta J. A. LOZANO. 2022. «Bayesian performance analysis for algorithm ranking comparison», *IEEE Transactions on Evolutionary Computation*, **26**(6), 1281-1292.

[9] B. EFRON. 1979. «Bootstrap methods: Another look at the jackknife», *The Annals of Statistics*, **7**(1), 1-26.

[10] G. CORANI eta A. BENAVOLI. «A bayesian approach for comparing cross-validated algorithms on multiple data sets», **100**(2), 285-304.

[11] M. ROSENBLATT. 1956. «Remarks on some nonparametric estimates of a density function», *The Annals of Mathematical Statistics*, **27**(3), 832-837.

[12] B. W. SILVERMAN, 1986, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

[13] W. N. VENABLES eta B. D. RIPLEY, 2002, *Modern Applied Statistics with S*, Springer, New York.

[14] B. CALVO, J. CEBERIO eta J. A. LOZANO. 2018. «Bayesian inference for algorithm ranking analysis», *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Orria 324-325.