eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

10-12 urteko haurrek informatikari buruz duten iritzian eragiten duten faktoreen bila

Egilea

Mikele Zurutuza Renom

2019

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Lana

10-12 urteko haurrek informatikari buruz duten iritzian eragiten duten faktoreen bila

Egilea

Mikele Zurutuza Renom

Zuzendaria(k)

Olatz Arbelaitz Gallego, Edurne Larraza Mendiluze

Laburpena

Ezaguna da gaur egun teknologiarekin eta bereziki informatikarekin lotutako ikasketa eta lanbideetan dagoen emakume eskasia, baina arrazoiak eta bokazioak noiz aldentzen diren ez dago argi. Gai horri argi printza batzuk emateko asmoz, konputagailuen programazioa gaztetxoen artean zabaltzeko zenbait ekimen existitzen dira. Mundu mailan hedatu den Kode Ordua oinarri hartuta, hiru urtez hainbat ikastetxetan 10-12 urte bitarteko gaztetxoen artean bildutako galdetegietako erantzun eta marrazkien azterketa egin dugu lan honetan. Datu-meatzaritzako prozesu oso bat egin behar izan dugu horretarako. Lortutako emaitzen artean, nabarmendu nahiko genuke badirudiela gure inguruko 10-12 urteko neskek oraindik hein handi batean ikusten dutela Informatikaria neska izan daitekeela. Mutilek, ordea, oro har gizonezkoak edo identifikatu ezinezko generoa duten gizakiak imajinatzen dituzte.

It is broadly known there is lack of women in fields related to technology and computer science. However, the reason causing this is unclear. In order to tackle this issue, there are several events that take place with the aim of making kids aware of the basic concepts of computer programming. Based on a worldwide expanded project called The Hour of Code, an analysis of the answers and drawings gathered in a survey conducted among 10-12-years-old youngsters has been performed. A complete data-mining process has been undertaken with the aim of interpreting the results. As a conclusion, it seems young girls still can figure women computer scientists, but young boys don't or at least they represent them as men or as unidentifiable genre.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Bada informatika arloan eztabaidagarri bihurtu den gai bat. Jakina denez, gaur egun informatikaren munduan sartzen direnen kopurua urria da beste arlo batzuekin alderatuta. Are gehiago, emakumearen partaidetza eskasak badu zeresana. Ondorioz, kezka handia dago zabalik.

Hainbat ikerketek agerian utzi dute emakumeek arlo honetan izan dezakeen konfiantza falta [Medel and Pournaghshband, 2017]. Informatikari baten estereotipoarekin bat ez egitea edota ingurugiro maskulinoa dira hauentzat mehatxu nagusiak esparru honetan sartzerakoan. Baina, zein unetatik aurrera aldatzen da pertsonen ikuspuntua? 10-12 urteko haurren artean, desberdina al da informatikariari buruz duten ikuspuntua generoaren arabera?

Litekeena da desberdintasunak egungo gizartearen eraginez sortzea, are zehatzago, egungo haurrek jasotzen duten hezkuntzarengatik. Mundu-mailan informatika derrigorrezko hezkuntzan txertatzeko mugimendu batzuk badauden arren, ez dira asko. Aipagarrienen artean, Erresuma Batua [Brown et al., 2014], Lituania [Dagiene, 2002] eta Finlandia [DeRuy, 2017] daude. Dena den, leku askotan ematen ari dira aurrerapausoak eta horietan, *Hour of Code* proiektuak rol aipagarria izan du [Code.org, a] [Code.org, b]. Proiektu hori 180 herrialde baino gehiagotan erabilia izan da eta bere helburua informatika eta programazioa gaztetxoen artean ezagutaraztea da.

UPV/EHUko Informatika Fakultateko hainbat ikasle eta irakaslek Kode Ordua aurrera eraman dute zenbait ikastetxetan. Ordubetez modu dibertigarri eta dinamikoan programatzen aritzea eta programazioko oinarrizko kontzeptuak barneratzea da Kode Orduaren helburua. Horretarako, code.org (www.code.org) tresna erabiltzen da, zeina 2015-16 ikasturtean euskarara itzuli zen UEUko Informatika Sailaren eta EHUko Donostiako Informatika Fakultatearen arteko elkarlanari esker.

Kode Ordua hiru urtez jarraian gauzatu da 10-12 urte bitarteko haurrekin. 2015/2016, 2016/2017 eta 2017/2018 ikasturteetan eraman da aurrera, zortzi, zazpi eta hamabi ikastetxeren parte hartzearekin, hurrenez hurren. Ekimen hau baliatu da haurrei inkesta bat

pasa eta informatikari buruz duten ezagutza eta ikuspuntua zein den aztertzeko. Horrez gain, inkestetan jasotako datuen bidez, ikasle bakoitzaren inguruak horretan duen eragina ere aztertu nahi izan da.

Lan honen helburu nagusia inkesta horretako emaitzak aurreprozesatu eta ikasketa automatikoaren bidez hainbat galdera erantzuten saiatzea da. Hala nola, bereizi al daitezke 10-12 urteko neskak eta mutilak informatikari buruz duten ikuspuntuagatik? Hurrengo ataletan, proiektuaren diseinua, lana aurrera eramateko haintzat hartutako oinarri teorikoak, erabilitako tresnak, proiektuaren garapena eta ateratako ondorioak azalduko dira.

2. KAPITULUA

Proiektuaren Helburuen Dokumentua

Atal honetan proiektuaren helburura iristeko plangintza zehaztuko da; alde batetik, proiektuko atazak zein diren deskribatuko da, eta bestetik, horiek garatzeko aurreikusten den denbora ere estimatuko da. Ondoren, proiektuan gerta daitezkeen arriskuak ere izendatuko dira. Azkenik, proiektu osoa burututakoan, aurreikusitako denbora-estimazioak eta errealitatean erabilitako denbora alderatu eta zenbait desbiderapenen inguruan haunaskerta egingo da.

2.1 Helburuen deskribapena

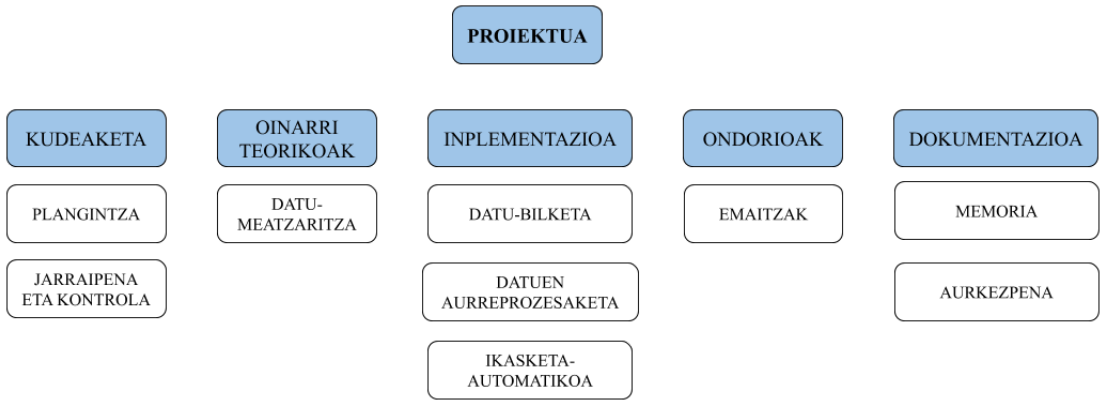
Zaila da jakitea zein diren informatikaren mundu zabalari buruzko haur baten iritziari eragin diezaioketen egunerokotasuneko faktoreak. Alde batetik, gerta daiteke haurrek eskolan informatikari buruz eskuratzen duten informazioa eskasegia edota desegokia izatea. Baina, bestalde, posible da etxean zein eskolatik kanpoko beste edozein esparrutan lortutako jakintzek ere zerikusia izatea.

Lan honen helburu nagusia aurretik aipatutako inkestetan lortutako emaitzak aurreprozesatu ondoren, ikasketa automatikoaren bidez hainbat galdera erantzuten saiatzea da. Hala nola, berdin irudikatzen al dute neskek eta mutilek informatikari bat? Edo, informatikariei buruz duten irudiak ba al du zerikusirik *freaky* estereotipoarekin? Zenbateraino irudikatzen dituzte emakume informatikariak adin txikiko pertsonek?

Horretarako, datu-meatzaritzako prozesu oso bat jarraituko da, ikasketa automatikoko zenbait teknika erabiliz. Helburua hau da: behin bildutako datuak aurreprozesatu eta ikasketa prozesurako prest utzi ondoren, datuak sailkatzea. Ondoren, ateratako emaitzak interpretatu eta proposatutako galderei azalpen bat eman ahal izateko.

2.2 LDE diagrama

Proiektua zenbait ataletan banatu da. Atal horiek proiektua zati sinpleagotan antolatzea ahalbidetu dute, azken batean, parte bakoitzak helburu egingarriago bat izan dezan. Horrez gain, aipatu beharra dago ataza bakoitzari denbora bat estimatu zaiola, behin proiektua amaituta, aurreikusitako epeak bete diren ikusi ahal izateko. Hona hemen LDE diagrama eta ataza bakoitzaren azalpen labur bat.



2.1 Irudia: *LDE diagrama.*

Kudeaketa (K): Proiektua aurrera eraman ahal izateko plangintzaren prestaketa eta proiektua aurrera doan heinean plangintza betetzen doan kontrolatu.

Oinarri teorikoak (OT): Datu-meatzaritzako prozesua ulertu eta ikasketa automatikorako erabili nahi diren algoritmoak ezagutu. Ondoren, erabiliko direnak hautatu.

Inplementazioa (I): Bildutako datuak aurreprozesatu, eta hautatutako algoritmoak aplikatu.

Ondorioak (O): Algoritmoak aplikatuz lortutako emaitzetatik ondorioak atera.

Dokumentazioa (D): Proiektuan zehar egindako urratsen azalpen-txostena idatzi eta proiektua bera laburtzen duen aurkezpena idatzi.

2.3 Atazen deskribapena

Ataza bakoitza, jarraian azaltzen diren lan-paketetan banatu da.

Kudeaketa (K): Plangintza (P)

P.1: Proiektuaren helburuak finkatu.
P.2: Proiektuaren plangintza idatzi.

Kudeaketa (K): Jarraipena eta kontrola (JK)

JK.1: Proiektuaren arriskuak identifikatu.
JK.2: Definitutako plangintza betetzen den kontrolatu.

Oinarri teorikoak (OT): Datu-meatzaritza (DM)

DM.1: Datu-meatzaritzako prozesua ulertu.
DM.2: Ikasketa-automatikoak nola funtzionatzen duen ulertu.
DM.3: Gainbegiratutako ikasketa eta gainbegiratu gabeko ikasketa bereiztu eta ulertu.
DM.4: Proiekturako erabilgarriak diren algoritmoak hautatu.

Inplementazioa (I): Datu-bilketa (DB)

DB.1: Proiektua inplementatzeko beharrezko datu-baseak eskuratu.

Inplementazioa (I): Datuen aurreprozesamendua (DA)

DA.1: Datu-baseen bateraketarako erabakiak hartu.
DA.2: Datu-baseei erabakitako aldaketak aplikatu.
DA.3: Azterketa estastitisko orokorra egin.

Inplementazioa (I): Ikasketa-automatikoa (IA)

IA.1: Gainbegiratutako ikasketarako algoritmoak aplikatu.
IA.2: Gainbegiratu gabeko ikasketarako algoritmoak aplikatu.

Ondorioak (O): Emaitzak (E)

E.1: Gainbegiratutako ikasketaren ondorioak atera.
E.2: Gainbegiratu gabeko ikasketaren ondoriak atera.

Dokumentazioa (D): Memoria (M)

M.1: Proiektuaren memoria idatzi.
M.2: Zuzendarien zuzenketak egin.

Dokumentazioa (D): Aurkezpena (A)

A.1: Proiektuaren ideia nagusiak bildu.
A.2: Aurkezpena prestatu.

2.4 Denbora-estimazioak

Hona hemen proiektuaren garapenerako aurreikusi diren atazen denbora-estimazioak. Ikus **2.1 Taula**

Ataza	Lan-paketea	Dedikazioa (h)	Guztira (h)
K	P JK	10 5	**15**
OT	DM	15	**15**
I	DB DA IA	5 80 100	**185**
O	E	50	**50**
D	M A	70 10	**80**
GUZTIRA			**345**

2.1 Taula: *Denbora-estimazioen* taula.

2.5 *Gantt* diagrama

2.2 Irudian proiektuaren *Gantt* diagrama azaltzen da.

2.2 Irudia: *Gantt diagrama*.

2.6 Arriskuak

Honelako lan zabal bat garatzerakoan aurrera eramateko oztopo izan daitezkeen arazoak suerta daitezke. Kasu honetan bi dira nabarmenenak:

1. **Datuak falta izatea:**

   Gerta daiteke bildutako galdetegietan erantzun gabeko datuak izatea. Ondorioz, hutsuneren bat duen galdetegiak nola kudeatu erabaki behar da.

   Konponbidea: Arrisku horri aurre egiteko, hainbat teknika daude; beraz, horiek aztertu eta datu-moten arabera, metodo bat aplikatu beharko da.

2. **Proposatutako galderak erantzun ahal ez izatea:**

   Posible da bildutako datuak informazio nahikoa ez izatea erantzun nahi ditugun galderak erantzuteko. Hau da, agian jasotako informazioa egokia da, baina gerta daiteke informazio gehiago edo datu zehatzagoak behar izatea.

   Konponbidea: Kasu horretan, azterketa beste modu batera planteatu beharko litzateke, edo bildutako informazioa desberdina izan beharko litzateke. Esaterko, haurrek heldu izandakoan zein ogibide izan nahiko luketen galdetzea aukeretako bat da.

2.7 Desbideratzeak

Atal honetan, proiekturako aurreikusitako dedikazioaren denbora-estimazioak errealitatearekin alderatu dira. Ondorioz, zenbait desbiderapen azaldu dira.

2.2 Taulan ikus daitekeen bezala, desbiderapen handiena **Inplementazioa (I)** atalean gertatu da. Funtsean, esan daiteke datu-base osoan oinarrituz eta zenbait moldaketa eginez, azpi datu-base asko sortu direla. Beraz, kontu handiz ibili behar izan da, ikasketa-automatikoa aplikatzerakoan buruhausteak gerta ez zitezen. Ikasketa-prozesuko atal batzuk behin baino gehiagotan hasi behar izan dira hasieratik. Izan ere, xehetasun txikiek desberdintzen zuten datu-base bat bestetik. Horren arira dator **Ondorioak (O)** ataleko desbiderapena. Beste era batera esanda, ez zegoen aurreikusita datu-base okerrarekin emaitzak ateratzea. Beraz, behin akatsaz ohartuta, berriro iritsi behar izan da beste ondorio batzutara.

Bestalde, **Dokumentazioa (D)** ataleko desbiderapenari dagokionez, datu-meatzaritzari eta ikasketa-automatikoari buruz esan daitekeenak ez du mugarik, eta hortaz, denbora behar izan da gehienbat proiektu honetarako oinarri teoriko nahikoak eta ulergarriak hautatzen. Horrekin lotuko nuke **Oinarri Teorikoak (OT)** ataleko desbiderapena.

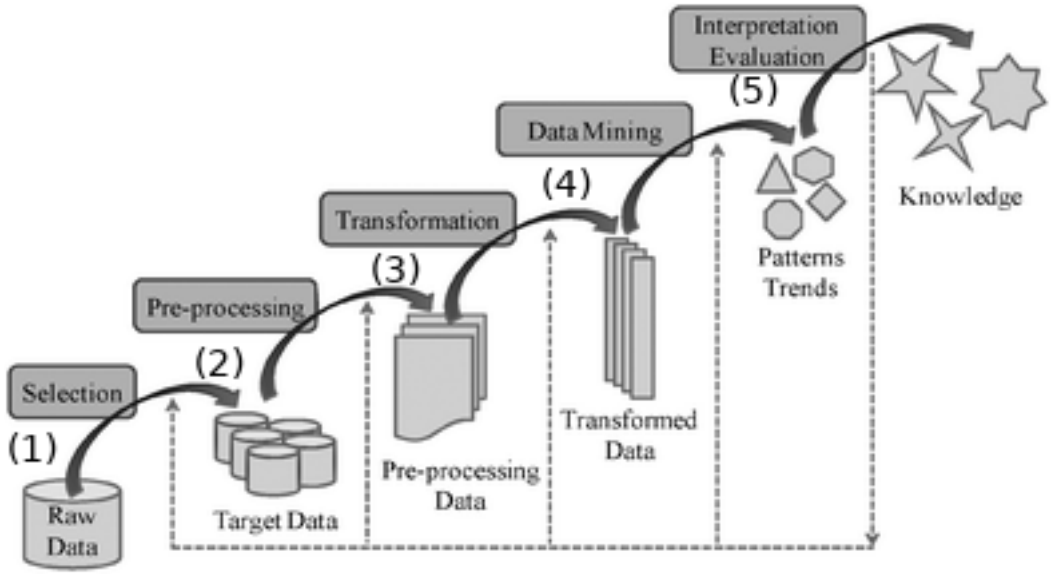
Ataza	**Lan-paketea**	**Dedikazioa (estim.)**	**Guztira (estim.)**	**Dedikazioa (erreala)**	**Guztira (erreala)**	Desbiderapena
K			**15**		**13**	**-2**
	P	10		8		-2
	JK	5		5		0
OT			**15**		**20**	**+5**
	DM	15		20		+5
I			**185**		**207**	**+22**
	DB	5		2		-3
	DA	80		95		+15
	IA	100		110		+10
O			**50**		**55**	**+5**
	E	50		55		+5
D			**80**		**70**	**-10**
	M	70		60		-10
	A	10		10		0
GUZTIRA			**345**		**365**	**+40**

2.2 Taula: *Desbiderapenen* taula.

3. KAPITULUA

Azalpen teorikoak

Mundu errealean, zenbait kasutan inkesta edo galdetegi bidez datu asko jasotzen dira. Bildutako emaitzetatik ezagutza lortzeko eta datuak interpretatuz informazio adierazgarria eskuratzeko, *Knowledge Discovery in Databases* (KDD) prozesu oso bat egiten da [Fayyad et al., 1996b]. Horretarako, ikasketa-automatikoko zenbait algoritmo aplikatu behar izaten dira [Clifton, 2017]. Labur esanda eta **3.1 Irudiari** erreparatuz, prozesu hori pauso hauetan oinarritzen da: lehenengoan, (1) bildutako datuen (2) aurreprozesaketa eta bateraketa egiten da. Bigarrenean, (3) ikasketa-automatikoko tresnek erabili ahal izateko prestakuntza, eta ondoren, (4) azalpena eman diezaguketen datu-meatzaritzako metodoak aplikatzen dira. Azkenik, (5) metodo horiek aplikatu ostean lortzen diren emaitzak interpretatuz, zenbait ondoriotara iristea da helburua.



3.1 Irudia: *Datu-meatzaritzako* prozesuaren urratsak [Figueiredo et al., 2016].

3.1 Datu-bilketa

Lehenik eta behin, prozesuaren helburu nagusia zehaztea garrantzitsua da, zer aztertu nahi den definitzea, hain zuzen ere. Horretarako, lagungarri izan daitekeen informazioa zein den determinatu behar da, datu horiek jasotzeko modua ere zehaztuz. Azken finean, prozesuan aurrera jarraitzeko datuak batu behar dira, helburua datu-base bat osatzea izanik.

Kasu bakoitzerako informazio aproposa lortzeko hainbat era daude [Osang et al., 2013]: egitura desberdineko galdera-sortak (aurrez aurreko elkarrizketa, telefono bidezko galdeketa, etab.), egoera zehatz batean egindako esperimentuen behaketak, arlo espezifiko baten inguruan egindako galdeketak eta aurretik existitzen diren datu-baseen azterketak [Johnson and Turner, 2003, Chapter 11], besteak beste.

3.2 Datuen aurreprozesaketa

Prozesuko atal honetan, datuak prestatzen dira: datuak ikasketa-automatikoko zenbait teknika aplikatzeko prest egon daitezen egiten da. Beste era batera esanda, urrats hau bildutako informazioa txukuntzean datza. Izan ere, normalean, mundu errealetik jasotako informazioa ez da guztiz "garbia" izaten eta beraz, zenbait egokitzapen egin behar izaten dira. Kontuan izan behar da ikasketa-automatikoan erabiltzeko sortu nahi den datu-basea jatorri desberdinetako azpi datu-baseetatik sor daitekeela. Hots, informazio berdina modu desberdinetan jaso daiteke, eta ondorioz, jasotako datuak desberdinak izan daitezke. Balioa jaso duen galdera bakoitzari ***atributu*** deritzo, prestatu nahi den datu-baseak zutabe gisa izango dituen elementuak, hain zuzen. Hortaz, haien bateraketa egitea beharrezkoa da. Mundu errealean atributuen adibide gisa, pertsona baten ezaugarriak daude; esaterako, pertsona baten adina, generoa edota izena, atributu gisa definitu ahalko lirateke.

Datuen prestaketan gauza asko dira kontuan hartu beharrekoak. Orokorrean, prozesuaren amaieran erantzun nahiko liratekeen galderak kontuan hartzea beharrezkoa da. Horretaz gain, atributuen kodeketa aintzat hartu beharreko xehetasuna da, eta beharbada, aldatu beharrekoa. Hau da, atributuek jaso dezaketen balio posibleak azpi datu-base guztietan berdinak izatea komeni da eta horren arabera, datu-basearen egitura berdina izan dadin, erabaki desberdinak hartu eta inplementatu. Gainera, sarrerako datu-baseko ***atributuen mota*** jakitea ere garrantzitsua da. Atributu bat, nominala, bitarra, ordinala edota zenbakizkoa izan daiteke [Jiawei Han, 2011, Chapter 2.1]. Bestalde, bildutako datuen gainean

azterketa estatistiko bat egitea ere komeni da, bereziki, jasotako balioak zein ehunekotan banatzen diren ikusteko.

Datuen prestaketaren zati ere bada datu-baseak osorik daudela ziurtatzea. Falta diren balio horiek tratatzeko, zenbait metodo daude eta ondoren aipatuko dira.

3.2.1 Balio-hutsak (*Missing values*)

Ohikoa da osatu gabeko datu-baseekin topatzea. Izan ere, mundu errealeko datuak ez baitira sarri osorik jasotzen. Falta diren balioei ***balio-hutsak*** edo *missing values* deritze, eta modu desberdinetan trata daitezke [Jiawei Han, 2011, Chapter 3.2]. Horietan egokiena aukeratzea, datuen domeinuaren eta datu-analisiaren helburuaren araberakoa izaten da. Besteak beste, hauek dira aukera posible batzuk:

1. **Balio-hutsak dituzten errenkadak ezabatu:** Errenkada batean atributu askoren informazioa falta bada, edota bereziki esanguratsua kontsideratzen den balio bat falta bada, normalean baztertu egiten da. Hala ere, ez da oso baliozkoa ezabatu beharreko datu kopurua oso handia den kasuetan.

2. **Konstante globalak erabili:** Falta den balio bat beste batekin ordezkatzea eraginkorra ez bada, “Erantzun gabe” bezalako aldagai konstante batekin ordezkatzea da egokiena. Hala ere, emaitza hori ohikoena izaten bada, ikasketa automatikoko prozesuan gaizki-ulertuak suertatu daitezke.

3. **Mediana edo batez-besteko balioarekin ordezkatu:** Zenbakizko datuei dagokienez, batezbestekoa ateratzea aukera ona da. Bestelako datuez ari bagara, berriz, maiztasun handienarekin jaso den erantzunarekin ordezkatzea izango litzateke egokia.

Dena den, badaude aukera adimentsuagoak eta, ondorioz, konplexuagoak direnak. *Multiple Imputation* (MI) [Rubin, 1978] algoritmoa, adibidez, balio-hutsen arazoa ebazteko baliozko metodo bat da. Metodologia horren bitartez, datu-baseko hutsuneak atributuaren balio posible desberdinekin betetzen dira, eta beraz, balio-hutsik gabeko datu-multzoak osatzen dira. Horietako multzo bakoitza estatistikoki analizatu eta emaitzak konbinatuz, inferentziak sortzen dira [Saar-Tsechansky and Provost, 2007].

3.2.2 Azterketa estatistikoa

Datuen aurreprozesaketa eraginkorra izan dadin, ezinbestekoa da batutako informazioaren irudi orokor bat izatea. Oinarrizko azterketa estatistiko batek zenbait datu esanguratsu azpimarratzen ditu. Izan ere, estatistikaren helburua ere bada datuak informazio bihurtzea. Datu-base zabal batetik atera daitezkeen datu-estatistikoak honakoak izan daitezke; esaterako, lortutako balioen ***batezbestekoa***, ***mediana*** eta ***moda***. Medianak, atributu batek jaso dituen zenbakizko datuak txikienetik handienera ordenatuz, erdian kokatzen den balioa adierazten du. Modak, berriz, atributu bakoitzerako nabarmendu den portaera globala edo gehien jasotako erantzuna adierazten du [Jiawei Han, 2011, Chapter 2.2].

3.3 Ikasketa-automatikoko algoritmoen aplikazioa

Datu-meatzaritzaren helburua datu-multzo batetik informazio erabilgarria ateratzea da. Zehatzago esanda, datu-multzo erraldoietan eredu adierazgarri eta ulergarriak aurkitzen dira, ondoren lortutako ereduak interpretatu eta azalpen bihurtzeko [Fayyad et al., 1996a]. Prozedura hori datu-meatzaritzako algoritmoen aplikazioan oinarritzen da, hau da, ikasketa automatikoko metodoez baliatuz egiten da.

3.3.1 Gainbegiratutako ikasketa

Gainbegiratutako ikasketa (*Supervised Learning*, ingelesez), datuen sailkapenean oinarritutako datu-meatzaritzako metodo bat da. Bere helburua kasu (edo instantzia) berrien klasea identifikatzea da. Datu-base bat zenbait zutabe eta errenkadaz osatutako fitxategi bat dela kontsideratuz, kasu edo instantzia bat datu-fitxategiko errenkada gisa definitu daiteke. Datuen sailkapenean, ***klasea*** datu-basean hautatu daitekeen atributu zehatz bat da. Horren balioa datuen sailkapenaren bidez aurresan behar da, instantzia berdineko gainerako atributuen balioak aintzat hartuz. Beste era batera esanda, sailkapen metodo horrek, zenbait instantzia dituen datu-base bat oinarritzat hartuz eta instantzia bakoitzaren klasea ezaguna izanik, kasu berrien klasea iragartzea du helburu. Prozesu horren lehenengo urratsean, sailkatzailea bera eraikitzen da eta ikasketa deritzo, eta bigarren urratsean, berriz, instantzia berrien klasea identifikatzen da, hots, sailkapena egiten da [Jiawei Han, 2011, Chapter 8].

Ikasketa (edo entrenamendu) fasean, sailkapen-algoritmo batek sailkatzailea eraikitzen

du, datu-basetik ateratako entrenamendurako kasu batzuk eta dagozkien klaseak analizatuz eta horietatik "ikasiz". Horretarako, erregela-multzo, sailkapen-zuhaitz edo formula matematikoen erabileraz baliatzen da. Metodo horiek entrenamendurako multzoa ez den beste instantzia batzuen klasea iragarriko dute. Entrenamendurako kasuen klasea ezaguna denez, gainbegiratutako ikasketa deritzo. Adibideen klasea ezezaguna denean, aldiz, geroago aztertuko den ***gainbegiratu gabeko ikasketa*** egiten da.

Gainbegiratutako ikasketan, hasteko, sailkatzaileak iragarpena egiteko duen gaitasuna estimatzen da. Estimazio hori kalkulatzeko modu bat, sailkatzailearen ***asmatze-tasa*** estimatzea da [Ali and Smith, 2006]. Aurrerago aipatuko dira existitzen diren beste modu batzuk. Horretarako, entrenamendurako kasuen desberdinak diren testerako kasu batzuk erabiltzen dira. Sailkatzaile baten asmatze-tasak egoki sailkatutako testerako kasuen ehunekoari egiten dio erreferentzia. Testerako adibide bakoitzari dagokion klasea, entrenamendu fasean ikasitako iragarpenarekin alderatzen da, eta konparaketa horren ondorioz sailkatzailearen asmatze-tasa determinatzen da [Jiawei Han, 2011, Chapter 8.1].

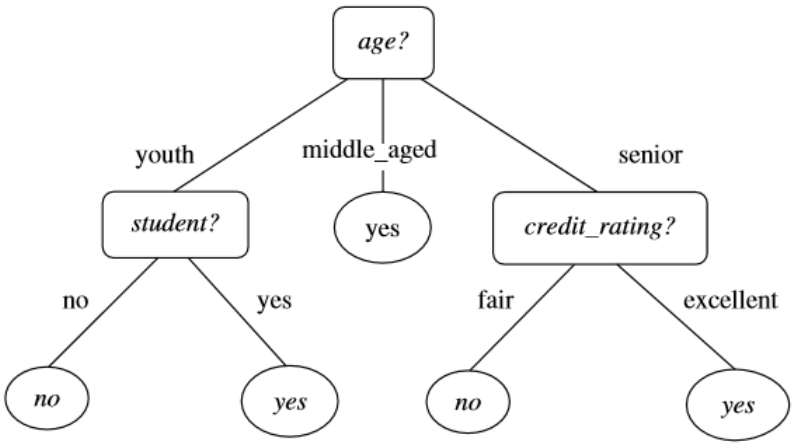
Datuen sailkapenean erabiltzen diren oinarrizko teknikei dagokienez, sailkapen-zuhaitzak, erregeletan oinarritutako sailkatzaileak edota sailkatzaile Bayestarrak adibide onak dira, beste batzuen artean. Dena den, horietako bakoitzaren eraginkortasuna konparatuz, datubase desberdinen gainean erabiltzea komeni den algoritmoa hautatzen da, ez baita existitzen teknika optimorik.

Sailkapen-zuhaitzak

Sailkapen-zuhaitzen bidez egindako indukzioa diagrama-egituran oinarritutako erabaki-zuhaitzen ikasketan datza. Ikasketa ezagunenen artean, ID3 algoritmoa dago; hartatik hedatutako C4.5 algoritmoa [Quinlan, 1993], hain zuzen ere. Entrenamendurako datu-multzo batetik abiatuz, zuhaitz bat osatzen da. Zuhaitz horren adabegiek datu-baseko atributuak adierazten dituzte. Horietatik hedatzen diren adarrek atributu horien balio posibleei egiten diete erreferentzia, azken batean klase-aldagaia adierazten duen azken adabegietara (hostoetara) iritsiz. Zuhaitz horiek goitik-beherako noranzkoan eraikitzen dira, banatu eta irabazi teknika erabiliz. Zuhaitzeko erroan datu-base osoa hartzen da kontuan eta, hurrengo banaketetarako, datu-basea datu-multzo desberdinetan banatzen joaten da, atributuaren balioen arabera. Adabegi bakoitza zatitzeko erabiliko den atributua hautatzeko metodo batzuk existitzen dira, eta aurrerago izendatuko ditugu.

3.2 Irudian sailkapen-zuhaitz baten adibidea azaltzen da. Adibide hau pertsona batek ordenagailu bat erosteko duen probabilitatean oinarritutako erabaki-zuhaitz bat da. Kasu ho-

netan, *age* (adina), *student* (ikaslea) eta *credit rating* (kredituaren balorazioa) datu-baseko atributuak izango lirateke. Horietatik ateratzen diren adarrak atributuek jaso dezaketen balio posibleak dira, eta azken adabegiek pertsona batek ordenagailu bat erosiko duen edo ez adierazten du.



3.2 Irudia: *Sailkapen-zuhaitz* baten adibidea [Jiawei Han, 2011, Chapter 8.2].

Esaterako, ezkerreko adarra segiz, adinez *youth* eta ikaslea den *no* erantzundakoek, ordenagailurik ez dutela erosten ulertzen da, eta ikaslea den *yes* erantzun dutenek ordenagailua erosten dutela. Zuhaitzaren erroa *age* (adina) da eta zuhaitzaren puntu horretan datu-base osoa dago. Datu-basea atributu horren balioaren arabera zatitzen da. Kasu honetan (ezkerreko bidea), *student* atributura iristen dira, adina *youth* esan dutenak. Azkenik, pertsona batek ordenagailu bat erosiko ez duela esaten da, baldin eta ikaslea ez dela esan badu, eta ikasleak direnak, berriz, erosiko dutela esaten da.

Honakoa da sailkapen-zuhaitz baten funtzionamendua sailkatzerako orduan: Testerako adibide bat emanik, zeinaren klasea determinatzeke dagoen, kasu horren atributuen balioak aurretik eratu den erabaki-zuhaitzarekin probatzen dira. Errotik hostoetaraino doan bide bat jarraitzen da, azken nodoan adibideari dagokion klasea iragarriz.

Zuhaitza eraikitzerako orduan, atributuen aukeraketa bat egiten da. Horretarako erabiltzen diren neurriak heuristikoak dira, zeinak datu-multzo bat klase indibidualen iragarpenetara iristeko “ondoen” banatzen duen irizpidea hautatzen duen. Irizpide “onenak” zera esan nahi du: partiziorako aukera onena, ondorioz, banatzen diren bide horiek jarraituko dituzten instantziek klase bera izateko probabilitatea izango dute. Beste era batera esanda, datuak hoberen sailkatzen dituen atributua hautatzen da zuhaitzeko erro gisa, eta proze-

su bera aplikatzen da goitik beherako gainerako atributuekin [Jiawei Han, 2011, Chapter 8.2]. Atributuen aukeraketarako zenbait neurri existitzen dira, eta aurrerago aipatuko dira.

ID3 algoritmoak [Quinlan, 1986] Claude Shannon-en laneko ***informazio-teorian*** (*information theory*, ingelesez) oinarritutako irizpidea erabiltzen du atributuen hautaketarako [Shannon, 1948]. Irizpide horrek mezuen "informazioaren edukia" edo balioa aztertzen du. Behin atributu bakoitzaren informazioaren balioa kalkulatuta, balio handiena duena hautatzen da. C4.5 algoritmoak, ID3 algoritmotik hedatutakoak, informazio-teorian oinarritutako neurri bat erabiltzen du, ***irabazi-ratio*** (*gain ratio*, ingelesez) izenez ezaguna. Hori ***informazio-irabazia*** edo *information gain* neurriaren hedapen bat da. Informazio-irabaziak honako balio posible asko jaso dezaketen atributuak hautatzen ditu. Esate baterako, demagun atributu gisa pertsonen NAN zenbakia hautatzen dela. NAN zenbakia atributuaren gaineko banaketa pertsonen identifikaziorako zenbaki adina adarretan egingo litzateke. Banaketa hori purua izango litzateke, eta ondorioz, atributu horretan egindako banaketarekin lortutako informazio-irabaziaren balioa, maximoa. Baina, argi dago horrelako partizio bat ez dela egokiena sailkapenerako.

Aurretik aipatutako irabazi-ratioak zera egiten du: informazio-irabaziak balio askoko atributuak aukeratzeko duen joera saihesten saiatzen da. Horretarako, atributu baten gaineko balio posible bakoitzerako, balio zehatz hori duten kasuak hartzen ditu kontuan, datubaseko kasu kopuru totalarekin batera. Azken batean, irabazi-ratio altuena duen atributua aukeratzen da banaketarako [Jiawei Han, 2011, Chapter 8.2.2].

Sailkatzaile Bayestarrak

Sailkatzaile Bayestarrak sailkatzaile estatistikoak dira. Kasu bakoitza zein klaseri dagokion adierazten duen probabilitatea kalkulatzen dute. Izenak dioen bezala, Bayes-en teoreman oinarritzen den metodo bat da [Hanson et al., 1991]. ***Naïve Bayes sailkatzailea*** teorema horretatik hedatu zen eta sailkapen-zuhaitzen edo neurona-sare artifizialen efizientziarekin konparagarri izan daiteke. *Naïve Bayes* sailkatzaileak dio atributu batek zehaztutako klasearekiko duen eragina, gainerako atributuek dutenarekiko independentea dela. Baldintzapeko independentziaren hipotesi hori gutxitan betetzen da mundu errealeko aplikazioetan; horren ondorioa da metodoari emandako "*naïve*" terminoa [Patil et al., 2013].

Demagun $\boldsymbol{X}$ atributuen informazioa duen adibide bat dela. Bayesen teoremari dagokionez, kasu hori "ebidentzia" da. H, berriz, hipotesi bat da, zeinak $\boldsymbol{X}$ adibidea C klase zehatz bati dagokiola esaten duen. Sailkatzerakoan, P($H|\boldsymbol{X}$) probabilitatea determinatu nahi da;

$\boldsymbol{X}$ "ebidentzia" izanik, H hipotesiak duen probabilitatea, hain zuzen ere. Beste era batera esanda, $\boldsymbol{X}$ osatzen duten atributuen deskribapena ezaguna dela suposatuz, $\boldsymbol{X}$ adibideari dagokion klasea C izateko probabilitatea.

Naïve Bayes sailkatzaileak honela funtzionatzen du. Demagun entrenamendurako datu-multzo bat eta horiei esleitutako klasearen balioak ditugula. Suposa dezagun m klase daudela, $C_1, C_2, ..., C_m$. $\boldsymbol{X}$ kasu bat emanik, sailkatzaileak aurresango du $\boldsymbol{X}$ adibideari P($C_i|\boldsymbol{X}$) balio altuena duen C_i klasea dagokiola. Hau da, *Naïve Bayes* sailkatzaileak $\boldsymbol{X}$ adibidea C_i klasekoa dela aurresango du, baldin eta soilik baldin

$$P(C_i|\boldsymbol{X}) > P(C_j|\boldsymbol{X}), \qquad non \quad 1 \leq j \leq m, j \neq i.$$

Ondorioz, P($C_i|\boldsymbol{X}$) maximizatu behar da:

$$P(C_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_i)P(C_i)}{P(\boldsymbol{X})}$$

non P($\boldsymbol{X}|C_i$) probabilitateak C klasea jakinda, $\boldsymbol{X}$ kasua gertatzeko probabilitatea estimatzen duen. P(C_i) datu-multzoko kasuak kontuan izanik C_i klasekoa izateko probabilitatea da, eta P($\boldsymbol{X}$) klase guztientzat konstantea den probabilitatea, zeinek datu-multzoko kasuek $\boldsymbol{X}$ adibidearen balioak izateko probabilitatea adierazten duen.

Hori horrela izanik, P($C_i|\boldsymbol{X}$) probabilitatea maximizatu ahal izateko, P($\boldsymbol{X}|C_i$) P(C_i) maximizatu behar da. $\boldsymbol{X}$ adibidearen klasea iragartzeko, probabilitate hori C_i klase bakoitzerako estimatzen da. Sailkatzaileak aurresango du $\boldsymbol{X}$ instantziaren klasea C_i dela, baldin eta soilik baldin

$$P(\boldsymbol{X}|C_i)P(C_i) > P(\boldsymbol{X}|C_j)P(C_j), \qquad non \quad 1 \leq j \leq m, j \neq i.$$

Metodoaren eraginkortasunari dagokionez, teorikoki, sailkaitzaile Bayestarrak errore-tasa txikiena dutenak dira, gainerako sailkatzaileekin alderatuz. Praktikan, aldiz, ez da beti horrela izaten. Izan ere, zehazgabetasunak nabarmentzen dira algoritmoa aplikatzerakoan egiten diren hipotesietan; hala nola, aurretik aipatutako baldintzapeko atributuen independentzian. Hala ere, sailkatzaile mota horiek azalpen teoriko ulergarriak emateko aukera ematen dute [Jiawei Han, 2011, Chapter 8.3].

Erregela-multzoak

Erregeletan oinarritutako sailkatzaileak baldintza-ondorio gisa errepresentatzen dira. Alde batetik, erregelaren baldintzaren zatia zenbait atributuz osatzen da, AND logikoaren bidez lotuz, eta bestetik, erregelaren ondorioak klase iragarpen bat du [Frank and Witten, 1998, Chapter 3.4]. Baldintza zati osoa egiazkoa bada sailkatu gabeko kasu batentzat, erregela egokia dela esaten da.

Aurretik aipatu bezala, sailkapen-zuhaitzak nahiko ulergarriak dira. Dena den, lanerako datu-multzoa oso handia den kasuetan, azalpenak bilatzea zaila izan daiteke. Kasu horietan, sailkapen-zuhaitzetatik ateratako erregelak ulerterrazagoak direla kontsideratu daiteke. Hortaz, esan daiteke azalpenak bilatu nahi direnean, erregelak sailkapen-zuhaitzen alternatiba ona direla.

Erabaki-zuhaitz batetik erregela bat lortzeko, zuhaitzeko errotik hostoraino doan bidea jarraitu behar da. Adabegien (atributuen) banaketa bakoitzeko AND logiko bat gehitzen da erregelan, baldintza aldea osatu dadin. Behin azken adabegira (hostora) iritsita, hura izango da erregelako ondorio zatia [Jiawei Han, 2011, Chapter 8.4]. Esate baterako, PART algoritmoa, erregelen algoritmo bat da, sailkapen-zuhaitzetatik erregelak ateratzen dituena eta erregela horiek ikasteko, banatu eta irabazi teknika erabiltzen duena. Izan ere, sailkapen-zuhaitz partzialak osatzen ditu; guztiz eratu gabeko C4.5 zuhaitzak, hain zuzen ere [Parsania et al., 2014].

Sailkapen-zuhaitzak ez dira erregelak lortzeko aukera bakarrak. Baldintza-ondorioko erregelak, *sequential covering* algoritmoaren bidez ere induzitu daitezke, non erregelak entrenamendurako datu-multzo batetik sekuentzialki (banan-banan) ikasten diren. Klase zehatz bati dagokion erregela batek klase horri dagozkion hainbat adibiderekin bat egingo du [Jiawei Han, 2011, Chapter 8.4.3]. Erregela bat ikasten den bakoitzean, erregela betetzen duten instantziak datu-multzotik ezabatzen dira, eta prozesua behin eta berriz errepikatzen da gainerako datuekin. Metodo ezagunenen artean, RIPPER algoritmo induktiboa dago [Cohen, 1995].

Erregelak orokorretik-zehatzerako joera izanik ikasten dira. Hau da, erregela bat ikasteko, ahalik eta erregela orokorrenetik hasten da. Kasu honetan, baldintza aldea hutsa dela suposatzen da, ondoren zehaztutako ondorioa betetzen duten baldintzak gehituz joateko.

k-NN sailkatzaileak

Orain arte aipatutako sailkapen-metodoak *eager learners* (ikasketa gogotsua, euskaraz) gisa kontsideratzen dira. Hau da, entrenamendurako kasu batzuk izanik sailkatzailea bera eratzen dute testerako kasu berriak izan baino lehen. *Lazy learner*-ek (Ikasketa alferrak, euskaraz), ordea, entrenamendurako kasuak jaso eta gero, adibide horiek gorde eta testerako kasuak jaso arte itxaroten dute. Ondoren, adibide berriak aurretik gordetakoekin alderatzen dituzte sailkapena egin ahal izateko [Jiawei Han, 2011, Chapter 9.5].

Ikasle alferraren adibide ezagunak dira k-NN sailkatzaileak (*k-Nearest-Neighbour Classifiers*). Esan bezala, testerako kasu berri bakoitza gordetako kasuekin alderatzen da eta haiekiko duen antzekotasuna kalkulatzen da. Antzekotasun edota gertutasun hori kalkulatzeko, atributuen arteko distantzia euklidearrean oinarritzen da. Horregatik, sailkatzaile mota horiek zenbakizko atributuekin aplikatzeko soilik balio dute [Peterson, 2009].

3.3.2 Sailkapenaren zehaztasuna handitzeko teknikak

Sailkatzaile baten doitasuna ezaugarri garrantzitsua da eta berau handiagotzeko badaude zenbait teknika, multisailkatzaileak deritzenak. Emaitza hobeak lortzeko, sailkatzaile soil desberdinak konbinatzen dituzte. Multisailkatzaileak sailkatzaileak berak baino zehatzagoak eta ziurragoak izan ohi dira [Jiawei Han, 2011, Chapter 8.6]. Besteak beste, *bagging* da teknika horien adibide bat. Multisailkatzaile gisa kontsideratzen ez den, baina sailkapenaren zehaztasuna handitzeko tekniken artean, *Consolidated Tree Construction* (CTC) teknika dago, sailkapen zuhaitzekin erlazionatutako algoritmo bat dena.

Testerako instantzia berri bat izanik, zeinaren klasea oraindik iragarri gabe dagoen, metodoko zati den sailkatzaile bakoitzak klase-aldagaiaren iragarpen bat egiten du. Ondoren, multisailkatzailearen “botoen” bidez, klasearen iragarpena egiten du. *Bagging* metodoaren kasuan, adibidez, sailkatzaile bakoitzaren “botoak” pisu bera hartzen du.

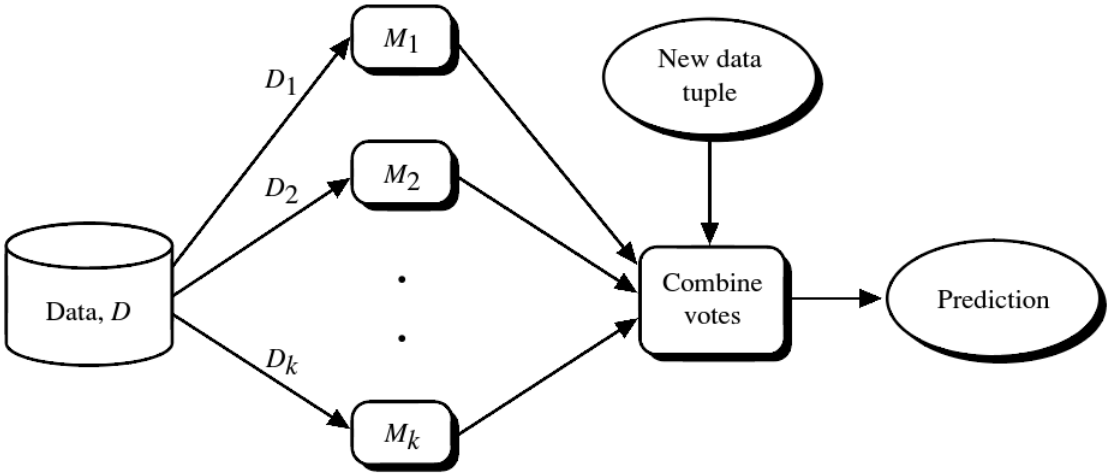
Bagging

Bagging teknika, [Breiman, 1996a] lanean aurkeztutako metodoa, adibide sinple batekin azaltzea erraza da. Demagun medikuarenera doan paziente batek bere sintometan oinarritutako diagnostiko bat jaso nahi duela. Horretarako, mediku bakar bati galdetu ordez, bat baino gehiagori galdetzea hobe izango litzateke. Diagnostiko bera zenbait medikuk iragarriz gero, hori izango litzateke aukeratuko litzatekeen iragarpena. Hau da, azken erabakia

gehiengoak bozkatu duenaren araberakoa da, non mediku guztiek bozkatzeko eskubide berdina duten. *Bagging* metodoaren ideia orokorrari dagokionez, mediku bakoitza sailkatzaile bat izango litzateke.

D datu-multzo bat emanik, *bagging* metodoak honela egiten du lan: i (i = 1, 2,...,k) iterazio bakoitzerako, d instantzia dituen D_i entrenamendurako datu-multzoa, jatorrizko D datu-multzotik erauzitako azpilagina izango da [Jiawei Han, 2011, Chapter 8.6.2]. Entrenamendurako datu-multzo hori jatorrizko datu-multzoaren tamaina berdinekoa da, baina kasu batzuk agian azaltzen ez diren bezalaxe, beste batzuk behin baino gehiagotan azaldu ahal dira. D_i azpilagin bakoitzetik ikasiz M_i sailkatzailea lortzen da. X kasu berri bakoitzerako, M_i sailkatzaile bakoitzak klase-aldagaiaren iragarpen bat egiten du, boto bat bailitzan kontsideratzen dena. M* sailkatzaileak boto guztiak kontatu eta boto gehien jaso dituen klase-aldagaia iragartzen du.

3.3 Irudian azaltzen da orain arte azaldu denaren laburpen bat, non M_1, M_2, ..., M_k sailkatzaile desberdinak diren. Sailkatzeke dagoen instantzia berri bakoitzerako (*New data tuple*), bozketaren arabera (*Combine votes*) klase-aldagaia aukeratzen da eta azkenik *Prediction*, hau da, iragarpena egiten da.



3.3 Irudia: *Multisailkatzaileen* jokaera, hainbat sailkatzaileez baliatuz [Jiawei Han, 2011, Chapter 8.6].

Consolidated Tree Construction (CTC)

Erabaki-zuhaitz bakar baten errepresentazioa nahiko intuitiboa eta erraza da ulertzeko, eta ondorioz, oso erabilgarria da azalpenak emateko. Baina, erabaki-zuhaitz askoren erre-

presentazioa, gizakien ulertzeko gaitasunetik urruti egon daiteke. Arazo horren aurrean, *Consolidated Tree Construction* (CTC) algoritmoa dago [Pérez et al., 2007a]. Metodo horrek entrenamendurako datu-multzotik zenbait azpilagin sortzen ditu nahi den klase-aldagairako. Badaude metodo batzuk adibide adina zuhaitz eraikitzen dituztenak, aurretik aipatutako *bagging* edo *boosting*, esaterako; CTC algoritmoak, ordea, sailkapen-zuhaitz bakarra induzitzen du, ulergarritasuna galdu ez dadin.

Azpilagin bakoitzak zuhaitzean kokatzea komeni den nodoa proposatzen du, ondoren hura banatu ahal izateko. Banatzeko funtzioa irabazi-ratioaren irizpidean ([Quinlan, 1993] C4.5 algoritmoan erabilitako berdina) oinarritzen da, beste edozein teknika erabili ahalko zen arren; hala nola, *Gini Index* [Mingers, 1989]. Banaketarako erabiliko den atributuaren hautaketa bozketa bidez egiten da. Prozesu hori nodoz nodo egiten da. Bozketarako zenbait modu daude; besteak beste, bozketa estandarra edota pisuaren araberako bozketa deritzona. Hautaketa horretan oinarrituz, aipatutako azpilagin guztiak atributu berdina erabiliz banatzen dira.

3.3.3 Aldagai hautaketa

Ikasketa-automatikoan, teorikoki, errepresentaziorako geroz eta atributu gehiago izanik, orduan eta sailkatzeko gaitasun handiagoa dagoela uste da. Hala ere, esperientzia praktiko askotan hori ez da kasua; gehienetan, nahiago da atributu gutxiko errepresentazio bat lortzea, non atributuen konbinazio egoki batekin, klase-aldagaia iragartzeko gai den. Datuen artean informazio "zaratatsu" eta baztergarri asko baldin badago, entrenamendurako fasea asko zailtzen da.

Ohikoa da datu-multzo handietan beharrezkoa ez den informazioa izatea, eta horrek emaitzen azalpenak konplexuago bihur ditzake. Azalpen sinpleagoak bilatzeko helburuarekin, [Garca et al., 2014] lanean gomendatzen den moduan, ikasketa automatikoaren ikuspuntutik esanguratsuenak diren aldagaiak automatikoki hautatzeko algoritmo bat proposatu da: *Correlation-based feature subset selection* [Hall, 2000] prozesuak korrelazioan oinarrituta datu-multzoan dagoen ahalik eta informazio erredundante gehiena identifikatu eta bertatik ezabatzen du. Hau da, datu-multzotik ikasteko informazio garrantzitsuena ematen duen atributuen aukeraketa bat egiten du, garrantzi gutxiko datuak ezabatuz. Prozesu horrek datu-multzoaren tamaina txikitzeaz gain, posible da ikasketarako prozesua azkartzea eta eraginkortasuna areagotzea. Kasu batzuetan, etorkizuneko sailkapenetan zehaztasuna handitu daiteke; beste batzuetan, ordea, lortutako errepresentazioari dagokionez, emaitza sinpleagoa da, eta hortaz, era errazean ulertzeko modukoa.

3.3.4 Sailkatzaileen ebaluazioa eta hautaketa

Behin sailkatzailea eratuta, garrantzitsua da jakitea nolako zehaztasun edo doitasunekin aurreikusi dezakeen sailkatzaile batek entrenatu gabeko kasu berri baten klasea. Hau da, sailkatzaileak kasu edo instantzia berriak zuzen sailkatzeko duen gaitasuna. Horretarako, alde batetik, sailkatzailea ebaluatzeko neurri batzuk behar dira, eta bestetik, ebaluazioa justua izan dadin, laginak banatzeko estrategia egokiak erabili behar dira. Esaterako, *holdout* eta balidazio-gurutzatua sailkatzaileen doitasuna neurtzeko ohiko bi teknika dira [Jiawei Han, 2011, Chapter 8.5].

Sailkatzaileen eraginkortasunaren neurketa

Badaude zenbait neurri sailkatzaile baten ontasuna ebaluatzen dutenak, besteak beste, asmatze-tasa. Termino horretan sakontzen hasi baino lehen, beste batzuetaz jabetzea komeni da; adibidez, adibide positiboak (intereseko klaseari dagozkien kasuak), P, eta adibide negatiboak (gainerako kasuak), N. Horiez gain, beste lau daude:

- *True positives* (TP): sailkatzailearen bidez zuzen sailkatutako kasu positiboak.
- *True negatives* (TN): sailkatzailearen bidez zuzen sailkatutako kasu negatiboak.
- *False positives* (FP): sailkatzailearen bidez oker sailkatutako kasu negatiboak (positibo gisa sailkatu direnak).
- *False negatives* (FN): sailkatzailearen bidez oker sailkatutako kasu positiboak (negatibo gisa sailkatu direnak).

Izendatutako terminoak konfusio-matrizea deritzon matrizean laburbiltzen dira. Ikus **3.1 Taula**.

	ald.1	***ald.2***
ald.1	TP	FN
ald.2	FP	TN

3.1 Taula: *Konfusio-matrizea.*

non *ald.1* eta *ald.2* intereseko klasearen benetazko balio posibleak diren eta ***ald.1*** eta ***ald.2***, iragarritako klasearen balioak.

TP eta TN balioek sailkatzailea egoki jokatzen ari dela esaten digute, eta FP eta FN balioek, berriz, kasuak gaizki sailkatu dituela. Taulan sailkapen bitarra azaldu da, baina klase-aldagai anitzen kasurako ere konfusio-matrizea irudikatu daiteke.

Orain, aurretik aipatutako ebaluazio-neurriei egingo zaie erreferentzia. Hasteko, testerako multzo baten gainean sailkatzaile baten asmatze-tasa zera da: datu-multzo horretatik egoki sailkatutako adibideen ehunekoa, eta honela kalkulatzen da:

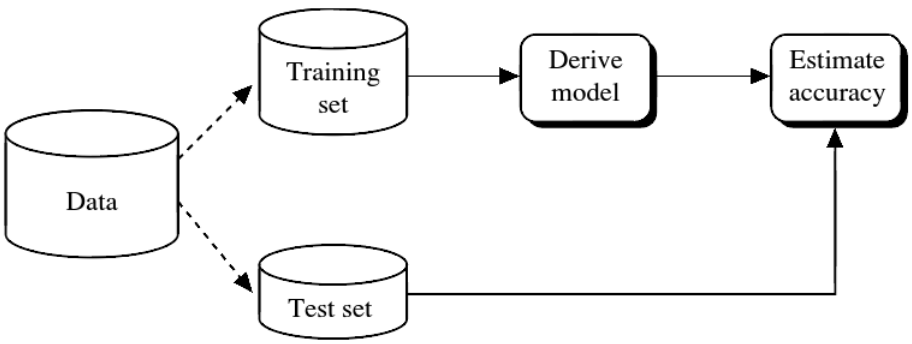
$$asmatze\text{-}tasa = \frac{TP+TN}{TP+FP+TN+FN}$$

Era berean, errore-tasa ere kalkulatu daiteke, hots oker sailkatutako adibideen ehunekoa. Era honetan:

$$errore\text{-}tasa = \frac{FP+FN}{TP+FP+TN+FN}$$

Holdout

Metodo honetan, datu-base osoa ausaz bi zati independenteetan banatzen da, entrenamendurako multzoa eta testerako multzoa. Ohikoa da datu-base osoko bi heren entrenamendurako erabiltzea, eta gainerako herena testerako. Entrenamendurako multzoaren laguntzaz, sailkatzailea eraikitzen da. Sailkatzailearen asmatze-tasa testerako multzoarekin determinatzen da. Lortzen den estimazioa ezkorra da, soilik hasierako datuen zati bat erabiltzen baita sailkatzailea lortzeko. Ondoren, *holdout* metodoaren prozesua laburbiltzen du **3.4 Irudiak**.



3.4 Irudia: *Holdout* metodoarekin, asmatze-tasaren kalkulua. [Jiawei Han, 2011, Chapter 8.5].

Data-k hasierako datu-baseari egiten dio erreferentzia, *Training set*-ak entrenamendurako datu-multzoari eta *Test set*-ak testerako datu-multzoari. *Derive model*-ek sailkatzailea eraikitzea esan nahi du, eta azkenik, *Estimate accuracy*-k asmatze-tasaren kalkuluari egiten dio erreferentzia.

k-geruzako balidazio-gurutzatua

k-geruzako balidazio gurutzatuak (*k-fold cross-validation*) datu-base osoa nahi adina partiziotan banatzeko aukera ematen du, tamaina berdineko multzoetan (D_1, D_2, ..., D_k). Entrenamenduak eta testak k aldiz egiten dira. i. iterazioan, D_i testerako multzo gisa gordetzen da, eta gainerako partizioak sailkatzailea entrenatzeko erabiltzen dira. *Hold-dout* metodoak ez bezala, *k*-geruzako balidazio-gurutzatuak azpimultzo bakoitza entrenamendurako aldi kopuru bera eta testerako behin erabiltzen ditu. Sailkapenari dagokionez, asmatze-tasa honela estimatzen da: k iterazioetan egindako sailkapen egokiak, hasierako datu-baseko instantzia kopuruarekin zatituta. Orokorrean, asmatze-tasa estimatzeko 10-geruzako balidazio-gurutzatua gomendatzen da; **3.5 Irudia**, horren adibide bat da. *Total number of dataset*, D_1, D_2, ..., D_k datu-multzoak dira eta *Number of experiments* datu-multzo horiekin egindako entrenamenduak.

3.3.5 Gainbegiratu gabeko ikasketa

Aurretik aipatu bezala, gainbegiratutako ikasketa gainbegiratu gabeko ikasketarekin desberdintasun batengatik bereizten da. Gainbegiratutako ikasketan, datu-multzoko adibideen menpeko aldagai edo klasea ezaguna da, gainbegiratu gabeko ikasketan, aldiz, ez dago instantzien menpeko klaserik.

Clustering algoritmoa

Clustering edo *cluster* analisia, datu-multzo bat azpimultzotan zatitzean datzan prozesu bat da. Azpimultzo horien banaketarako ezinbestekoa da adibideen arteko bereizketa egitea. Bereizketa hori, adibideen arteko distantzia aintzat hartuz lortzen da. Azpimultzo bakoitza *cluster* bat da. Cluster batean biltzen diren adibideek azpimultzo bereko gainerakoekin antzekotasunen bat dutela esan nahi du, hau da, adibideen arteko distantzia txikia da. Beste clusterretako instantziekiko, berriz, distantzia handiagoa dago, hau da, horiekin antzekotasun txikiagoa dute. Datu-multzoa zenbait azpimultzotan banatzeko, zenbait

3.5 Irudia: *k-geruzako balidazio-gurutzatua*-ren prozesua [Talpur, 2017], *3.9 Irudia*

metodo daude, eta horietako bakoitzak datu-multzo osoa modu desberdinetan banatu ahal dezake [Jiawei Han, 2011, Chapter 10]. Zenbat eta gertutasun handiagoa izan cluster bateko adibideek haien artean, eta zenbat eta gertutasun txikiagoa gainerako clusterreko adibideekin, orduan eta banaketa hobea da [Sharma et al., 2012].

Cluster analisia gauzatzeko algoritmo asko daude, eta metodo ezagunenen artean, **partiziozko clusteringa** dago. n adibideko datu-multzo bat izanik, partiziozko metodo batek k banaketa egiten ditu, bakoitzak cluster bat errepresentatuz ($k \leq n$). Cluster bakoitzak gutxienez instantzia bat izan behar du. Partiziozko clusteringaren teknika, cluster analisia egikaritzeko modurik errazena eta oinarrizkoena da, eta algoritmo ezagunenen artean, *k-Means* [MacQueen et al., 1967] metodoa dago, zentroideetan oinarritutako algoritmoa.

k-Means algoritmoa

k-Means algoritmoak cluster bereko adibideen arteko distantzia minimizatzea eta cluster arteko distantzia maximizatzea du helburu. Demagun D datu-multzoa n kasuz osatuta da-

goela. Partiziozko cluster analisiak D-ko adibide guztiak k cluster desberdinetan banatzen ditu, C_1, C_2, ..., C_k, non $C_i \subset D$ eta $C_i \cap C_j = 0$ ($1 \leq i, j \leq k$). k parametroak jaso dezakeen balio minimoa 2 da, datu-multzoa gutxienez bitan banatu ahal izateko, eta maximoa, berriz, $\sqrt{n}$, non n datu-multzoaren tamaina den. Algoritmo honek, clusterretako zentroideak erabiltzen ditu, C_i, cluster desberdinak errepresentatzeko, zentroideak azpimultzoen erdigunea izanik.

Algoritmo honen prozesua ondorengoa da. Lehenik eta behin, D datu-multzotik, ausaz k kasu hautatzen dira, non horietako bakoitzak cluster desberdinetako erdigune edo zentroidea adierazten duen. Gainerako kasuekin zera egiten da: kasu bakoitzari, zentroide antzekoena duen clusterra esleitzen zaio, horien arteko distantzia kontuan izanik. Beste era batera esanda, kasu bakoitzaren eta aurretik hautatutako clusterren zentroideen arteko distantzia kalkulatzen da, aurrerago azalduko den moduan, eta kasu bakoitzari, harengandik gertuen dagoen clusterra esleitzen zaio. *k-Means* algoritmoak iteratiboki zentroideen eta cluster bereko kasuen antzekotasuna hobetzen du. Iterazio bakoitzean eta cluster bakoitzerako, zentroide berria clusterreko instantzia guztiak erabiliz kalkulatzen da, aurrerago azalduko den moduan. Azpimultzo bakoitzeko adibideei berriro esleitzen zaie cluster bat, eguneratutako zentroide berria kontuan izanik. Iterazioen amaiera, clusterren esleipena behin egonkorra denean iristen da, aurreko iterazioan sortutako clusterren esleipenak mantentzen direnean, alegia [Jiawei Han, 2011, Chapter 10.2].

Aurretik esan bezala, clusteringeko algoritmoek datu-multzoko kasuen antzekotasunaren arabera sortzen dituzte azpimultzo desberdinak. Horretarako, adibideen arteko distantzia kalkulatzea beharrrezkoa da, adibideak bata bestearekin konparatu ahal izateko, hain zuzen ere. Kalkulu matematiko horiek datu-baseko atributuek jaso dituzten balioen artean egiten dira, atributuz atributu. Aldagaiak zenbakizkoak direnean erraza da distantzia estimatzea, baina gehienetan, datu-baseak zenbakizko aldagaiez eta zenbakizkoak ez direnez osatuta daude. Horregatik, teknika desberdinak existitzen dira datu-multzo mota horiekin clustering algoritmoak aplikatu ahal izateko.

Zenbakizkoak ez diren aldagaiak kudeatzeko teknika ezagun bat, *Gower* [Gower, 1971] distantziaren kontzeptua aplikatzea da. Kontzeptu hori sinplea da: zenbakizkoa ez den aldagai mota bakoitzeko, distantzia desberdin bat kalkulatzen da, zeinak 0 eta 1 arteko balio bat jasoko duen. Aldagai nominalen kasuan, q balio posible dituen atributu bakoitzetik, q zutabe bitar sortzen dira. Atributuak adibide zehatz horretan hartutako balioaren zutabean 1 egongo da, eta atributu horren gainerako balio posibleetan 0.

Ondoren, instantzien arteko distantzia kalkulatzea lan errazagoa izango litzateke. Izan ere,

balio guztiak izango lirateke zenbakizkoak. Distantzia euklidearra eta Manhattan distantziak, adibidez, ohikoak dira cluster analisian, beste hainbeste ere erabili ahal diren arren [Singh et al., 2013].

Demagun $X = x_1, x_2, ..., x_n$ datu-multzoko elementuak direla, eta $V = v_1, v_2, ..., v_c$ clusterren zentroideak.

- Datu-multzo osoko adibide bakoitza eta zentroide bakoitzaren arteko **distantzia euklidearra** honela kalkulatzen da:

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

- Adibide bakoitza eta zentroide bakoitzaren arteko **Manhattan distantzia**, berriz, honela kalkulatzen da:

$$Dist_{XY} = |x_{ik} - x_{jk}|$$

Behin distantzia kalkulatuta, aurretik esan bezala, adibide bakoitzari cluster bat esleitzen zaio, non horren zentroidearekiko distantzia minimoa duen. Ondoren, clusterren zentroide berria kalkulatzen da, modu honetan:

$$V_i = \left(\frac{1}{C_i}\right) \sum_{1}^{c_l} x_i$$

3.3.6 Cluster balidazioa

Aurretik aipatu bezala, clustering prozesuaren helburua, datu-multzo orokorretik antzekoak diren kasuez osatutako clusterrak sortzea da. Algoritmo gehienek k parametroa dute, cluster kopurua definitzeko. Cluster kopuru desberdina ezarriz, partizio desberdinak lortzen dira, baina zein da partiziorik onena? Clustering algoritmo optimo bat ez da existitzen. Beste era batera esanda, partizio bat bera ere ez da onena kasu guztietan. Gainbegiratutako ikasketan, jatorrizko datu-baseko kasu batzuk gorde ohi dira eraiki den sailkatzailearen eraginkortasuna neurtzeko. Gainbegiratu gabeko ikasketan, hau da, kasu honetan, ordea, hori ez da posible. Beraz, clustering eraginkor bat egikaritzeko asmoz, datu-multzo

berdinaren gainean partizio desberdinak gauzatzea gomendagarria da. *k-Means* algoritmoaren kasuan, ohikoa da k (cluster kopurua) parametroaren zenbait balio desberdinekin probak egitea.

Cluster balidazioko indizeak

Cluster balidaziorako indizeak edo *Cluster Validation Indices* (CVI), clustering partizioak ebaluatzeko erabiltzen diren neurriak dira. Indize horiek partizio onena aukeratzeko erabiltzen dira [Arbelaitz et al., 2013]. Ohikoa da indize guztiek ez partizio bera proposatzea onena gisa. Indize asko daude eta ez dago argi zein den onena. Hala ere, jarraian horietako batzuk azalduko dira.

Notazioa

Demagun X datu-multzo bat dela, bektore gisa errepresentatutako N kasuz osatutakoa: $X = x_1, x_2, ..., x_n$. X gaineko K partizioak honela adieraziko dira: $C = c_1, c_2, ..., c_K$, non $\bigcup_{k \in c} c_k = X$, $c_k \cap c_l = 0 \; \forall \; k \neq l$. c_k cluster baten zentroidea. Haren batez-besteko bektorea da, $\overline{c_k} = 1/|c_k| \sum_{x_i \in c_k} x_i$ eta antzeko modura, datu-multzo osoko zentroidea, datu-multzo osoko batezbesteko bektorea da, $\overline{X} = 1/|N| \sum_{x_i \in X} x_i$.

Bi adibideen, x_i eta x_j, arteko distantzia euklidearra $d_e(x_i, x_j)$ gisa errepresentatuko da. Indize bakoitzean deskribatzen diren bi neurriak hauek dira: kohesioa (Δ) eta banaketa (δ).

- **Silhouette indizea (Sil)** [Rousseeuw, 1987]

  Indize hau bi neurrietan oinarritzen da: alde batetik, kohesioan, hau da, cluster bereko kasu guztien arteko distantzietan eta bestetik, banaketan, cluster bakoitzetik gertuen dagoen beste clusterrerainoko distantzietan. Jasotzen duen balioa 0 eta 1 arteko neurri bat da, eta zenbat eta indize altuagoa izan, orduan eta partizio hobea dela adierazten du.

  $$Sil(C) = 1/N \sum_{c_k \in c} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}}$$

  ,

  non

  $$a(x_i, c_k) = 1/|c_k| \sum_{x_j \in c_k} d_e(x_i, x_j)$$

$$b(x_i,c_k) = min_{c_l \in C c_k} \left\{ 1/|c_l| \sum_{x_j \in c_l} d_e(x_i,x_j) \right\}$$

- **Davies-Bouldin-en indizea (DB)** [Davies and Bouldin, 1979]

  Neurri hau CVI ohikoenetako bat dela esan daiteke. Kohesioa cluster bereko kasuek bertako zentroidearekiko duten distantzietan oinarritzen da eta banaketa, cluster desberdinen zentroideen arteko distantzietan. Zenbat eta indize baxuagoa, orduan eta partizio hobea dela adierazten du.

$$DB(C) = \frac{1}{K} \sum_{c_k \in C} max_{c_l \in C c_k} \left\{ \frac{S(c_k) + S(c_l)}{d_e(\overline{c_k}, \overline{c_l})} \right\}$$

  ,

  non

$$S(c_k) = 1/|c_k| \sum_{x_i \in c_k} d_e(x_i, \overline{c_k})$$

- **Calinski-Harabasz-en indizea (CH)** [Caliński and Harabasz, 1974]

  Indize honen kasuan, kohesioa cluster bereko kasuen eta cluster horretako zentroidearen arteko distantziekin kalkulatzen da, eta banaketa, berriz, clusterretako zentroideetatik, existitzen den beste zentroide orokor baterainoko distantziekin. Zenbat eta indize altuagoa, orduan eta partizio hobea dela esan nahi du.

$$CH(C) = \frac{N-K}{K-1} \frac{\sum_{c_k \in C} |c_k| d_e(\overline{c_k}, \overline{X})}{\sum_{c_k \in C} \sum_{x_i \in c_k} d_e(x_i, \overline{c_k})}\}$$

- **Dunn-en indizeak** [Dunn, 1973]

  Kasu honetan, kohesioa cluster batetik gertuen dagoen beste clusterrerako distantzian oinarrituta estimatzen da, eta banaketa cluster diametroaren distantzia maximoarekin. Zenbat eta indize altuagoa, orduan eta partizio hobea dela esan daiteke. Indize mota honetan oinarrituta, beste hainbeste neurri deribatzen dira. Hala nola, **Orokortutako Dunn-en Indizeak** (*Generalized Dunn Indices*), honako izenez ezagunak direnak: **gD31**, **gD41**, **gD51**, **gD33**, **gD43**, **gD53**. Neurri horiek guztiak, δ (banaketarako estimazioa) parametroaren hiru estimazioren eta Δ (kohesiorako estimazioa) parametroaren bi estimazioren konbinaketan oinarritzen dira.

$$\delta^3(c_k,c_l) = \frac{1}{|c_k||c_l|}\sum_{x_i\in c_k}\sum_{x_j\in c_l} d_e(x_i,x_j)$$

,

$$\delta^4(c_k,c_l) = d_e(\overline{c_k},\overline{c_l})$$

,

$$\delta^5(c_k,c_l) = \frac{1}{|c_k|+|c_l|}\left(\sum_{x_i\in c_k} d_e(x_i,\overline{c_k}) + \sum_{x_j\in c_l} d_e(x_j,\overline{c_l})\right)$$

eta

$$\Delta^1(c_k) = \Delta(c_k)$$

,

$$\Delta^3(c_k) = 2/|c_k|\sum_{x_i\in c_k} d_e(x_i,\overline{c_k})$$

Esate baterako, **gD33** indizean, cluster bereko kasuen arteko distantzia gisa (kohesioa), bi aldiz zentroidetik kasu guztietarako batez-besteko distantzia kontsideratzen da. Eta banaketa gisa, hau da, cluster arteko distantzia gisa, batez-besteko clusterren arteko loturen distantzia. **gD43** indizean, kohesio gisa aurreko kasuko distantzia bera kontsideratzen da. Banaketa gisa, ordea, clusterren zentroideen arteko distantzia. Eta, **gD53** indizean, kohesioa aurreko bi kasuetan kontsideratu den distantzia bera izan da. Eta banaketari dagokionez, clusterren zentroide bakoitzetik gainerako clusterretako kasu guztietarako distantzien batezbestekoa.

- **COP indizea (COP)** [Gurrutxaga et al., 2010]

  Indize honetan honakoa hartzen da kontuan: kohesioa estimatzeko, cluster bereko kasuetatik horko zentroidera duten distantziak hartzen dira kontuan, eta banaketarako, cluster batetik urrutien dagoen clusterrerainoko distantzia. Kasu honetan, zenbat eta balio txikiagoa, orduan eta partizio hobea dela adierazten du.

$$COP(C) = \frac{1}{N}\sum_{c_k\in C}|c_k|\frac{1/|c_k|\sum_{x_i\in c_k} d_e(x_i,\overline{c_k})}{min_{x_i\notin c_k} max_{x_j\in c_k} d_e(x_i,x_j)}$$

4. KAPITULUA

Proiektuaren garapena

Proiektu hau aurrera eramatearen helburua zenbait galdera erantzutea izan da, **Proiektuaren Helburuen Dokumentua** atalean aipatu bezala. Inplementatu den datu-meatzaritzako prozesua hiru atal nagusitan banatzen da; alde batetik, datuen bilketa, hau da, galdetegien bidez emaitzak jasotzea. Bestetik, datuen aurreprozesaketa, non eskuarteko informazioa prozesatzeko prest utzi den eta azkenik, datuen analisia, non datuak interpretatu diren. Kasu honetan, lehenengo urratsa aurretik egina zegoen, hau da, ekimena aurrera eraman den urte bakoitzeko datu-bilketa gauzatu da. Hala ere, ondoren prozesuko hiru atal horiek deskribatuko dira.

4.1 Datu-bilketa

Ikerketa hau egiteko erabili diren datuak, 10-12 urte bitarteko hainbat haurri eskaini zaien kode orduan jasotako inkestak dira. Galdeketa horiek UPV/EHUko Donostiako Informatika Fakultateak antolatutakoak dira. 2015/2016, 2016/2017 eta 2017/2018 ikasturteetan, zortzi, zazpi eta hamabi ikastetxetan egin da ekimena. Ikasleei etxean betetzeko inkesta bat eman zitzaien. Inkesta hori, ikastetxe batzuetan bete dute baina ez guztietan. Hain zuzen ere, 2015/2016 ikasturtean 5 ikastetxe izan ziren inkestak bueltatu zizkigutenak (267 inkesta), 2016/2017 ikasturtean hiru (207 inkesta), eta 2017/2018 ikasturtean berriz, 7 (393 inkesta). Ikastetxe gehienak Gipuzkoan kokatuta daude, nahiz eta badiren Bizkaia eta Nafarroakoak, eta bakarra da hiru urteetan zehar inkestak itzuli dituen ikastetxea. Azken hori, kasu partikular gisa aztertuko da.

Prestatu den galdera-sortaren helburua haurrek informatikari buruz duten ezagutza eta ikuspuntua zein den eta ikasle bakoitzaren inguruak horretan duen eragina aztertzea da. Horren gaztetxoak diren ikasleekin askotan kosta egiten da asmatzea nola jaso haien iritzi eta ikuspuntuak. Horretarako sortu ziren hain zuzen ere *"Draw A Scientist Test"* edo DAST deritzotenak [Chambers, 1983] eta Ingeniarientzako ere erabiliak izan direnak [Knight and Cunningham, 2004]. Hori kontuan izanik, lan honetarako, beste galdera

batzuen artean ikasleei eskatu zaie batetik Informatikaria eta bestetik Informatikariaren lantokia marraztea. Irudi horien bidez saiatuko gara aztertzen ikasleek horren adin gaztean dagoeneko informatikariei buruzko genero eta *freaky* estereotipoak barneratuta ote dituzten. Irudiaren generoa, betaurrekoak ote dituzten eta informatikariak lanean, bakarka ala taldean imajinatzen ote dituzten aztertuko dugu horretarako.

Zehazki, honako datuak eskatu zitzaizkien ikasleei:

- Ikasturtea
- Eskola
- Ikaslearen generoa
- Ikaslearen adina
- Guraso bakoitzaren generoa
- Guraso bakoitzaren ikasketa-maila
- Guraso bakoitzaren lanbidea
- Irudi bidez adierazitako ogibide multzo bat, haietan informatikari batek lan egin ote lezakeen galdetzeko (medikuntza, artea, kimika, autogintza, jantzigintza, etxegintza, bideo-jokoak, musikagintza, zuzenbidea, hizkuntzalaritza, ordenagailu-sareak, hezkuntza)
- Irudikatu informatikari bat:
  - Irudikatutako pertsonaren generoa
  - Irudikatutako pertsonak betaurrekoak zeramatzan edo ez
- Irudikatu informatikarien lantokia:
  - Irudikatutako lanlekuan zeuden ordenagailu kopurua, informatikaria bakarrik edo taldean imajinatzen duen

4.1.1 Galdetegiek jasandako aldaketak

Arestian aipatu bezala, hiru ikasturte desberdinetan eraman da aurrera ekimen hau; 2015/16, 2016/17 eta 2017/18 ikasturteetan, hain zuzen. Urtetik urtera, aurreko esperientziatik ikasiz eta galdetegietan zenbait birmoldaketa egin ondoren, horien eraginkortasuna hobetu

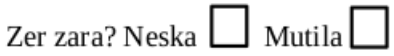
da. Beraz, esan daiteke jaso nahi izan zen informazioa ez zela era berean bildu. Ikus **A-Eranskina** eta **B-Eranskina**.

Hona hemen zenbait aldaketen eraginez desberdin jaso ziren informazio zatiak:

Generoa

Generoaren inguruko aldaketa hau galdetegia inklusiboagoa izateko egin da.

- **2015/16** ikasturtean egindako galdera sortan, ikaslea *neska* edo *mutila* den erantzuna jaso zen. Ikus **4.1 Irudia**.



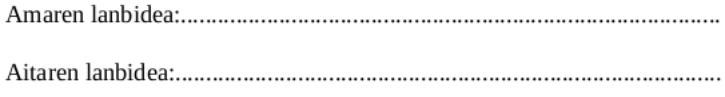
4.1 Irudia: *2015/16 ikasturte*ko galdetegian ikaslearen generoari buruzko galdera.

- **2016/17 eta 2017/18**-ko galdeketan, aldiz, hirugarren aukera bat gehitu zen, eta beraz, hauek izan ziren aukerak: *neska*, *mutila* edo *ez bitarra*. Ikus **4.2 Irudia**.

4.2 Irudia: *2016/17 eta 2017/18 ikasturteta*ko galdetegian ikaslearen generoari buruzko galdera.

Gurasoen ikasketa-maila eta lanbidea

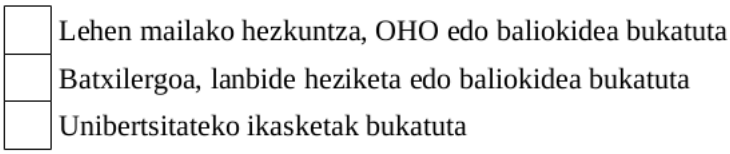
- **2015/16**: Zuzenean amaren eta aitaren lanbideak bereizi ziren, gizonezkoa eta emakumezkoa zirela suposatu. Ikus **4.3 Irudia**.



4.3 Irudia: *2015/16 ikasturte*ko galdetegian gurasoen lanbideari buruzko galdera.

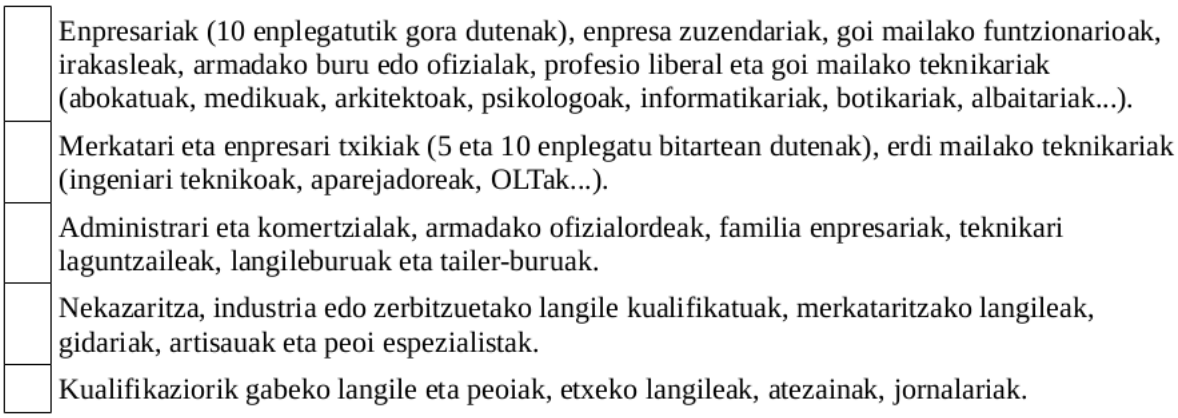
Atal honetan, ikasleak eman zezakeen erantzuna nahiko orokorra izan zitekeenez, nahiko erantzun anbiguoak jaso ziren. Gainera, lanbidea bakarrik eskatu zenez, guraso bakoitzaren ikasketa-maila ezezaguna zen.

- **2016/17 eta 2017/18**: Gainerako galdeketetan, ordea, informazio hori zehaztasun handiagoarekin eskatu zen. Alde batetik, gurasoaren generoa zein den jaso zen, *emakumezkoa*, *gizonezkoa* edo *ez-bitarra* izanik aukera posibleak. Hori ere, galdetegia inklusiboagoa izan zedin egin zen. Bestetik, gurasoaren ikasketa-maila zehazterakoan hiru talde bereizi ziren. Ikus **4.4 Irudia**.



4.4 Irudia: *2016/17 eta 2017/18 ikasturteta*ko galdetegian gurasoaren ikasketa-mailari buruzko galdera.

Lanbideari dagokionez, bost talde bereizi ziren. Ikus **4.5 Irudia**.



4.5 Irudia: *2016/17 eta 2017/18 ikasturteta*ko galdetegian gurasoaren lanbideari buruzko galdera.

Horietatik, guraso bakoitzarekin bat zetorren hautagaia aukeratu zuten.

Era horretan, 2015/16 ikasturteko galdeketan sortutako arazoa ekidin ahal izan zen, erantzunaren anbiguotasuna, hain zuzen. Izan ere, ikasketak eta lanbidea maila desberdinetan taldekatu ziren.

4.1.2 Digitalizazio prozesua

Galdera desberdinak egin izanak informazio desberdina jasotzea eragin zuen. Galderasorta hori paper bidezko galdetegietan bete zuten ikasleek eta nolabait digitalizatu egin

behar izan zen. Ikasturte bakoitzeko emaitzak era desberdinetan pasa ziren modu digitalera, hori ere aurreko urteko esperientziatik ikasita. Digitalizazio prozesu horren aurretik ez zen behar besteko haurnasketa egin, eta ondorioz, informazioa era ezberdinetan gorde zen. Hala ere, azken batean, hiru urtetako datuak taula bidez errepresentatu daitezkeen egituran batu ziren. Lehenengo urtean egindako galdera sortako bost ikastetxetako erantzunak, eskuz pasa ziren modu digitalera, datu-fitxategi bakar batera, eta bigarren eta hirugarren urteko bederatzi ikastetxetakoak, aldiz, txantiloi bidez, fitxategi bakar batera ere. Hortaz, bi datu-base osatu zirela ondoriozta daiteke, bakoitzeko atributuen balioak berdinak zertan izan gabe.

2015/16 ikasturteko galdeketen digitalizazioa

Emaitza hauek eskuz eta banan-banan pasa behar izan zirenez ordenagailura, modu sinpleena eta laburtuena hautatu zen. Ikus **4.1 Taula**.

Eskatutako datuak	**Digitalizazioa**
Eskola	Ikastetxe bakoitzarentzat zenbaki bat
Ikaslearen generoa	0 = Zehaztu ezin daitekeen generoa 1 = Emakumezkoa 2 = Gizonezkoa
Irudikatu informatikari bat (generoa)	0 = Zehaztu ezin daitekeen generoa 1 = Emakumezkoa 2 = Gizonezkoa
Irudikatu informatikari bat (betaurrekoak)	0 = betaurrekoak ez 1 = betaurrekoak bai
Irudikatu informatikarien lantokia (lanlekuko ordenagailu kopurua)	1 = ordenagailu bat 2 = ordenagailu asko
Irudi bidez adierazitako ogibideetan informatikarien parte-hartzea	0 = ez 1 = bai

4.1 Taula: *2015/16 ikasturteko* galdeketen digitalizazioa.

Ordezkapenik jasan ez zuten datuak honakoak izan ziren: ikasturtea, ikaslearen adina eta guraso bakoitzaren lanbideak, ikaslearen erantzun berak idatzi baitziren.

2016/17 eta 2017/18 ikasturteetako galdeketen digitalizazioa

Emaitza hauek, ordea, txantiloi bidez pasa ziren modu digitalera, hau da, galdeketen emaitzak paperean izanik, *Google*-ko formulario bat sortu zen, galdera berdinak zituena, eta ikasle bakoitzaren emaitzak sartu ziren banan-banan. Ondoren, Googlek berak taula bat sortu zuen automatikoki, bildutako informazioarekin.

Kasu honetan, informazioa erantzunaren arabera zuzenean jaso zen, hau da, galdera bakoitzaren emaitza posibleak inongo zenbakirekin ordezkatu gabe. Ikus **4.2 Taula**.

Eskatutako datuak	**Digitalizazioa**
Eskola	Ikastetxe bakoitzaren izena bera
Ikaslearen generoa	"Ez bitarra" "Emakumezkoa" "Gizonezkoa"
Guraso bakoitzaren generoa	"Ez bitarra" "Emakumezkoa" "Gizonezkoa"
Guraso bakoitzaren ikasketa-maila(*)	Talde bakoitzarentzat zenbaki bat
Guraso bakoitzaren lanbidea(*)	Talde bakoitzarentzat zenbaki bat
Irudikatu informatikari bat (generoa)	"Ez bitarra" "Emakumezkoa" "Gizonezkoa"
Irudikatu informatikari bat (betaurrekoak)	"Ez" = betaurrekoak ez "Bai" = betaurrekoak bai
Irudikatu informatikarien lantokia (lanlekuko ordenagailu kopurua)	"Bat" = ordenagailu bat "Anitz" = ordenagailu asko

4.2 Taula: *2016/17 eta 2017/2018 ikasturteetako* galdeketen digitalizazioa.

(*) Gurasoen ikasketa-mailari eta lanbideari dagokienez (**4.2 Taula**), bereizitako hiru eta bost taldetako hautagai bakoitzari zenbaki bat egokitu zitzaion. Dena den, 2017ko ikasturteko gurasoen lanbideak, ez daude digitalizaturik.

Ikasleen ustean informatikariak behar diren ogibideetarako, zerrenda bat sortu zen ikasleak informatikariak behar direla erantzundako lanbideekin, zutabe bakar batean: ogibide batean informatikariak behar direla zehaztu bazen, zerrendan azalduko litzateke ogibidea, eta bestela, ez.

4.2 Datuen aurreprozesaketa

Esan daiteke datu-bilketaren ondorio gisa bi datu-fitxategi eskuratu zirela. Hortaz, lanmateriala honakoa izan da: alde batetik, 2015/16 ikasturteko datuak gordetzen dituen datu-basea, eta bestetik, 2016/17 eta 2017/18 ikasturteetako datuena, haien arteko desberdintasunak aurreko atalean aipatutakoak izanik. Beraz, behin datu-bilketa burutua izanik, eskuratutako datu-baseak bateratu egin behar izan dira, atributu bakoitzarentzat kodeketa

komun eta zuzen bat egokituz. Azken finean, datu-fitxategi bat sortu nahi izan da, non hiru urteetako datu guztiak biltzen diren. Horretaz gain, datuak urteka ere banandu dira eta hiru urteetan zehar parte hartu duen ikastetxe bateko datuak biltzen dituen datu-fitxategia ere eratu da, Lazkaoko San Benito Ikastolako datu-basea, hain zuzen ere.

4.2.1 Hartutako erabakiak

Egiturari dagokionez, datu-baseak izan beharreko atributuak edo zutabeak zein diren erabaki behar izan da. Nahiz eta bi datu-fitxategietako zutabe gehienak berdinak izan, erantzun batzuen digitalizazioan zenbait desberdintasun nabarmendu dira; hala nola, proposatutako ogibideetan informatikari baten parte-hartzeari buruzko galderan. Txukuntasuna mantentzearren, ogibide bakoitzari zutabe bat egokitzea erabaki da, 2015/16 ikasturtean erabili zen egitura mantenduz.

Bestalde, erabaki garrantzitsuagoak ere hartu behar izan dira. Bi datu-fitxategien arteko desberdintasun nagusia gurasoen lanbidean eta ikasketa-mailan egon da. Lehenengo urtean jasotako gurasoen lanbideari buruzko erantzunak, gainerako urteetan bereizitako lanbide eta ikasketa-mailen taldeetan sailkatzea erabaki da. Horretarako, zenbait suposizioez baliatu behar izan gara. Esate baterako, ikaslearen erantzuna “Medikua” izan baldin bada, suposatu da, alde batetik, guraso hori “Unibertsitateko ikasketak bukatuta” taldekoa dela, eta bestetik, talde hau ere dagokiola: “Enpresariak (10 enplegatutik gora dutenak), enpresa zuzendariak, goi mailako funtzionarioak, irakasleak, armadako buru edo ofizialak, profesio liberal eta goi mailako teknikariak(abokatuak, medikuak, arkitektoak, psikologoak, informatikariak, botikariak, albaitariak...)”. Erantzunak banan-banan sailkatu dira, eta ondorioz, datuak txukuntzerakoan lan gehigarria izan da hori. Dena den, funtsean ikasketa-maila bakarrik hartu da kontuan. Izan ere, 2016/2017 ikasturtean ez ziren lanbideak jaso eta urte osoko atributu horren hutsuneak sinesgarritasunean eragina izango luke. Bestalde, gurasoen ikasketa-maila lanbidea bera baino garrantzitsuagoa dela uste dugu.

4.2.2 Eragindako aldaketak

Aipatutako erabakiak ardatz izanik, jarraian zehaztuko diren egokitzapenak egin dira datu-fitxategietan. Aldaketak egiteko, *Python* programazio lengoaian inplementatutako programa txikiak garatu dira. Dena den, **5. Erabiltako tresnak** atalean sakonduko da horren inguruan.

2015/16 ikasturteko fitxategia

Lehenik eta behin, beste datu-fitxategien egitura mantenduz, zenbakizko datuak, datu nominal bihurtu dira. Adibidez, "Marrazkiko informatikariaren generoa" atributua 0, 1 edo 2 balioekin adierazi bada, 0koa zutenei "Ezinzehaztu" balioa egokitu zaie, 1ekoa zutenei "Emakumezkoa" eta 2koa zutenei "Gizonezkoa"; berdina egin da "Ikaslearen Generoa" atributuarekin. Ondoren, gurasoen lanbideari buruzko emaitza anbiguotik haien ikasketamaila eta lanbidea taldeetan sailkatuta izanik, talde desberdinei dagozkien zenbakiak egokitu zaizkie. Labur esanda, zenbakiz adierazitako atributuak, adina izan ezik, dagozkien izenarekin ordezkatu dira.

2016/17 eta 2017/18 ikasturteetako fitxategiak

Datu-multzo honetan egindako aldaketa nagusia ogibideekin erlazionatutako galderen emaitzen egituran egin da: informatikariak behar diren ogibideen zerrenda desegin da, eta ogibide bakoitzari zutabe bat egokitu zaio — ogibidea zerrendan azaldu bada, zutabe horren balioa 1ekoa izan da, eta bestela, 0koa —. Funtsean, 2015/16 ikasturteko taularen estruktura erabili nahi izan da. Bi ikasturte horietako galderen artean gurasoen generoari buruz ere galdetu zenez, datu-base honetan zutabe bat gehitu da. Atal horrek ikaslearen gurasoak "Emakumezkoa-Emakumezkoa", "Gizonezkoa-Gizonezkoa", "Emakumezkoa-Gizonezkoa" edota "Guraso bakarrak" diren zehaztu du. Dena den, etorkizuneko prozesamenduan atributu hori ez da gogoan izan, 2015/16 ikasturteko gurasoen generoei buruzko informazioa ez baitzen guztiz ziurra. Izan ere, galdetegi horretan zehaztuta zegoen "Amaren lanbidea" eta "Aitaren lanbidea" idazteko, **4.3 Irudian** ikus daitekeen bezala.

4.2.3 Balio-hutsak (*missing values*)

Azpimarratzekoa da datu-fitxategiak ez daudela guztiz osatuta, hau da, kasu batzuetan datuak falta dira (*missing values*). Datu-base osoari erreparatuz, 866 ikaslek erantzun zituzten inkestak, baina horietatik 636 dira datu bat bera ere falta ez zaizkienak. Beste era batera esanda, ikasleen %26.56-k galderaren bat erantzun gabe utzi zuen. **4.3 Taulan** ikus daitezke datu zehatzagoak.

2015/16 ikasturteko datuei erreparatuz, 267 ikaslek erantzun zituzten inkestak, baina horietatik 240 dira datu bat bera ere falta ez zaizkienak. Hots, ikasleen %10.11-k galderaren bat erantzun gabe utzi zuen. Ikus **4.4 Taula**.

Datuak	**Ikasle kopurua**	**Ehunekoa**
Datu guztiak	866	%100
Missing values baztertuta	636	%73.44
Missing values dituztenak	230	%26.56

4.3 Taula: *Missing values*-en eragina datu-base osoan.

Datuak	**Ikasle kopurua**	**Ehunekoa**
2015/16 datu guztiak	267	%100
Missing values baztertuta	240	%89.89
Missing values dituztenak	27	%10.11

4.4 Taula: *Missing values*-en eragina *2015/16 ikasturteko* datu-basean.

2016/17 ikasturteko datuei erreparatuz, 207 ikaslek erantzun zituzten inkestak, eta horietatik 109 dira balio-hutsak ez dituztenak. Hau da, ikasleen %47.34-k galderaren bat erantzun gabe utzi zuen. Ikus **4.5 Taula**.

Datuak	**Ikasle kopurua**	**Ehunekoa**
2016/17 datu guztiak	207	%100
Missing values baztertuta	109	%52.66
Missing values dituztenak	98	%47.34

4.5 Taula: *Missing values*-en eragina *2016/17 ikasturteko* datu-basean.

2017/18 ikasturteko datuei erreparatuz, 392 ikaslek erantzun zituzten inkestak, eta horietatik 287 dira balio-hutsak ez dituztenak; ikasleen %26.79-k galderaren bat erantzun gabe utzi zuen. Ikus **4.6 Taula**.

Azkenik, Lazkaoko San Benito Ikastolako datu-multzo partikularrean ere, beste bi azpi datu-base bereizi dira. **4.7 Taulari** erreparatuz, ikus daiteke zentro horretako 207 ikasleren emaitzak lortu zirela eta horietatik 162 direla datu bat bera ere falta ez zaizkienak. Hau da, ikasleen %21.74-k galderaren bati ez zion erantzunik eman.

Funtsean, balio-hutsak tratatzerakoan, bi aukera izan dira aintzat hartu direnak. Alde batetik, balio hutsak dituzten errenkadak ezabatzea, beti ere, kontuan hartuz datu baztertuen zenbatekoa oso handia ez izatea, eta bestetik, gehien jaso den erantzunarekin ordezkapena egitea. Ondorioz, bi datu-base bereizi dira; bata, hasieran jasotako informazio guztia duen datu-fitxategia (*missing values* ordezkatuta), eta bestea, *missing values* ezabatuta, gainerakoena.

Datuak	Ikasle kopurua	Ehunekoa
2017/18 datu guztiak	392	%100
Missing values baztertuta	287	%73.21
Missing values dituztenak	105	%26.79

4.6 Taula: *Missing values*-en eragina *2017/18 ikasturteko* datu-basean.

Datuak	Ikasle kopurua	Ehunekoa
Lazkaoko datuak	207	%100
Missing values baztertuta	162	%78.26
Missing values dituztenak	45	%21.74

4.7 Taula: *Missing values*-en eragina Lazkaoko datu-basean.

Lehen ikasketa-automatikoko esperimentu batzuk egin dira, balio-hutsak baztertuta sortutakoak, horiek ezabatutako datu-baseak bereiztuz. Bi datu-base horiekin probak egin ostean, antzeko asmatze-tasak lortu dira. Ondorioz, datuen fidagarritasunagatik, *missing values* ezabatutako datu-basea erabiltzea erabaki da. Hortaz, ikasketa-automatikoko teknikak aplikatzeko datu-baseak honela geratu dira: 636 errenkada, datu-base osoan, eta 162 errenkada, Lazkaoko datu-basean.

4.2.4 Datuen analisi-estatistikoa

Datu-fitxategietan egindako moldaketa guztien ostean, lortutako datu-baseetan oinarrituta azterketa-estatistiko orokor bat egin da. Analisi estatistikoa, aurreko atalean esan bezala, datuak falta ez zaizkien errenkadekin egin da, eta jarraian, azterketa horretatik ateratako informazio esanguratsuena adieraziko da.

Aurreko ataletan aipatu den bezala, badaude zenbait galdera erantzun nahi direnak; horietako bat honakoa litzateke: 10-12 urteko haurren generoa bereizi al daiteke haien inkestetako erantzunen arabera? Beste era batera esanda; adin horretako neskek eta mutilek informatikari buruz duten ezagutza eta iritzia desberdinak al dira? Horretarako, ikaslearen generoa estatistikoki aztertu da, datu-basea orekatuta dagoen jakin ahal izateko. **4.8 Taulan** ikus daitekeen bezala, neska eta mutilen kopurua nahiko orekatua izan da.

Erantzun nahi den hurrengo galdera, haurrek marraztu duten informatikariaren generoarekin erlazionatuta dago. Hau da, 10-12 urteko gaztetxoek informatikaria emakume irudikatzen duten jakin nahi da, eta horrez gain, ba ote dagoen ezaugarri komunen bat emaku-

Ikaslearen generoa	Ikasle guztiak	Ehunekoa	Lazkaokoak	Ehunekoa
Neska	328	%51.57	89	%54.94
Mutila	308	%48.43	73	%45.06
Ez bitarra	0	%0	0	%0

4.8 Taula: Parte hartu duten ikasleen generoa.

me marrazten dituztenen artean. Marrazteko eskatu zitzaien informatikariari erreparatuz, gehiengoak gizonezkoa irudikatu du. Izan ere, emakumezkoa marraztu dutenen ehunekoa nahiko baxua da gizonezkoa marraztu duten proportzioarekin alderatuta. **4.9 Taulan** ikus daitekeen bezala, datu-basean desoreka nabarmentzen da marrazkiaren generoari dagokionez.

Marrazkiaren generoa	Ikasle guztiak	Ehunekoa	Lazkaokoak	Ehunekoa
Emakumezkoa	186	%29.25	58	%35.80
Gizonezkoa	363	%57.07	90	%55.56
Ezin zehaztu	87	%13.68	14	%8.64

4.9 Taula: Ikasleak irudikatutako informatikariaren generoa.

Horiez gain, 10-12 urteko haurrei informatikariak duen *freaky* estereotipo hori iritsi zaien aztertzeko, bi datu hartu dira kontuan: alde batetik, informatikariaren marrazkian ea betaurrekoak irudikatu dituzten edo ez, eta bestetik, informatikaria lanlekuan bakarrik edo taldean irudikatzen duten. **4.10 Taulan** ikusten den bezala, betaurrekoei dagokienez, datubase osoan ikasleen erdiak baino gutxiagok irudikatzen dute informatikaria betaurrekoekin, baina Lazkaoko kasuan, erdiak baino zertxobait gehiagok. Bi kasuetan ikasleek emandako erantzunaren ehunekoa, gizartean betaurrekoak erabiltzen dituen jendearen ehunekoaren azpitik dago, eta beraz, ez du atentziorik ematen [Netherlands, 2013]. Informatikariaren lantokiari dagokionez, **4.11 Taulan** zehazten den bezala, bi datu-baseetan gehiengoak konputagailu bakarra marraztu du, hots gehiengoak informatikaria bakarrik irudikatzen du.

Betaurrekoak	Ikasle guztiak	Ehunekoa	Lazkaokoak	Ehunekoa
Bai	283	%44.50	88	%54.32
Ez	353	%55.50	74	%45.68

4.10 Taula: Ikasleak irudikatutako informatikariari betaurrekoak jantzi dizkion edo ez.

4.8, **4.9**, **4.10** eta **4.11 Taulei** erreparatuz, esan daiteke 10-12 urteko haurrek informatika-

Ordenagailuak lanlekuan	Ikasle guztiak	Ehunekoa	Lazkaokoak	Ehunekoa
Bat	490	%77.05	124	%76.54
Asko	146	%22.95	38	%23.46

4.11 Taula: Ikasleak irudikatutako informatikariaren lanlekuan ordenagailu bat edo asko dauden.

ria errazago irudikatzen dutela gizonezko gisa, emakumezkoa baino. Gainera, lanlekuan bakarrik irudikatzen dituzte informatikariak, taldean baino. Betaurrekoak, aldiz, ez dituzte informatikarien ezaugarri bereizgarri gisa hartzen.

4.3 Gainbegiratutako ikasketa

Azterketaren helburua bikoitza da, proposatu diren galderentzat asmatze-tasa altuak dituzten sailkatzaileak lortzen saiatzea alde batetik, eta, hori posible bada, aukera bakoitzarentzat azalpen bat topatzea. Horretarako, azalpena ematen duten sailkatzaileak erabili dira, sailkapen zuhaitzak eta erregela-multzoak, hain zuzen ere. Hala ere, beste sailkatzaile batzuk ere erabili dira asmatze-tasa hobetzen ote duten aztertzeko. Esperimentuak [Witten et al., 2011] liburuan aurkezten den *Weka* software ingurunean egikaritu dira, eta aurrerago hitz egingo da horretaz.

4.3.1 Sailkatzaileak

Hona hemen erabili diren sailkatzaileen zerrenda:

- [Quinlan, 1993] lanean aurkeztutako J48 (sailkatze-zuhaitzak eraikitzeko **C4.5** algoritmoaren inplementazio librea). Sailkapen-zuhaitzak datu-basearen ondoz ondoko banaketa eginez eraikitzen dira, urrats bakoitzean hori egiteko aldagai eta banaketa egokiena hautatzen dira. Hostoek klase bat adierazi ohi dute (bertan erori diren adibide gehienei dagokiena) eta adabegien arteko loturek aldagaiak eta baldintzak.
- Eibe Frank eta Ian Witten-ek proposatuko **PART** algoritmoa, erregelak sortzeko erregela-algoritmoen bi mota nagusiak konbinatzen dituena [Frank and Witten, 1998]: sailkapen-zuhaitzetatik erregelak sortzea, eta erregelak ikasteko ohikoa den banatu eta irabazi teknika. Algoritmoak sailkapen-zuhaitz partzialak sortzen ditu; erabat garatu gabeak dauden C4.5 zuhaitzak, hain zuzen ere.

- **Consolidated Tree Construction (CTC)** [Pérez et al., 2007b]. Lagin anitzetan oinarrituta sailkapen-zuhaitzak sortzen dituen algoritmoa, klase desoreka dagoen kasuetarako egokia eta azalpen egonkorra ematen duena.

- [John and Langley, 1995] lanean aurkeztutako **Naïve Bayes** (**NB**) oinarrizko sailkatzailea.

- Sailkapen-zuhaitzetan (J48) oinarritutako multisailkatzaile bat, [Breiman, 1996b] lanean aurkeztutako **Bagging**.

Sailkatzaile guztiak defektuzko parametroak erabiliz eraiki dira eta haien eraginkortasuna neurtzeko hamar iteraziodun balioztatze-gurutzatua (*10-fold cross-validation*) metodoa erabili da.

4.12 eta **4.13 Taulak** aztertuz hainbat ondorio atera ditzakegu. Lehenik eta behin, esan dezakegu, asmatze-tasei dagokienean, oro har antzekoak direla datu-base osorako eta Lazkaoko datu-baserako lortzen diren emaitzak. Hau da, problemak eta sailkatzaileak berak eragin handiagoa dutela, lokalizazioak baino. Sailkatzaile desberdinen errendimenduan kokatzen bagara, PART algoritmoa erabiliz lortutako erregela-multzoek lortzen dituzten asmatze-tasak, gainerako algoritmoek lortzen dituztenekin alderatuta, txikiagoak dira. Bestalde, azalpenik ematen ez duten sailkatzaileek, bagging eta NB, ez dute asmatze-tasa nabarmenki igotzea lortzen. Hori dela eta, hemendik aurrera baztertu egingo ditugu eta oro har, asmatze-tasarik handienak lortzen dituen C4.5ek (J48) sailkatzailean zentratuko gara.

Galdera	C4.5	CTC	PART	Bagging	NB
Neska, Mutila (ikaslea)	%66.04	%65.25	%63.80	%64.31	%66.98
Emakumea, EzEmak (irudia)	%72.17	%70.44	%69.97	%74.37	%68.40
Betaurrekoa Bai, Ez (irudia)	%57.07	%56.60	%54.09	%57.86	%59.90
Taldean, Bakarrik (irudia)	%77.04	%61.48	%67.29	%75.47	%72.96

4.12 Taula: Datu-base osoko galdera eta algoritmo desberdinentzat asmatze-tasak.

Galdera	C4.5	CTC	PART	Bagging	NB
Neska, Mutila (ikaslea)	%61.73	%68.52	%66.05	%63.58	%64.81
Emakumea, EzEmak (irudia)	%70.37	%69.75	%66.67	%69.14	%67.28
Betaurrekoak: Bai, Ez (irudia)	%56.17	%54.32	%54.32	%53.70	%61.11
Taldean, Bakarrik (irudia)	%75.93	%54.32	%70.37	%72.84	%64.81

4.13 Taula: Lazkaoko datu-baseko galdera eta algoritmo desberdinentzat asmatze-tasak.

Bestalde, galderen erantzunei dagokienean, asmatzen zailena informatikariak betaurrekoekin ala betaurrekorik gabe irudikatzen ote dituzten da (ausaz adieraziko bagenu betaurrekorik baduten ala ez, asmatze-tasa %50ekoa izango litzateke eta kasu honetan %57.07koa da), eta asmatzen errazena, aldiz, informatikariek taldean ala bakarrik lan egiten duten iragarri nahi duena, %77,04ko asmatze-tasarekin.

Emaitza horiek adierazten dute, egin diren galdera gehienetan badaudela desberdintasunak aukera bat eta bestearen artean eta, gainera, galdeketan lortutako erantzunetan oinarrituta antzeman daitezkeela desberdintasun horiek.

4.3.2 Atributuen aukeraketa

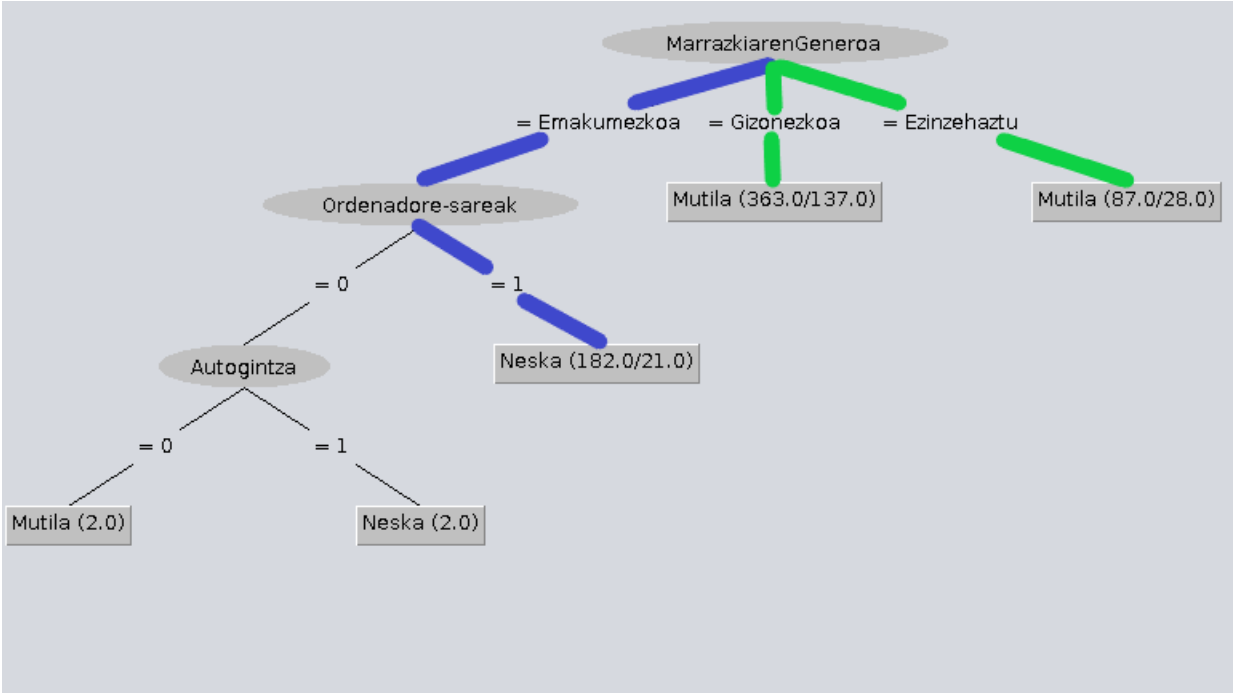
Azalpena lortzeko zuhaitzen egitura aztertu aurretik, azalpen sinpleagoak bilatzeko helburuarekin, eta ahal dela asmatze-tasa txikiagotu gabe, [Garca et al., 2014] lanean gomendatzen den moduan, ikasketa automatikoaren ikuspuntutik esanguratsuenak diren aldagaiak automatikoki hautatzeko algoritmo bat erabili dugu: [Hall, 2000] laneko *Correlation-based Feature Subset Selection* (CFSS) algoritmoa, non korrelazioan oinarriturik aldagai azpimultzo egokia aukeratzen den. Asmatze-tasa mantentzea lortzen badugu, emaitza hobea izango da, aldagai gutxiagorekin lortutako sailkatzailea sinpleagoa izango baita.

4.12 eta **4.13 Taulak 4.14 Taulako** emaitzekin konparatzen baditugu zera ondoriozta daiteke: asmatze-tasetan ez dago diferentzia handirik; izan ere, hainbat kasutan hobetu ere egiten da aldagai hautaketarekin, eta bestalde, eraiki diren sailkatzaileak askoz sinpleagoak izango dira. Menpeko aldagaia eta beste 20 aldagai erabiltzetik, hiru edo lau erabiltzera pasatu baikara. Hautatutako aldagaiak goi-mailan aztertuz, hainbat ondorio orokor atera ditzakegu. Lehenik eta behin, esan dezakegu adin horietan haurren generoak dagoeneko baduela eragina gaztetxoek informatika eta informatikarion inguruan duten ikuspuntuarengan. Izan ere, kasu askotan ageri da aldagai hori hautatua. Estereotipoak guk uste baino lehenago iristen dira gaztetxoengana. Datu-base osoan egindako aldagai hautaketari begiratuz esan dezakegu, haur hori hezi den inguruak ere eragiten duela bere ikuspuntuan. Izan ere, eskola aldagai gisa agertzen da informatikariaren lantokia deskribatzerakoan eta baita betaurrekoak jarri ala ez erabakitzerakoan.

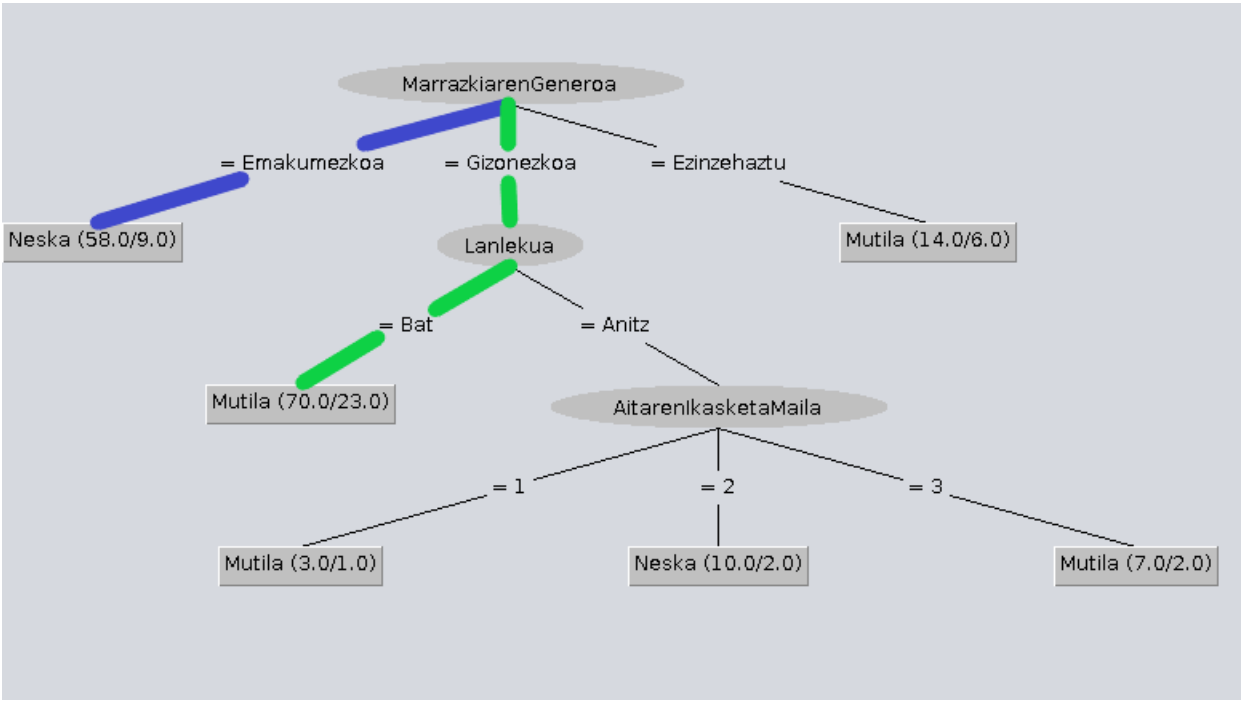
Galdera	**C4.5**	**Hautatutako atributuak**
	DB osoa	
Neska, Mutila (ikaslea)	% 70.44	aitaren ikasketa-maila, irudiaren generoa, autogintza, ordenagailu-sareak
Emakumea, Ez emakumea (marrazkia)	% 75.62	generoa, betaurrekoak, zuzenbidea
Betaurrekoak: Bai, Ez (marrazkia)	% 56.76	eskola, adina, irudiaren generoa, ordenagailu-sareak
Taldean, Bakarrik (marrazkia)	% 77.04	eskola, jantzigintza, etxegintza
	Lazkao DB	
Neska, Mutila (ikaslea)	% 70.37	aitaren ikasketa-maila, lantokia
Emakumea, Ez emakumea (marrazkia)	% 74.70	urtea, generoa, betaurrekoak
Betaurrekoak: Bai, Ez (marrazkia)	% 54.32	irudiaren generoa, artea, kimika, jantzigintza, ordenagailu-sareak
Taldean, Bakarrik (marrazkia)	% 76.54	generoa, amaren ikasketa-maila, autogintza

4.14 Taula: Galdera desberdinentzat asmatze-tasak CFSS aldagai-hautaketa egin ondoren

Aldagai horien eta menpeko aldagaiaren arteko erlazioari buruz gehiago jakin dezakegu eraiki ditugun zuhaitzen egiturari erreparatuz. Adibide gisa, datu-base osorako eta Lazkaoko kasu partikularrerako sortutako bana sailkapen-zuhaitz aztertuko ditugu: **4.6** eta **4.7 Irudietakoak** ikasleen generoak bereizteko eraikitakoak, eta **4.8** eta **4.9 Irudietakoak**, irudiaren generoa bereizteko eraikitakoak. Sailkapen-zuhaitz horien adabegi bakoitzak klase bat adierazten du eta errotik adabegietarainoko bideak aldiz, egindako sailkapenari dagokion azalpena.



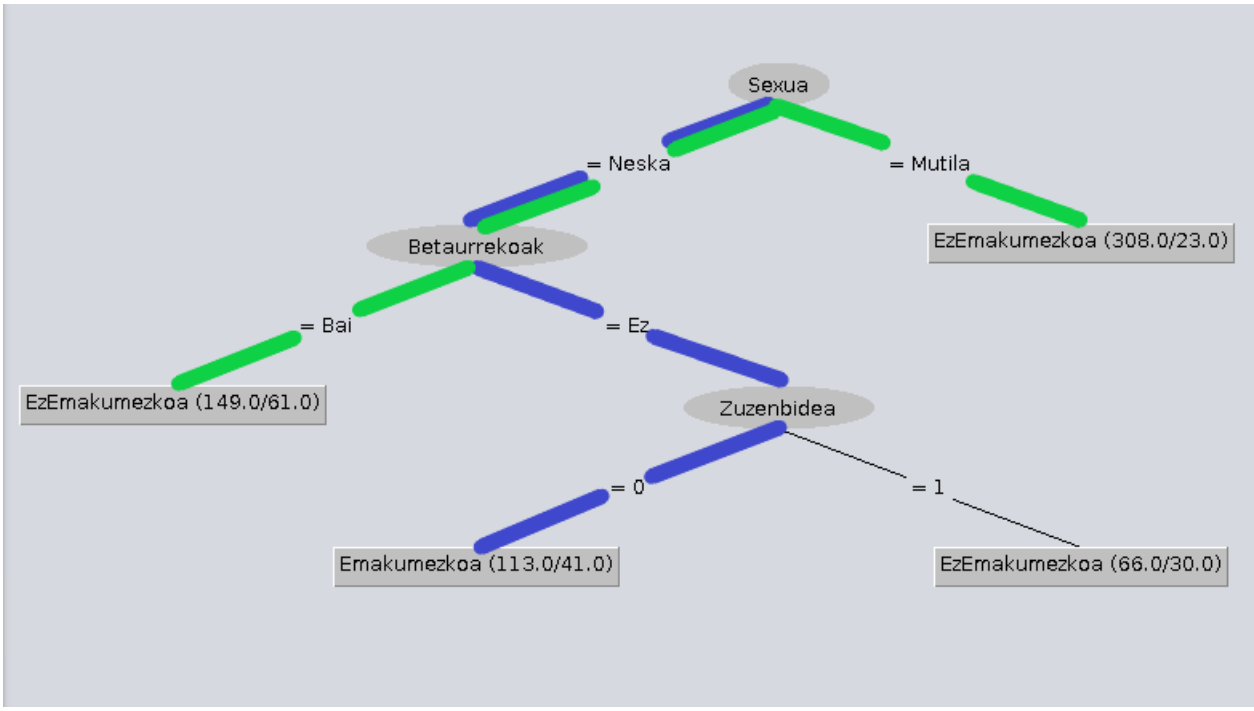
4.6 Irudia: *Datu-base osoan* **ikaslearen generoa** sailkatzeko erabilitako zuhaitza.



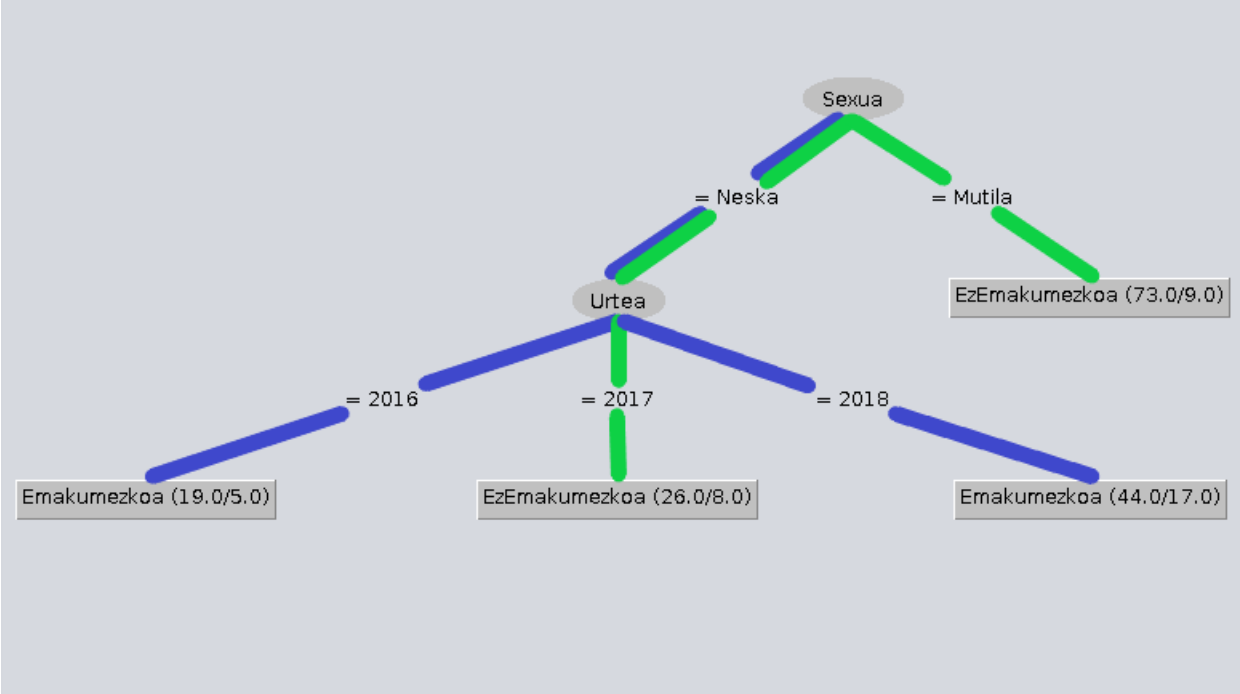
4.7 Irudia: *Lazkaoko datu-basean* **ikaslearen generoa** sailkatzeko erabilitako zuhaitza.

4.6 eta **4.7 Irudiak** aztertuz, esan liteke datu-base osoan zein Lazkaoko datu-basean, ikasleen generoa determinatzeko garrantzia gehien duen aldagaia hauek marrazkiari ematen dioten generoa dela. Nola nahi ere, neskak (urdinez markatuak) eta mutilak (berdez markatuak) bereizteko lortzen diren erregela nagusiak, edo kasu gehien biltzen dituztenak, ez dira berdinak bi kasuetan.

Datu-base osoaren kasuan ondoko bi erregela nagusi ondoriozta genitzazke: (1) neskak dira informatikaria emakumezkoa marraztu eta ordenagailu-sareetan informatika badagoela diotenak; (2) mutilak dira informatikaria gizonezko edo definitu gabe marrazten dutenak. Lazkaoko datu-basearen kasuan, aldiz, lehenengo erregela berdin errepikatzen da baina bigarrenak, aldaketa batzuk ditu, (2) mutilak dira informatikaria definitu gabe marrazten dutenak eta gizonezko marrazteaz gain, lantokia pertsona bakarrekoa marrazten dutenak. Badago aukera gizonezkoa marrazten dutenen artean neskak identifikatzeko. Adibidez, **4.9 Irudian** eskuinera doan marra berdea segiz, (73.0/9.0) zenbakietara iristen gara. Horrek esan nahi du 73 mutiletatik, 9 pertsonak informatikaria emakumezkoa irudikatu dutela.



4.8 Irudia: *Datu-base osoan* **marrazkiaren generoa** sailkatzeko erabilitako zuhaitza.



4.9 Irudia: *Lazkaoko datu-basean* **marrazkiaren generoa** sailkatzeko erabilitako zuhaitza.

4.8 eta **4.9 Irudietako** zuhaitzak aztertzen baditugu, informatikaria emakumezko gisa zeren arabera marrazten duten ondorioztatu ahal izango dugu. Datu-base osoa kontutan hartuz, (1) informatikaria emakume imajinatzen dute neskek baldin eta betaurrekorik gabe imajinatzen badute eta zuzenbidean ez dela informatikarik erabiltzen pentsatzen badute (urdinez markatua). Bestalde, (2) informatikaria ez dute emakume imajinatzen mutilek eta betaurrekoekin irudikatzen duten neskek (berdez markatua). Lazkaoko datu-basean, aldiz, badirudi taldeak edo urteak eragina duela. Kasu honetan, neskek 2016. eta 2018. urtean informatikaria emakume irudikatu zuten (urdinez markatua) baina ez aldiz, 2017. urtean (berdez markatua). Mutilek, aldiz, ez dute emakumezkoa imajinatzen informatikaria (berdez markatua).

4.4 Gainbegiratu gabeko ikasketa

Gainbegiratutako ikasketaz gain, proiektu honetan beste azterketa mota bat ere egin da, gainbegiratu gabeko ikasketa, hain zuzen ere. Kasu honetan, *Clustering* izeneko prozesua erabili da, datuak multzokatzean oinarritzen dena. Datuen multzo osotik horien artean antzekoak diren instantziak azpimultzo edo cluster berdinetan biltzen dira. Horrez gain,

azpimultzo bakoitzeko adibideak, gainerako azpimultzotako adibideekin zertan bereizten diren ikusi nahi izan da. Horretarako, proposatutako galderei lotutako aldagaiak hartu dira kontuan.

Aipatutako cluster horiek sortzeko, *k-Means* algoritmoa erabiltzea erabaki da. Algoritmo hori datu-baseko kasuen arteko distantzien kalkuluan oinarritzen da, instantzien gertutasuna estimatzeko. Zenbakizko aldagaien arteko distantzia kalkulatzea erraza da, hala nola, adina, baina gure datu-basea ez dago zenbakizko aldagaiez soilik osatua. Izan ere, gehienbat aldagai nominalak daude gure datuetan; eskola, ikaslearen generoa, marrazkiaren generoa, etab. Kasu horietarako aukera desberdinak dauden arren, lan honetan oinarrizkoenarekin soilik egin dugu lan.

Atributu nominal bakoitzarekin zera egin da: aldagai horren balio posible bakoitzerako zutabe bitar bat eratu da. Zutabean 1 balioa egongo da aldagaia aukeratua izan bada, eta 0, aldiz, kontrako kasuan. Hori horrela izanik, atributu guztiek zenbakizko balio bat jasotzen dute, eta aldagai nominalek sortzen zuten arazoa konponduta geratzen da.

k-Means algoritmoa aplikatzeko k parametroa definitu behar izan da, cluster analisiaren prozesuarekin aurrera segi ahal izateko. Jakin badakigu, ez dela existitzen partizio bat onena dena kasu guztietan. Beraz, k parametroari balio desberdinak emanez, partizio desberdinak egin dira. Hauek izan dira erabilitako k balioak: 2, 5, 10, 15, 20, 25. Balio horiek hautatu dira, alde batetik, 2 delako clustering prozesurako k parametroak jaso dezakeen balio minimoa, eta 636 izanik datu-basearen tamaina, $\sqrt{636} \approx 25$ delako. Barruti horretako balioak ausaz hautatu dira.

4.4.1 Cluster balidazioa

Egindako partizio desberdinen ontasuna zein den jakiteko, cluster balidaziorako indizeak (*Cluster Validation Index* (CVI)) kalkulatu dira. CVI-ek horietatik partizio onena zein den jakitea ahalbidetu dute. [Arbelaitz et al., 2013] artikuluan zenbait CVI-en emaitzak alderatuz, hauek izan dira onenen artean sailkatutakoak: Silhouette (Sil), Davies-Bouldin (DB), Calinski-Harabasz (CH), COP, Dunn. Horregatik, horiek erabili dira lan honetan.

4.10 Irudian azaltzen dira partizio desberdinetarako lortu diren indizeak. *k-Means* algoritmoa eta distantzia euklidearra erabiliz lortutako indizeak dira hauek. Lehenengo zutabean, probatu diren cluster kopuruak azaltzen dira, eta indizeen artean, berdez azpimarratuta daude indize mota bakoitzeko balio onena.

	Sil	DB*	CH	DB	COP	Dunn		
						gD33	gD43	gD53
2	0.0648	9.776	39.4842	3.7635	1.5093	1.5005	0.5297	1.131
5	0.0691	7.5252	37.8435	3.0331	1.3994	1.4814	0.5881	1.1368
10	0.0745	6.7674	29.2414	2.7664	1.3003	1.408	0.6201	1.1136
15	0.0583	6.5519	22.7578	2.7384	1.2675	1.3438	0.5659	1.0552
20	0.0548	6.6065	18.7782	2.7298	1.269	1.3008	0.5707	1.0236
25	0.0581	6.0978	16.6353	2.6133	1.2239	1.3382	0.5999	1.0562

4.10 Irudia: *Cluster balidaziorako indizeak*

Silhouette (Sil) indizearen balioak, orokorrean, oso baxuak dira. Horrek esan nahi du banatzen diren clusterrak banatzen direla, haien artean ez dutela antzekotasun handirik. Kasu honetan, 10 clusterreko partizioa egin ostean lortzen da balio altuena (0.07).

Davies-Bouldin (DB) eta **Davies-Bouldin* (DB*)** indizeen kasuan, zenbat eta cluster kopuru gehiagoko banaketa egin, orduan eta partizio hobea da. Argi dago balio onena 25 clusterrekin lortu dela, baina gainerakoak ez daude balio horretatik oso urrun.

Calinski-Harabasz (CH) indizeari dagokionez, zenbat eta handiagoa denean cluster kopurua, orduan eta partizio okerragoa dela esan daiteke. Emaitza onena 2 clusterreko partizioa dela ondorioztatu da.

COP indizearen kasuan, lortutako balioa zenbat eta txikiagoa, orduan eta partizio hobea kontsideratzen da. Ateratako emaitzak kontuan izanik, 25 clusterreko partizioa da aukerarik onena. Esan beharra dago indizeen balioen artean ez dagoela desberdintasun nabarmenik, hau da, indizeen balioak oso antzekoak direla.

Dunn indizeak kalkulatzerakoan, hiru zutabetan kalkulatu dira **(gD33, gD43, gD53)**. Hiru zutabe horietan antzeko zerbait gertatu da. Zenbat eta cluster kopuru txikiagoa, orduan eta partizio hobea lortu da. 2, 5 eta 10 clusterreko partizioak izan dira aukera onenak.

Gehiengoari erreparatuko bagenio, esango genuke 25 clusterreko partizioa dela aukera onena, gehiengoak hala diolako. Hala ere, aipatu beharra dago ez direla gainerako balioengandik asko urruntzen. Hau da, oso antzekoak direla lortutako emaitzak. Egia da 2 clusterreko partizioa ere nabarmendu dela, gainerakoekiko aldea oso handia izan gabe; beraz, kontuan izan daitekeen partizio bat ere izan daiteke.

Azpimarratzekoa da, indizeen balioek adierazten dutela partizioen kalitatea ez dela oso ona eta posible da (probabilitate handiz) hori lotuta egotea aldagai nominal kopurua eta horiek kodetzeko erabili den moduarekin. Beraz, etorkizunerako, lanean sakontzeko esparru zabal bat egon daiteke ildo horretan.

4.4.2 Cluster analisirako partizioak

CVI indizeak aztertu ostean, honako partizioak irudikatzea hautatu da: 2, 10 eta 25 cluster. Gainbegiratutako sailkapenean, aurretik proposatutako galderei erantzun bat emateko zenbait aldagai hartu dira kontuan, hala nola, ikaslearen generoa, marrazkiaren generoa eta lanlekua. Gainbegiratu gabeko sailkapenean, multzokatutako datuak atributu horien arabera nola banatzen diren aztertu nahi izan da. Horiez gain, haurrak hezten diren inguruak haiek emandako erantzunetan zerikusirik baduen aztertzeko, eskola ere aztertzea erabaki da. Aldagaiek balio desberdinak hartzen badituzte partizio bereko cluster desberdinetan, aldagai hori partizioa egin ahal izateko esanguratsua izan dela esan ahal izango genuke, aldagai horrek ikasle batzuk besteetatik bereizten dituela, alegia.

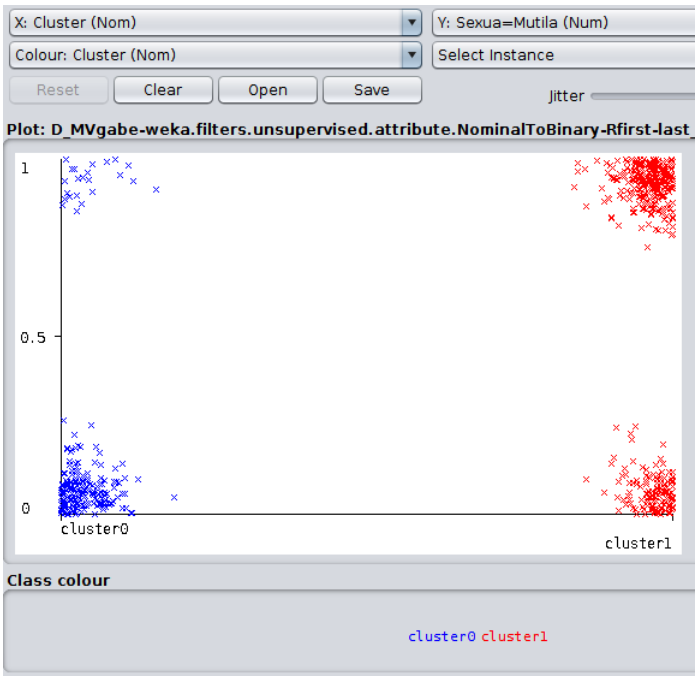
25 clusterreko partizioak aztertu eta gero, ondorioztatu da ezin dela ezer berezirik ikusi, hau da, cluster gutxiagoko partizioetatik atera denaz gain ez dela ezer ateratzen. Ondorioz, 2 eta 10 clusterreko partizioak lehenetsi dira azterketa egiterako orduan. Era berean, eskolen araberako partizioak ere aztertu dira, baina ez da ezer garbia bereiztea lortu.

Partizioa X eta Y ardatzeko espazioan irudikatu da. X ardatzean sortutako cluster desberdinak azaltzen dira (kolore desberdinez), eta Y ardatzean, berriz, aztertu den aldagaiaren balioak, kasu honetan 0 edo 1 balioak jaso dituztenak.

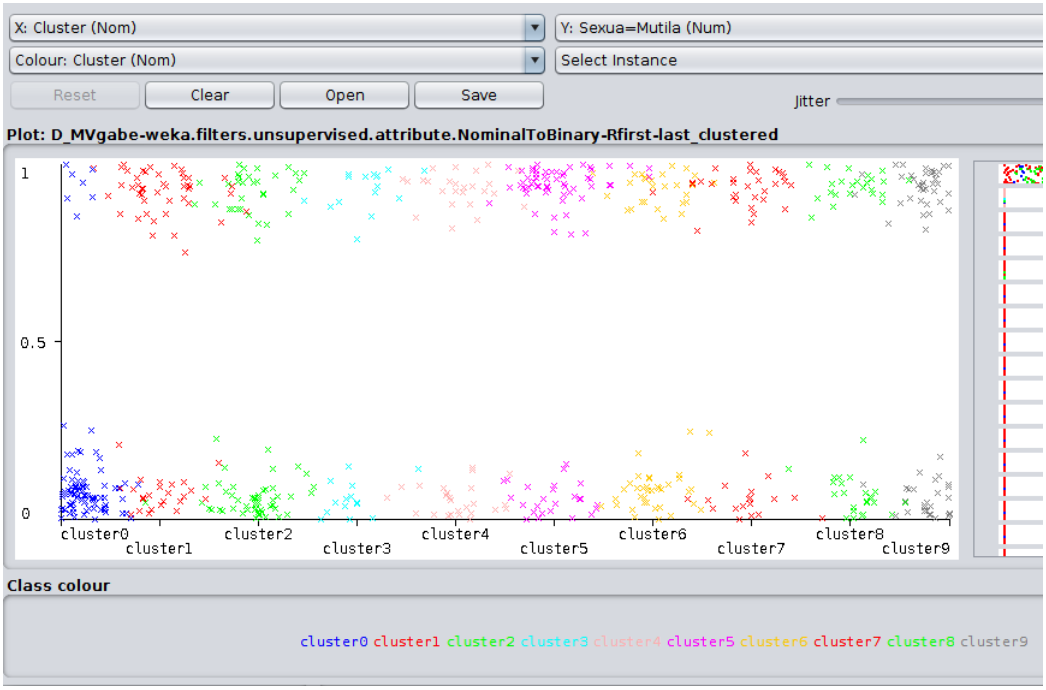
Ikaslearen generoa

Ikaslearen generoak honako balioak hartu ditu: 1 = mutila eta 0 = neska. **4.11 Irudian** ikus daitekeen bezala, Y ardatzean goiko partean dauden instantziak mutilak dira, 1 balioa dutelako, eta beheko partean daudenak, berriz, neskak, 0 balioa dutelako. Dena den, bi clusterretan banatu dira gero, urdina eta gorria. Bietan 0 eta 1 daudenez (neskak eta mutilak), partizio honetan ez dira neskak eta mutilak bereiztu, baina esan daiteke neskak gehienbat cluster urdinean bildu direla eta mutilak, aldiz, gorrian. Beraz, ikaslearen generoa atributu esanguratsua, baina ez determinantea dela esan dezakegu.

Hamar clusterreko partizioan (**4.12 Irudia**), esan beharra dago, cluster gehienetan neska eta mutil kopurua parekatua dagoela. Hala ere, azpimarratzekoa da, adibidez, *cluster0* eta *cluster2* azpimultzoetan gehienak neskak direla. *Cluster5*-n, berriz, mutilak daude gehienbat. Beraz, partizio honetan, bi clusterreko partizioan ateratako konklusio bera ondorioztatu dezakegu, neska edo mutil izateak, ikasleak multzo desberdinetan banatzen dituela neurri batean.



4.11 Irudia: *2 clusterreko* partizioa, **ikaslearen generoaren** arabera irudikatuta.

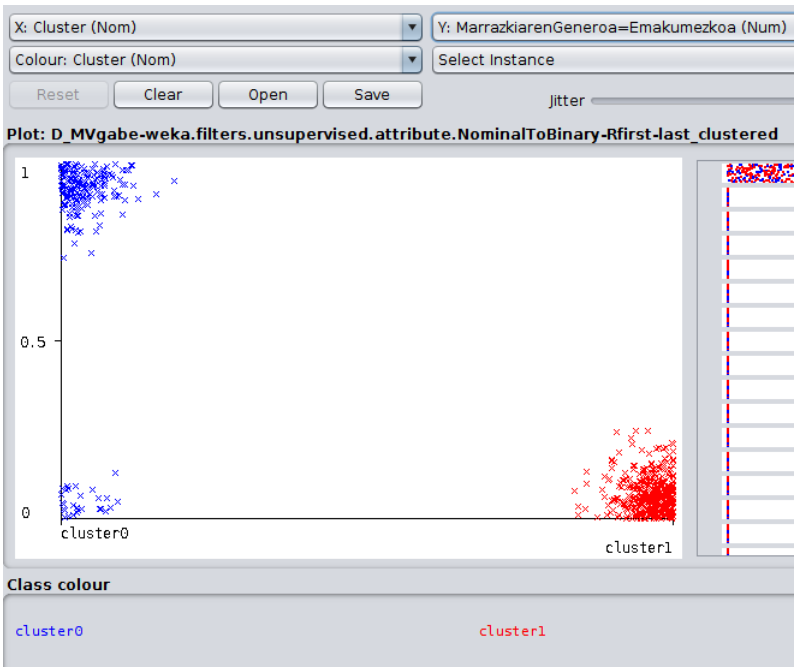


4.12 Irudia: *10 clusterreko* partizioa, **ikaslearen generoaren** arabera irudikatuta.

Marrazkiaren generoa

Marrazkiaren generoak honako balioak hartu ditu: 1 = emakumezkoa eta 0 = gizonezkoa edo ezin zehaztu.

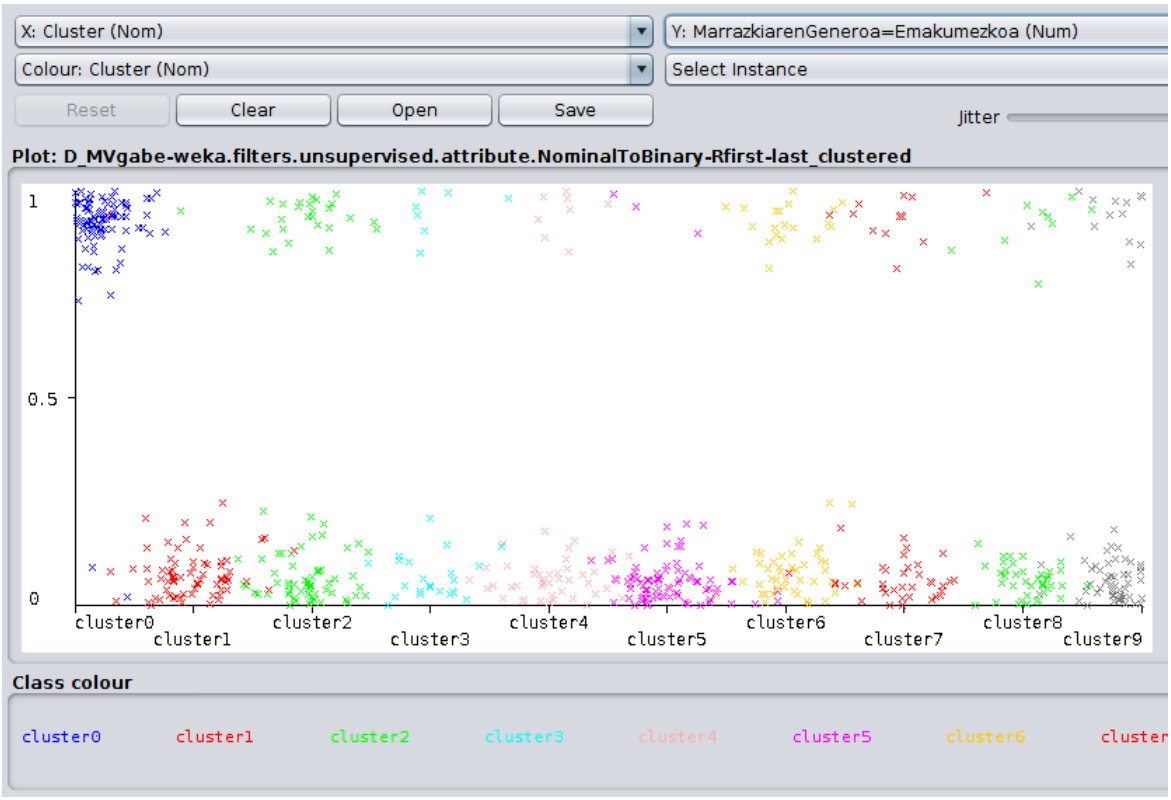
Bi clusterreko partizioan (**4.13 Irudia**), nabarmena da marrazkiaren generoa aldagai esanguratsua dela. Izan ere, cluster banatan zatitu dira, alde batetik, marrazkian informatikaria emakumezkoa irudikatu dutenak (*cluster0*), eta bestetik, marrazkian informatikaria gizonezko edota ezin zehaztu daitekeen generokoa irudikatu dutenak (*cluster1*). Dena den, *cluster0*-n badaude batzuk gizonezko edota ezin zehaztu daitekeen generokoa marraztu dutenak. Horiek zein beste aldagairen arabera desberdintzen diren aurrerago aztertuko da. Ondorioz, esan daiteke marrazkiaren generoa aldagai oso esanguratsua dela, bi clusterren artean ez baitago antzekotasunik.



4.13 Irudia: *2 clusterreko* partizioa, **marrazkiaren generoaren** arabera irudikatuta.

4.14 Irudiari erreparatuz, kasu batzuetan argi ikusten da cluster bateko instantzia guztiak marrazkiaren generoari balio bera eman diotela. Adibide gisa, *cluster0*-n bertako instantzia ia guztiek irudikatu dute informatikaria emakume, Y-ardatzeko 1 balioan kokatuta baitaude gehienak, 0 balioan kokatuta dauden salbuespen batzuk dauden arren. Kontrako kasua adierazteko, esaterako *cluster1* edo *cluster5*-en gehienak informatikariaren generoa emakumezkoa ez den genero batez irudikatu dute (gizonezkoa edo ezin zehaztu). Aurretik gertatu den bezala, partizio honetan badaude ere beste hainbeste kasu, non aldagaiari

balio desberdinak eman dizkieten instantzia kopuruak parekatuta dauden. Ondorioztatu daiteke, marrazkiaren generoa esanguratsua dela.



4.14 Irudia: *10 clusterreko* partizioa, **marrazkiaren generoaren** arabera irudikatuta.

Ikaslearen generoa eta marrazkiaren generoa aldagaien erlazioa

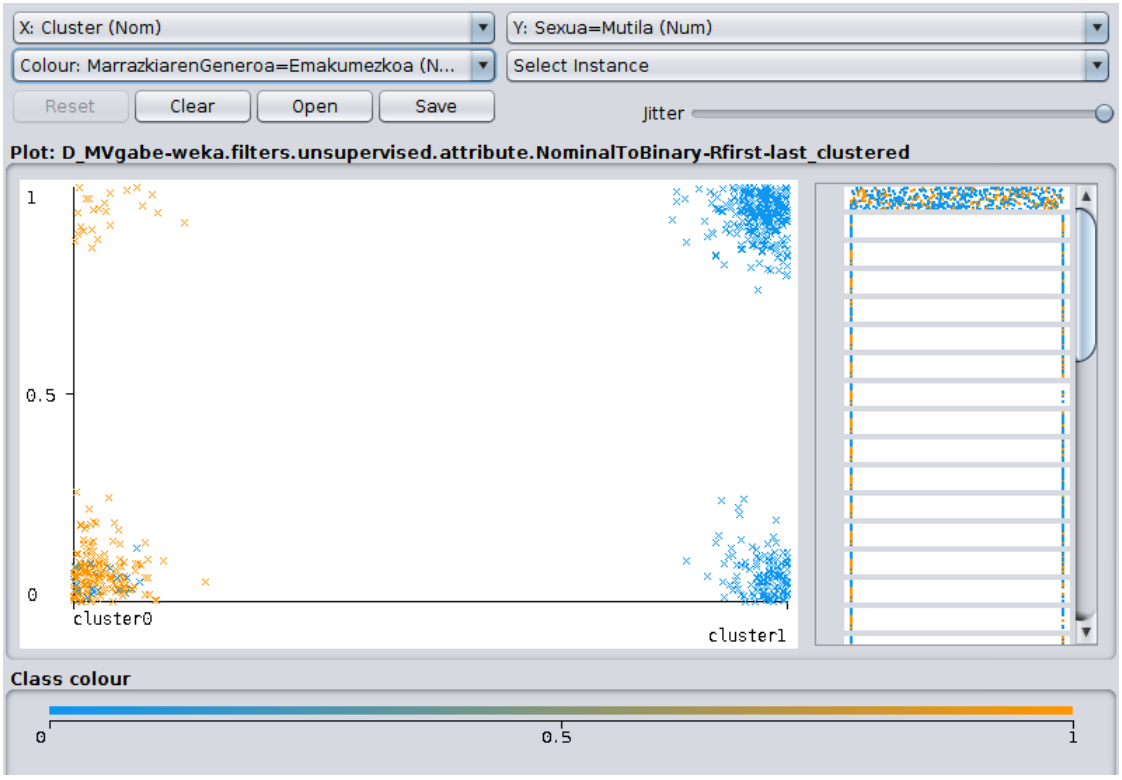
Ikaslearen generoa eta marrazkiaren generoa aldagaien erlazioa ikusi ahal izateko, bi clusterreko partizioarekin proba egin da.

4.15 Irudiari begiratuz, aurreko kasuetan ikusi den bezala, X ardatzean clusterrak ditugu eta Y ardatzean, berriz, aldagaiari eman zaizkion balioak. Kasu honetan, ikaslearen generoa aztertu da (1 = mutila eta 0 = neska edo zehaztu gabea). Horretaz gain, marrazkiaren generoa koloreztatu da. Beste era batera esanda, marrazkian informatikaria emakumezko gisa irudikatu dutenak laranjaz koloreztatu dira, eta gizonezko edo zehaztu ezin den generokoak irudikatutako instantziak, berriz, urdinez koloreztatu dira. Beraz, irudiko goiko partean eta ezkerrean, informatikaria emakumezko irudikatu duten mutilak daude. Goian eta eskuinean, informatikaria gizonezko irudikatu duten mutilak. Beheko aldean eta ezke-

rrean, gehienbat informatikaria emakumezko irudikatu duten neskak pilatu dira. Azkenik, beheko aldean eta eskuinean, informatikaria gizonezko irudikatu duten neskak daude.

Nabaria da partizioa ikaslearen generoaren araberakoa izan dela. Izan ere, *cluster0*-n gehiengoa neskak dira (Y=0) eta *cluster1*-n gehienak mutilak dira (Y=1). Gainera, koloreari dagokionez, *cluster0*-n dauden ia denek, marraztu dute informatikaria emakumezkoa (laranja), eta *cluster1*-n daudenek, aldiz, gizonezkoa edo zehaztu gabekoa.

Ondorioz, esan daiteke orokorrean neskak izan direla informatikaria emakumezkoa irudikatu dutenak, eta mutilak izan direla informatikaria gizonezko edo zehaztu ezin den generokoa marraztu dutenak. Hala ere, aipatu beharra dago, salbuespen batzuk badaudela: badaude mutil batzuk informatikaria emakumezkoa marraztu dutenak eta beste neska batzuk informatikaria gizonezkoa irudikatu dutenak. Dena den, neska gehiago izan dira informatikaria gizonezko edo zehaztu ezineko generokoa irudikatu dutenak, emakumezko marraztu duten mutilak baino. Horrek guztiak gainbegiratutako ikasketan ateratako ondorioak egiaztatzen ditu.

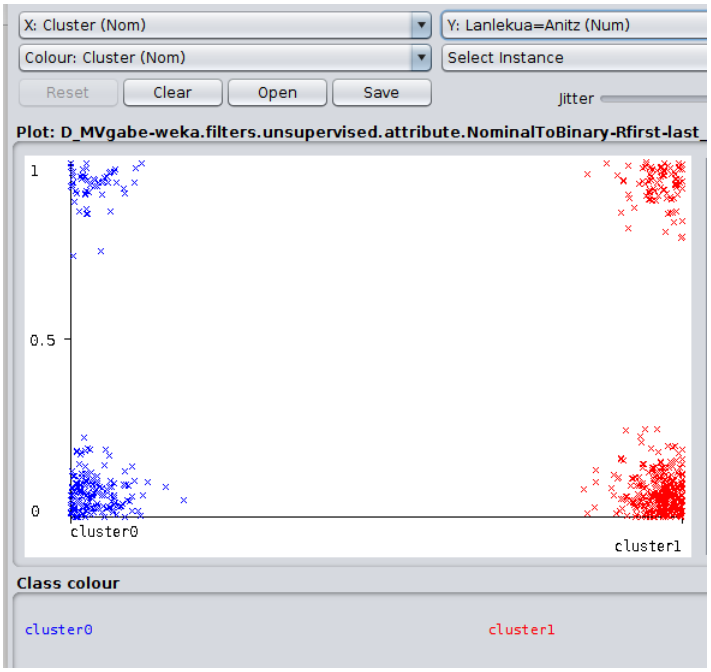


4.15 Irudia: *2 clusterreko* partizioa, **ikaslearen generoa eta marrazkiaren generoa** aldagaien arabera irudikatuta.

Lanlekua

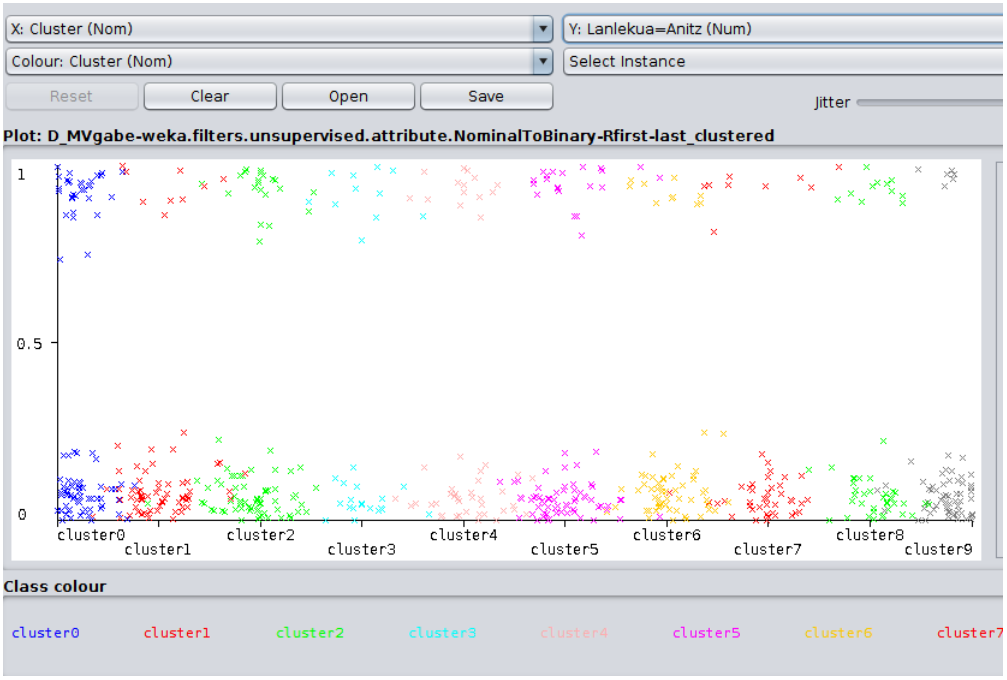
Haurrei informatikari baten lanlekua irudikatzeko eskatu zitzaien, eta beraz, lanlekua aldagaiak honako balioak hartu ditu: 1 =“ordenagailu anitz lanlekuan” eta 0 =“ordenagailu bakarra lanlekuan”. Hortik ondorioztatu da gazteek informatikaria bakarrik edo taldean imajinatzen duten lantokian.

Partizio honetako bi clusterren instantzien joera berdina dela esan daiteke (**4.16 Irudia**). Bi azpimultzoetan instantzia gehienetan hau erantzun da: informatikaria bakarrik lanlekuan. Beraz, esan daiteke informatikarien lanlekua deskribatzen duen atributua ez dela bereziki esanguratsua.



4.16 Irudia: *2 clusterreko* partizioa, **lanlekua** aldagaiaren arabera irudikatuta.

Aurreko partizioaren antzera, kasu honetan ere argi ikusten da partizioa ez dela egin aldagai honen arabera (**4.17 Irudia**). Izan ere, clusterretan nabaritzen da adibide gehienetan, gehiengoak lanlekuan informatikari bakarra irudikatu duela. Hala ere, salbuespena *cluster0* dela esango genuke, cluster horretan lanlekuan ordenagailu anitz eta ordenagailu bat irudikatu dutenak parekatuago baitaude.



4.17 Irudia: *10 clusterreko* partizioa, **lanlekua** aldagaiaren arabera irudikatuta.

5. KAPITULUA

Erabilitako tresnak

Atal honetan, proiektua garatzeko erabili diren tresnak aurkeztuko dira. Behin datuak paper bidezko inkesta bidez bilduta izanik, **.csv** formatuan digitalizatu dira hasierako datu-baseak, hau da, taula bidezko errepresentazioa duen formatuan.

5.1 *Python* programazio lengoaia

Datu-multzo horiek kudeatzeko, zenbait programa txiki sortu dira. ***Python*** interpretatutako lengoaia bat da, zeinak sintaxi garbia, erraza eta ulergarria duen. [Python.org, 2019] Tresna hori arrazoi honengatik aukeratu da: taula bidezko errepresentazioa duten fitxategiekin lan egitea oso erosoa da, datu-analisirako liburutegi egokiak baititu, hala nola, ***Pandas*** liburutegia. Liburutegi hori, datu-analisirako erabiltzen den NumPy paketearen hedapena da. Bereziki, datu-egiturak eta zenbakizko taulak eta denbora sekuentziak manipulatzeko eragiketak eskaintzen ditu [Pandas.pydata.org, 2019].

Alde batetik, ikasturte desberdinetako galdetegiak bateratzerako orduan, zenbait ordezkapen egin behar izan dira. Bestalde, 2016/17 eta 2017/18 ikasturteetako datu-basean zutabeak ere aldatu behar izan dira. Hori, informatikariak ogibide zehatzetan duen parte-hartzearekin lotutako galderagatik egin da. Aurreko atalean aipatu bezala, hori izan baita desberdintasun nagusia.

Orokorrean, ***Python*** programazio lengoaiaz baliatu gara datu-baseak txukundu eta ikasketa-automatikorako prest uzteko. Ikasketarako erabili den tresna kontuan izanik, datu-baseen formatua ere aldatu behar izan da, hau da, .csv batetik **ARFF** formatura pasa behar izan da. Fitxategi mota horrek testu formatua du eta edozein testu editoreren bidez ikusi edo moldatu daiteke. Ezaugarri gisa esan daiteke .arff fitxategi bat bi zatitan bereizten dela: bata atributuei buruzko informazioa duen atala, eta bestea, datuak berak.

5.2 Weka softwarea

Ikasketa-automatikoko algoritmoak aplikatzerako orduan, ***Weka*** [Hall et al., 2009] izeneko software bat erabili da. Wekak datu-meatzaritzako zereginak betetzeko hainbat algoritmo eskaintzen ditu. ***Python*** programazio lengoaiaren bidez sortutako .arff fitxategiak sartuz, ondorengoa egiteko aukera eman du software horrek:

Gainbegiratutako ikasketari dagokionez, alde batetik, datuen sailkapenerako zenbait sailkatzaile eskaintzen ditu, horien artean, guk erabilitakoak. Sailkapen-zuhaitzei dagokienez, zuhaitzak berak irudikatzeko aukera ere ematen du. Bestalde, aldagai aukeraketa ere Wekako beste funtzio batekin egin da.

Gainbegiratu gabeko ikasketari dagokionez, berriz, cluster analisia software horren laguntzaz egin ahal izan da. Lehenik eta behin, partizioak guk emandako parametroen eta .arff-ko datuen arabera egin dira eta ondoren, cluster bakoitzaren irudikapen argia eskaini digu.

6. KAPITULUA

Ondorioak eta etorkizunerako lana

Atal honetan, proiektuaren ondorioak zein diren adieraziko da, hasiera batean planteatu ziren galderekiko. Proiektuaren motibazioa informatika ikastera iristen diren emakumeen kopuru urriaren jatorria aztertzea izanik, lan honetan hiru urtetan zehar Euskal Herriko hainbat eskoletan 10-12 urte arteko gaztetxoei egindako inkesta batzuez baliatu gara, adin horretan duten iritzia aztertzeko. Datu-meatzaritzako prozesu oso bat egin behar izan da horretarako: datuak bildu, aurreprozesatu, erantzun nahi genituen galderak erantzuteko prestatu, sailkatzaileekin erabiltzeko egokitu, sailkatzaileak eraiki edo cluster analisia egin, eta aztertu. Prozesu horri esker, jasotako inkestetan parte hartu duten ikasleei buruz hainbat ondorio atera dira.

Lehenik eta behin, neskek eta mutilek informatika eta informatikariei buruz duten ikuspuntua adin horretan desberdina dela ondorioztatu da, eta %70eko asmatze-tasarekin bereizteko gai gara. Hau da, %70eko ziurtasunarekin bereiztu daitezke mutilak eta neskak gainerako galderei emandako erantzunak kontuan izanik. Asmatze-tasa hori zerbait igotzen da haurrek informatikariari irudikatzen dioten generoa eta lanean bakarrik edo taldean imajinatzen ote duten asmatzen saiatzen garenean, eta jaitsi, aldiz, betaurrekoekin irudikatzen ote dituzten asmatzen saiatzen garenean. Ikus **4.3.1** atala. Horrez gain, aldagai-hautaketa eta sailkapen-zuhaitzak aztertuz, esan dezakegu haurren bizi-inguruneak (eskolak) eta taldeak badutela eragina haurrek gai horren inguruan duten iritziaren gainean. Ikus **4.3.2** atala.

Bestalde, *Freaky* estereotipoari dagokionez, informatikariaren marrazkian betaurrekoen agerpena kontuan izanik, esan daiteke ez dela bereziki atentzioa ematen duen ezaugarri bat. Izan ere, betaurrekoekin irudikatutako informatikarien ehunekoa ez da oso altua izan. Ikus **4.2.4** atala.

Azpimarratzekoa da gainbegiratutako ikasketan zein clustering prozesuan ateratako ondorioak antzekoak direla. Lortutako emaitzen artean, nabarmentzekoa da badirudiela gure inguruko 10-12 urteko mutilek oro har, gizonezkoak edo identifikatu ezinezko generoa duten gizakiak irudikatzen dituzten arren, neskek oraindik hein handi batean ikusten dutela informatikaria neska izan daitekeela. Emaitza horrek adierazten du, adin-tarte horretara-

ko prestatutako ekintzek eragina izan dezaketela etorkizunean informatika ikasiko duten emakume kopuruarengan.

Egindako datuen azterketatik haratago, posible da beste ikasketa mota batzuk egitea. Adibidez, *missing values* edo balio-hutsak kudeatzeko beste teknika adimentsuagoak erabili daitezke, edo gainbegiratu gabeko ikasketan aldagai nominalak tratatzeko modua ere alda daiteke, instantzien arteko antzekotasuna kalkulatzeko distantzia mota alda daitekeen bezalaxe. Dena den, gainbegiratutako eta gainbegiratu gabeko ikasketa automatikoko prozesua hasieran planteatu diren galderek baldintzatu dute, eta baliteke, hori horrela izanik, datuetan benetan existitzen ez zen egitura bat bilatzen aritu izana. Bildutako datuei dagokienez, agian jasotako informazioa ez da nahikoa izan egin nahi izan den azterketarako, eta posible da datu gehiago biltzea egokiagoa izatea. Hala nola, haurrek zer izan nahi duten helduak direnean galdetzea.

Bibliografia

[Ali and Smith, 2006] Ali, S. and Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119–138.

[Arbelaitz et al., 2013] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

[Breiman, 1996a] Breiman, L. (1996a). Bagging predictors. *Machine learning*, 24(2):123–140.

[Breiman, 1996b] Breiman, L. (1996b). Bagging predictors. *Mach. Learn.*, 24(2):123–140.

[Brown et al., 2014] Brown, N., Sentance, S., Crick, T., and Humphreys, S. (2014). Restart: The resurgence of computer science in uk schools. *ACM Transactions on Computing Education (TOCE)*, 14(2):9.

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

[Chambers, 1983] Chambers, D. W. (1983). Stereotypic images of the scientist: The draw-a-scientist test. *Science education*, 67(2):255–265.

[Clifton, 2017] Clifton, C. (2017). Data mining. https://www.britannica.com/technology/data-mining.

[Code.org, a] Code.org. Code.org: About us. https://code.org/about. [Online; 2019ko otsailaren 13an eskuratua].

[Code.org, b] Code.org. The hour of code. https://hourofcode.com/es/en. [Online; 2019ko otsailaren 13an eskuratua].

[Cohen, 1995] Cohen, W. W. (1995). Fast effective rule induction. In *Machine Learning Proceedings 1995*, pages 115–123. Elsevier.

[Dagiene, 2002] Dagiene, V. (2002). The model of teaching informatics in lithuanian comprehensive schools. *Journal of Research on Computing in Education*, 35(2):176–185.

[Davies and Bouldin, 1979] Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

[DeRuy, 2017] DeRuy, E. (2017). In finland, kids learn computer science without computers. https://www.theatlantic.com/education/archive/2017/02/teaching-computer-science-without-computers/517548/. [Online; 2019ko Otsailaren 13an eskuratua].

[Dunn, 1973] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

[Fayyad et al., 1996a] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

[Fayyad et al., 1996b] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et al. (1996b). Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88.

[Figueiredo et al., 2016] Figueiredo, M., Esteves, L., Neves, J., and Vicente, H. (2016). A data mining approach to study the impact of the methodology followed in chemistry lab classes on the weight attributed by the students to the lab work on learning and motivation. *Chemistry Education Research and Practice*, 17(1):156–171.

[Frank and Witten, 1998] Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. *Generating Accurate Rule Sets Without Global Optimization*, pages 144–151. cited By 46.

[Garca et al., 2014] Garca, S., Luengo, J., and Herrera, F. (2014). *Data Preprocessing in Data Mining*. Springer Publishing Company, Incorporated.

[Gower, 1971] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

[Gurrutxaga et al., 2010] Gurrutxaga, I., Albisua, I., Arbelaitz, O., Martín, J. I., Muguerza, J., Pérez, J. M., and Perona, I. (2010). Sep/cop: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition*, 43(10):3364–3373.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

[Hall, 2000] Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Hanson et al., 1991] Hanson, R., Stutz, J., and Cheeseman, P. (1991). Bayesian classification theory.

[Jiawei Han, 2011] Jiawei Han, Micheline Kamber, J. P. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 225 Wyman Street, Waltham, USA, 3rd edition.

[John and Langley, 1995] John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Johnson and Turner, 2003] Johnson, B. and Turner, L. A. (2003). Data collection strategies in mixed methods research. *Handbook of mixed methods in social and behavioral research*, pages 297–319.

[Knight and Cunningham, 2004] Knight, M. and Cunningham, C. (2004). Draw an engineer test (daet): Development of a tool to investigate students' ideas about engineers and engineering. In *ASEE Annual Conference and Exposition*, volume 2004.

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Medel and Pournaghshband, 2017] Medel, P. and Pournaghshband, V. (2017). Eliminating gender bias in computer science education materials. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*, pages 411–416. ACM.

[Mingers, 1989] Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. *Machine learning*, 3(4):319–342.

[Netherlands, 2013] Netherlands, S. (2013). More than 6 in 10 people wear glasses or contact lenses. https://www.cbs.nl/en-gb/news/2013/38/more-than-6-in-10-people-wear-glasses-or-contact-lenses. [Online; 2019ko uztailaren 22an eskuratua].

[Osang et al., 2013] Osang, J., Udoimuk, A., Etta, E., Ushie, F., and Offiong, N. (2013). Methods of gathering data for research purpose and applications using ijser acceptance rate of monthly paper publication (march 2012 edition-may 2013 edition). *IOSR Journal Of ComputerEngineering (IOSR-JCE)*, 15(2):59–65.

[Pandas.pydata.org, 2019] Pandas.pydata.org (2019). https://pandas.pydata.org/. [Online; 2019ko uztailaren 25ean eskuratua].

[Parsania et al., 2014] Parsania, V., Bhalodiya, N., and Jani, N. (2014). Applying naïve bayes, bayesnet, part, jrip and oner algorithms on hypothyroid database for comparative analysis.

[Patil et al., 2013] Patil, T. R., Sherekar, S., et al. (2013). Performance analysis of naive bayes and j48 classification algorithm for data classification. *International journal of computer science and applications*, 6(2):256–261.

[Pérez et al., 2007a] Pérez, J. M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., and Martín, J. I. (2007a). Combining multiple class distribution modified subsamples in a single tree. *Pattern Recognition Letters*, 28(4):414–422.

[Pérez et al., 2007b] Pérez, J. M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., and Martín, J. I. (2007b). Combining multiple class distribution modified subsamples in a single tree. *Pattern Recogn. Lett.*, 28(4):414–422.

[Peterson, 2009] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

[Python.org, 2019] Python.org (2019). Welcome to python.org. https://www.python.org/. [Online; 2019ko uztailaren 25ean eskuratua].

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[Rubin, 1978] Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.

[Saar-Tsechansky and Provost, 2007] Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research*, 8(Jul):1623–1657.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

[Sharma et al., 2012] Sharma, N., Bajpai, A., and Litoriya, M. R. (2012). Comparison the various clustering algorithms of weka tools. *facilities*, 4(7):78–80.

[Singh et al., 2013] Singh, A., Yadav, A., and Rana, A. (2013). K-means with three different distance metrics. *International Journal of Computer Applications*, 67(10).

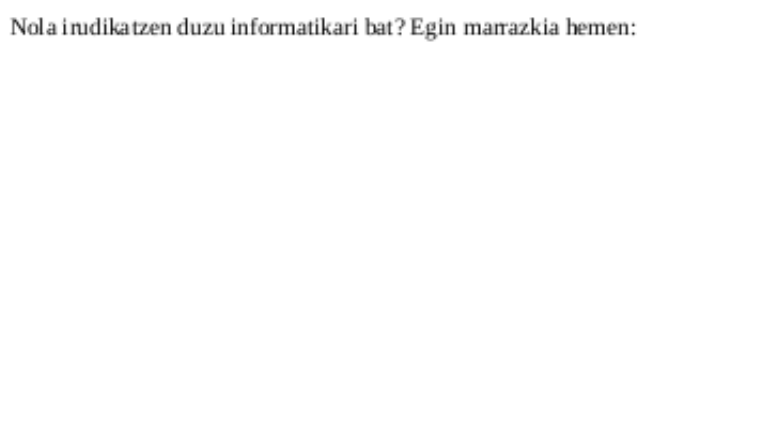
[Talpur, 2017] Talpur, A. (2017). *Congestion Detection in Software Defined Networks using Machine Learning*. PhD thesis, Master's thesis, University of Bremen, Germany.

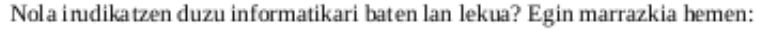
[Witten et al., 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Eranskinak

A. ERANSKINA

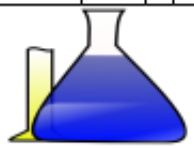
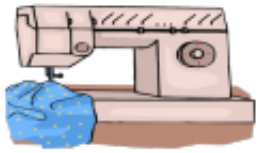
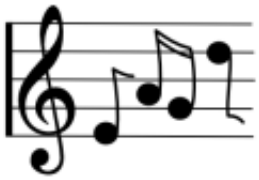
2015/16 ikasturteko galdetegia





A.1 Irudia: *2015/16* ikasturteko galdetegia.

Ondoren lanbide batzuk aurkeztuko dizkizugu. Horietako zeinean aritu zitekeen informatikari bat zure ustez? Markatu Bai ala Ez ondoan dagoen koadrotxoa.

Medikuntza	Bai		Ez		**Artea**	Bai		Ez		**Kimika**	Bai		Ez	
														
Autogintza	Bai		Ez		**Jantzigintza**	Bai		Ez		**Etxegintza**	Bai		Ez	
														
Bideo-jokoak	Bai		Ez		**Musikagintza**	Bai		Ez		**Zuzenbidea**	Bai		Ez	
														
Hizkuntzalaritza	Bai		Ez		**Ordenadore sareak**	Bai		Ez		**Hezkuntza**	Bai		Ez	
														

Zenbat urte dituzu?.....

Zer zara? Neska ☐ Mutila ☐

Amaren lanbidea:.....

Aitaren lanbidea:.....

A.2 Irudia: *2015/16* ikasturteko galdetegia.

B. ERANSKINA

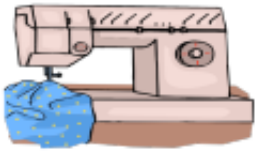
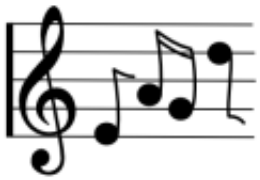
2016/17 eta 2017/18 ikasturteetako galdetegia

Nola irudikatzen duzu informatikari bat? Egin marrazkia hemen:

Nola irudikatzen duzu informatikari baten lan lekua? Egin marrazkia hemen:

B.1 Irudia: *2016/17 eta 2017/18* ikasturteetako galdetegia.

Ondoren lanbide batzuk aurkeztuko dizkizugu. Horietako zeinean aritu zitekeen informatikari bat zure ustez? Markatu Bai ala Ez ondoan dagoen koadrotxoa.

Medikuntza	Bai ☐	Ez ☐	**Artea**	Bai ☐	Ez ☐	**Kimika**	Bai ☐	Ez ☐
								
Autogintza	Bai ☐	Ez ☐	**Jantzigintza**	Bai ☐	Ez ☐	**Etxegintza**	Bai ☐	Ez ☐
								
Bideo-jokoak	Bai ☐	Ez ☐	**Musikagintza**	Bai ☐	Ez ☐	**Zuzenbidea**	Bai ☐	Ez ☐
								
Hizkuntzalaritza	Bai ☐	Ez ☐	**Ordenadore sareak**	Bai ☐	Ez ☐	**Hezkuntza**	Bai ☐	Ez ☐
								

Zenbat urte dituzu?.....

Zer zara? Neska ☐ Mutila ☐ Ez bitarra ☐

B.2 Irudia: *2016/17 eta 2017/18* ikasturteetako galdetegia.

Galdetegiaren atal hau gurasoen laguntzarekin bete:

Guraso baten **ikasketa maila**

- ☐ Lehen mailako hezkuntza, OHO edo baliokidea bukatuta
- ☐ Batxilergoa, lanbide heziketa edo baliokidea bukatuta
- ☐ Unibertsitateko ikasketak bukatuta

Guraso horren generoa

- ☐ Emakumezkoa
- ☐ Gizonezkoa
- ☐ Ez bitarra

Adieraz ezazu ondorengo zein taldetan dagoen **guraso horren lanbidea**:

- ☐ Enpresariak (10 enplegatutik gora dutenak), enpresa zuzendariak, goi mailako funtzionarioak, irakasleak, armadako buru edo ofizialak, profesio liberal eta goi mailako teknikariak (abokatuak, medikuak, arkitektoak, psikologoak, informatikariak, botikariak, albaitariak...).
- ☐ Merkatari eta enpresari txikiak (5 eta 10 enplegatu bitartean dutenak), erdi mailako teknikariak (ingeniari teknikoak, aparejadoreak, OLTak...).
- ☐ Administrari eta komertzialak, armadako ofizialordeak, familia enpresariak, teknikari laguntzaileak, langileburuak eta tailer-buruak.
- ☐ Nekazaritza, industria edo zerbitzuetako langile kualifikatuak, merkataritzako langileak, gidariak, artisauak eta peoi espezialistak.
- ☐ Kualifikaziorik gabeko langile eta peoiak, etxeko langileak, atezainak, jornalariak.

Beste guraso baten **ikasketa maila**

- ☐ Lehen mailako hezkuntza, OHO edo baliokidea bukatuta
- ☐ Batxilergoa, lanbide heziketa edo baliokidea bukatuta
- ☐ Unibertsitateko ikasketak bukatuta

Guraso horren generoa

- ☐ Emakumezkoa
- ☐ Gizonezkoa
- ☐ Ez bitarra

Adieraz ezazu ondorengo zein taldetan dagoen **guraso horren lanbidea**:

- ☐ Enpresariak (10 enplegatutik gora dutenak), enpresa zuzendariak, goi mailako funtzionarioak, irakasleak, armadako buru edo ofizialak, profesio liberal eta goi mailako teknikariak (abokatuak, medikuak, arkitektoak, psikologoak, informatikariak, botikariak, albaitariak...).
- ☐ Merkatari eta enpresari txikiak (5 eta 10 enplegatu bitartean dutenak), erdi mailako teknikariak (ingeniari teknikoak, aparejadoreak, OLTak...).
- ☐ Administrari eta komertzialak, armadako ofizialordeak, familia enpresariak, teknikari laguntzaileak, langileburuak eta tailer-buruak.
- ☐ Nekazaritza, industria edo zerbitzuetako langile kualifikatuak, merkataritzako langileak, gidariak, artisauak eta peoi espezialistak.
- ☐ Kualifikaziorik gabeko langile eta peoiak, etxeko langileak, atezainak, jornalariak.

B.3 Irudia: *2016/17 eta 2017/18* ikasturteetako galdetegia.