EKAIA (2024), artikulua prentsan/article in press.
https://doi.org/10.1387/ekaia.26335
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255

Adimen Artifiziala. PTIA eredu kimioinformatikoa endekapenezko gaixotasun neurologikoen kontrako botikak aurkitzeko.

(Artificial Intelligence. PTIA cheminformatic model for finding drugs against degenerative neurological diseases.)

*Leire Llona[1], Ane Ibañez[1], Humberto González-Díaz[1,2,3], Harbil Bediaga[4,5], Sonia Arrasate[1]**

[1]Kimika Organikoa eta Ezorganikoa Saila, UPV/EHU, 48940, Leioa.

[2]Biofisika Institutua, CSIC-UPV/EHU, 48940, Leioa.

[3]IKERBASQUE, Basque Foundation for Science, 48011, Bilbao.

[4]IKERDATA S.L., ZITEK, UPV/EHU, Errektoretza Eraikina, 6, 48940 Leioa.

[5]Pintura saila, Arte Ederren Fakultatea, UPV/EHU, 48940, Leioa

LABURPENA: Kimioinformatika, kimika teorikoaren atal bat da, eta, zientzia kimiko farmazeutikoetako arazoetan teknika informatikoen erabileran datza. Lan honetan, IFPTIA (Informazioa Fusionatzea + Perturbazio Teoria + Ikasketa Automatikoa) metodologiak Kimika Medikoaren esparruan izan dezakeen aplikazioa aurkezten da. Zehazki, Alzheimer, Parkinson, Alboko Esklerosi Amiotrofikoa (AEA), Friedreich ataxia eta Huntington endekapenezko gaixotasun neurologikoak aztertu dira. Eraikitako eredu kimioinformatikoan, horiekin erlazioa izan ditzaketen proteinen sekuentziak eta garunaren eskualde desberdinen (kortex entorrinalaren (KE), hipokanpoaren (HIP), erdiko loki bihurgunearen (ELB), atzeko zingulu kortexaren (AZK), goi bekoki bihurgunearen (GBB) eta kortex bisualaren (KB)) Proteinen Elkarrekintza Sarea (PES) kontuan hartu dira. Lortutako eredu egokiena diskriminante lineala izanik, doitasun metrikak onak izan dira; Sn (%) = 77,76, Sp (%) = 72,69 eta Ac (%) = 73,83 entrenamendurako, eta Sp (%) = 72,66, Sn (%) = 77,95 eta Ac (%)= 73,84 balioen berrespenerako. Gaixotasun hauek sendaezinak dira eta sintomak agertzen diren momentutik aurrera, gaixoak ahultzen doaz, garunean dauden neuronak hil egiten diren arte. Honek dakarren arazo ohikoen artean mugimendu eta garun funtzionamendu okerren arazoak dira. Zentzu horretan, eredu kimioinformatikoak teknika guztiz erabilgarriak izan daitezke farmako berrien garapenean, entseguetarako animaliak murrizteko eta orotar, baliabideak saihesteko ahaleginetan. Ereduak, endekapenezko gaixotasun neurologikoetan aktiboa izateko edo ez izateko konposatu baten probabilitatea iragar dezake. Beraz, eredu hauek, gaixotasun hauen atzean dauden mekanismoak ulertzeko erreminta oso baliagarriak direla frogatu da eta horri esker, etorkizun handiko bide terapeutikoak zabaltzen dira.

HITZ GAKOAK: Ikasketa Automatikoa, Adimen Artifiziala, Kimioinformatika, Endekapenezko gaixotasun neurologikoak

ABSTRACT: *Cheminformatics is part of theoretical chemistry and consists of the use of computer techniques in pharmaceutical chemical science problems. This work presents the possible application of the IFPTIA methodology*

***Harremanetan jartzeko/ Corresponding author**: Sonia Arrasate, Kimika Organikoa eta Ezorganikoa Saila, Euskal Herriko Unibertsitatea (UPV/EHU), 48940, Leioa, Bizkaia. https://orcid.org/0003-2601-5959, sonia.arrasate@ehu.eus

Nola aipatu / How to cite: Llona, Leire; González-Díaz, Humberto; Bediaga, Harbil; Arrasate, Sonia (2024). << Adimen Artifiziala. PTIA eredu kimioinformatikoa endekapenezko gaixotasun neurologikoen kontrako botikak aurkitzeko >>, Ekaia, Ale berezia, xx-xx. (https://doi.org/10.1387/ekaia.26335)

Jasoa: maiatzak 17, 2024; Onartua: uztailak 16, 2024

ISSN 0214-9001-eISSN 2444-3225 / © 2024 UPV/EHU

(Information Fusion + Perturbation Theory + Machine Learning) in the field of medical chemistry. Specifically, Alzheimer's, Parkinson's, Amyotrophic Lateral Sclerosis (ALS), Friedreich ataxia and Huntington degenerative neurological diseases have been studied. In the predictive model, the sequences of proteins that may be related to them and the Protein Interaction Network (PES) of the different regions of the brain (entorrinal cortex (CR), hippocampus (HIP), central temple curve (CEP), rear cortex (CSC), upper forehead curve (BB) and visual cortex (CB) have been considered. The statistical parameters of the model obtained have been good; Sn (%) = 77.76, Sp (%) = 72.69 and Ac (%) = 73.83 for training; and Sp (%) = 72.66, Sn (%) = 77.95 and Ac (%) = 73.84 for validation. These diseases are incurable and from the moment the symptoms appear, the patients are weakened until the neurons in the brain die. Among the usual problems that this entails are problems of poor movement and brain functioning. In this sense, cheminformatic models can be a very useful technique in the development of new drugs, in efforts to reduce animals for testing and, in general, to avoid resources. They can predict the probability that a compound may or may not be active in degenerative neurological diseases. These models have therefore been shown to be very useful tools for understanding the mechanisms behind these diseases, and this has led to the opening of promising therapeutic pathways.

KEYWORDS: Machine Learning, Artificial Intelligence, Cheminformatics, Degenerative neurological diseases

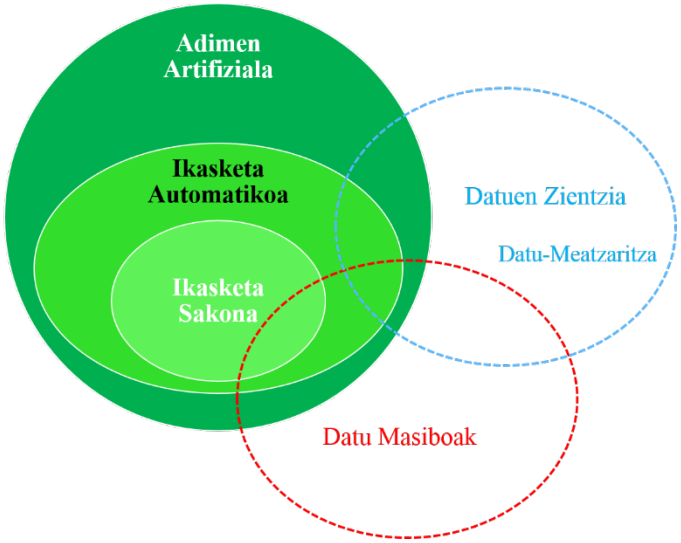
1. SARRERA

Artificial Intelligence (*AI*) edo Adimen Artifiziala (AA), makinek giza garunaren funtzioak simulatzean datza eta entitate adimendunak/makinak ulertzeaz eta eraikitzeaz arduratzen da. Makina horiek, egoera berri anitzen aurrean modu eraginkorrean eta seguruan jarduteko kalkuluak egin ditzakete [1]. Alan Turing-ek AAren inguruan ikerketa esanguratsua egin zuen lehena izan zen [2]. Adimen-muga oso zabala da; izan ere, AAk jakintza arlo asko barne hartzen ditu, hala nola, medikuntza, kimika, farmazia, fisika, biologia, bioteknologia, materialak, nanozientzia, eta abar.

Machine Learning (*ML*) edo Ikasketa Automatikoa (IA), AAren azpimultzotzat hartzen da (**1. irudia**). Makinek ematen diegun informaziotik ikastea nahi dugunean, IA teknikak erabiltzen dira. Makinek, gizakiek bezala ikas eta pentsa dezaten lortzeko dagoen moduetako bat da. Ikasketa prozesu hori, behaketa, proba eta errore bidez egiten da. Zeregin jakin eta mugatu bat ikasi ahal izateko adina datu ematen zaizkie, eta gero ezagutza hori datu berrietan aplikatzeko gai dira, denborarekin gehiago zuzenduz eta ikasiz [3–7]. Makina bati ikasten irakasteko modu asko daude: gainbegiratutako, gainbegiratu gabeko, erdi-ikuskatutako eta errefortzuzko ikaskuntza-teknikak. Gainbegiratutakoren kasuan, ikasten ari den bitartean algoritmoari etiketa edo sailkapen zuzena ematen zaio. Gainbegiratu gabekoan ez zaio emaitzarik ematen. Erdi-ikuskatutako teknikan emaitza batzuetan ematen zaio eta, errefortzuzko ikaskuntza teknika erabiltzen denean, puntuazioa desberdina ematen zaio ondo edo gaizki egitearen arabera. Era berean, hainbat arazotarako erabil daitezkeen algoritmoak daude: iragarpena, sailkapena, erregresioa, eta abar. Funtzioekin lortutako emaitzan datza diferentzia. Balio horrek erregresio-ereduetan bezala balio kuantitatibo bati edo sailkapen-ereduetan bezala talde bati esleitzea bezalako erantzun kualitatibo bati egin diezaioke erreferentzia. Zehazki, erregresio lineal sinpleak edo polinomikoak, euskarri bektore-makinak, erabaki-zuhaitzak, Random Forest metodologiak, K hurbileneko bizilagunak, adibidez, IAn erabiltzen diren algoritmo arruntetako batzuk dira, baina askoz gehiago ere badaude [8–13]. Horietan guztietan aipagarria da, aldez aurretik, oso garrantzitsua dela datuak lortzen eta kargatzen ikastea, esate baterako horien esplorazio-analisia egitea eta informazioa garbitzea; ikaskuntzaren kalitatea datuen kalitatearen araberakoa izango baita.

IAaren barruan *Deep Learning* (*DL*) edo Ikasketa Sakona (IS) deritzogun adarra dago. Bere teknikak Neurona-Sare Artifizialak (NSA) erabiltzean oinarritzen dira. Teknika horietan asko aurreratu da 2010. urtetik hona eta, egun, sakonera neurona-geruza ugariko sareak sortzeko gai dira. Azken urteotan, ikasketa sakona aplikatu izan da, besteak beste, botiken diseinuan, bioinformatikan, irudi medikoen analisian, materialetan, ahotsaren ezagutzan eta hizkuntzaren prozesaketan. Lerro hauek

irakurtzerakoan erraz bururatzen zaizkigu Netflixen gomendioak, ahots-laguntzaileak (Alexa, Siri edo Googleren laguntzailea) edo aurpegia aldatzen duten mugikorraren aplikazioak; horiek guztiek IS erabiltzen dute [14–17].



1. **irudia.** AAk barne hartzen dituen ezagutza arlo adierazgarrienak.

Big Datak (*BD*) edo Datu Masiboak (DM), datuak eskala handian eta sistematikoki atzemateko, biltegiratzeko, homogeneizatzeko, transferitzeko, kontsultatzeko, bistaratzeko eta aztertzeko beharrezkoak diren teknikak biltzen ditu. DMaren ezaugarri nagusia ingelesetik datozen lehengo 3V dira (*Volum, Velocity, Variety*); egun, 5V+C (*Volum, Velocity, Variety, Value, Veracity and Complexity*) dira [1]. Zentzu horretan, ikerketa lan honen datu sorta kuasi-DM kontsidera daiteke. Tamainari dagokionez, ez luke DM gisa sailkatuko, baina konplexutasunagatik bai. Bestalde, datuen zientzia (DZ), datu-multzoetatik informazio garrantzitsua ateratzeaz arduratzen da. Era berean, ezagutza ateratzeko datuen tratamendua inplikatzen duten metodoei, prozesuei eta sistemei ere erreferentzia egin diezaioke. Besteak beste, teknika estatistikoak eta datu-analisiak erabil daitezke, baita beren kabuz ikasten duten eredu adimendunak ere (gainbegiratu gabeak). Beraz, IAren zati bat ukituko luke. Izan ere, termino hori datu-meatzaritzarekin edo IArekin berarekin nahas daiteke.

Gure ikerkuntza taldean Kimioinformatikan dihardugu. Hala, medikamentu batek gaixotasun zehatz batean eduki dezakeen aktibitatea, nanopartikula batek medikamentu garraiatzaile sistema batean duen aktibitatea edota erreakzio organiko baten etekina edo enantiomero-gaindikina aurresan ditzaketen

eredu kimioinformatikoak garatzen ditugu [18,19]. Kimioinformatika, kimika teorikoaren atal bat da, eta, zientzia kimiko farmazeutikoetako arazoetan teknika informatikoen erabileran datza [20]. *In silico* teknika horiek farmakoak aurkitzeko prozesuan erabiltzen dira farmazia-enpresetan. Ez da kimika konputazionalarekin nahastu behar, azken hau, kimikaren adar bat da (**2. irudia**) eta simulazio informatikoa erabiltzen du arazo kimiko konplexuak konpontzen laguntzeko [21]. Kimika teorikoko metodoak ustiatzen ditu, programa informatiko eraginkorretan txertatuak, molekulen egiturak, interakzioak eta propietateak kalkulatzeko [22–26]. Kimiometria sistema kimikoetatik informazioa ateratzea helburu duen diziplina kimikoa da eta behaketa esperimentaletatik datozen datuen tratamendu estatistikoaz eta matematikoaz arduratzen da. Kimika teorikoaren barruan ere, molekulen modelatze molekularra dago, eta horrek, molekula kimikoen eta biologikoen 3-D egituraren sorrera, irudikapena edota manipulazioa deskribatzen du. Kimika matematikoa, bestalde, fenomeno kimikoen eredu matematikoarekin zerikusia du. Kimika matematikoari kimika informatikoa ere deitu izan zaio batzuetan, baina ez da kimika konputazionalarekin nahastu behar [27,28]. Azkenik, atomoen eta molekulen mugimendu fisikoak aztertzeko dinamika molekularra ordenagailuen bidez egiten den simulazio-metodoa da (**2. irudia**).

2. **irudia.** Kimioinformatika kimika teorikoaren atala da.

Eredu kimioinformatikoek teknika informatikoak eta estatistikoak integratzen dituzte aktibitate biologiko zehatzen iragarpen teorikoa egiteko eta balizko farmako berrien diseinu teorikoa ahalbidetzeko. Hala, sintesi organikoaren proba eta errore prozesua adibidez, saihesten da. Ingurune

birtual batean bakarrik dagoen zientzia denez, zenbait baliabide alde batera uzteko aukera ematen du, hala nola ekipoak, tresnak, materialak eta laborategiko langileak. Egitura kimikoaren eta aktibitate biologikoen arteko erlazioak ezartzen direnez, farmako berrietarako hautagaien diseinua askoz azkarragoa eta merkeagoa suertatzen da. Farmakoen aurkikuntzan oinarritutako metodologia kimioinformatikoa diziplinartekoa da; eta, beraz, Kimika Organikotik eta Farmakologiatik jasotzen du informazioa. Tresna konputazionalen bidez egindako simulazio molekularrak eskatzen duen denbora, konposatu berrien sintesiak eta biosaiakuntzek daramaten denbora baino askoz txikiagoa da, azken hauek hilabeteak edo urteak ere izan baitaitezke. Abantaila horri esker, molekula segida batzuk har daitezke, eta emaitzak nahiko azkar lortzen direnez, sintesi-laborategira zuzenean pasa daitezke sintesia gidatzeko. Hala, eredu kimioinformatikoek konposatu berriak aurreikusten dituzte eta sintetiza ditzaten kimikari organikoei proposatzen dizkiete; eta, ondoren farmakologoei, biosaiakuntzetara eraman ditzaten. Hortik, eredu kimioinformatikoak aurreikusitako balioak berresten edo ezeztatzen dituzten emaitzak lortzen dira. Kasu optimo batean, ziklo operatibo honen bidez proba eta errore hutsez baino hautagai hobeak lortzen dira. Horrek denbora, dirua eta baliabideak aurrezten ditu, eta farmako berriak garatzen dituztenen porrota saihesten du.

Farmako bat itu terapeutikoarekin edo itu molekularrarekin elkarreragiten duen molekula gisa defini daiteke, haren portaera nolabait aldatuz. Farmako ezagunek diana ezagunetan eragiten dute, baina gaixotasun baten bilakaera alda dezaketen edo dauden tratamenduen eraginkortasuna hobetu dezaketen farmako berriak aurkitzea da kimikaren eta biologiaren arloko ikerketaren helburu nagusietako bat. Farmako berri baten garapenak 12 urte arte iraun dezake, eta batez besteko kostua, merkatura iritsi arte, mila milioi euro ingurukoa dela kalkulatzen da. Parte hartzen duten denbora eta kostuak, hein handi batean, beren garapenaren etapa batean edo gehiagotan huts egiten duten molekulen kopuru handiarekin lotuta daude. Izan ere, 5.000 sendagaietatik 1 bakarrik iristen da azkenean merkatura. Aurreko estatistikek erakusten dutenez, farmako berriak aurkitzea eta garatzea oso prozesu konplexua eta garestia da. Prozesu hori denbora luzez egin da soilik metodo esperimentalak erabiliz. Azken hamarkadetako aurrerapen teknologikoek *in silico* terminoaren sorrera sustatu dute. Termino horrek birtualki gauzatzen diren prozesu biologikoen simulazio informatikoen bidezko prozesuei erreferentzia egiten dio. Ez da organismo bizi baten gainean (*in vivo* esperimentuak) zuzenean egiten den esperimentua, ezta entsegu-hodi batean edo organismotik kanpoko beste ingurune artifizial batean (*in vitro* izeneko esperimentuak) egiten dena. Hala, konputazio-tresna horiek ezinbestekoak dira esperimentazio biologikorako; izan ere, eredu teorikoak zehaztasun handiz kodetzeko aukera ematen dute, eta informazio-kopuru handiak prozesatzeko gai dira, farmako berrien garapen-prozesua erraztuz eta azkartuz.

Oro har, eredu kimioinformatikoak aplikazio txikiko domeinu batekin sortzen dira, soilik baldintza multzo batean zentratuta, adibidez, propietate edo aktibitate biologiko espezifiko batean, proteina objektibo batean edo lerro zelular batean. Baina oso interesgarria da aldi berean saiakuntza-baldintza ugari kontuan hartu ahal izatea, eta hori, perturbazio teoriaren bidez errazten da. Hala, gure ikerkuntza taldean, IFPTIA metodologia garatu dugu, zeinak, Informazioa Fusionatzea + Perturbazio Teoria + IA adierazten duen [29,30]. Metodologia hori erreferentzia-funtzio batekin hasten da, farmako bat baldintza jakin batzuetan aktiboa izateko probabilitatea neurtzen duena. Horretaz gain, PT Operadoreak (PTO) erabiltzen ditu farmako horren, baldintza berberetan entseatutako farmakoen populazio batekiko, sarrera-aldagaien desbiderapenen (perturbazioen) berri emateko. IFPTIA metodoa hainbat jakintza arlotan aplika daiteke. Alde batetik, medikamentuak askatzeko nanopartikula sistemetan aplikatu da, hots, medikamentu/estaldura-agentea/metal edo metal oxido nanopartikula sistema egokiak hautatzeko lehen PTIA etiketa anitzeko eredua garatu da [18]. Bestetik, gaixotasun kardiobaskularrak edo neurodegeneratiboak tratatzeko kalmodulinaren inhibitzaileen medikamentuak aurkitzeko eredu baliagarria sortu da [31]. Era berean, Flaviviridae familiako gaixotasunen aurkako tratamendu berriak garatzeko sendagaiak aurkitzeko ahaleginetan lagungarria izan daitekeen eredua lortu da [32]. Kimika organiko arloari dagokionez, molekulen arteko α-amidoalkilazio erreakzioen enantioselektibitatea optimizatzeko eredu kimioinformatikoa sortu da eta web zerbitzari batean ezarri da. Kimikari organiko esperimentalentzako eskuragarri dagoen tresna konputazional berri honek, erreakzio-baldintzen optimizazio jasangarria, katalizatzaile berrien diseinua eta produktu berrien sintesia ahalbidetzen ditu[19].

Lan honetan, IFPTIA metodologiak kimika mediko esparruan izan dezakeen aplikazioa aurkezten da. Zehazki, endekapenezko gaixotasun neurologikoetara zuzenduta dago. Gaixotasun hauek sendaezinak dira eta sintomak agertzen diren momentutik aurrera, gaixoak ahultzen doaz, garunean dauden neuronak hil egiten diren arte. Horrek dakarren arazo ohikoen artean mugimendu eta garun funtzionamendu okerren arazoak dira. Zentzu horretan, eredu kimioinformatikoak teknika guztiz erabilgarriak izan daitezke farmako berrien garapenean, entseguetarako animaliak murrizteko eta orohar, baliabideak saihesteko ahaleginetan.

2. METODOLOGIA

Corwin Herman Hansh, ikerlaria da ordenagailuz lagundutako diseinu molekularraren aitzindaria, *Quantitative Structure-Activity Relationship* (*QSAR*) edo Egitura Aktibitate Erlazio Kuantitatiboa (EAEK) sortzeagatik ezaguna. EAEKak erantzun kimiko baten eta analizatutako molekulen

berezitasunak definitzen dituzten ezaugarri kimiko kuantitatiboen artean erlazio matematiko bat ezartzen du. Erlazio matematiko hori, egitura-erlazio bat duten konposatuen artean ezartzen da. Ikerketa horrek, berez, erlazio matematiko bat ezartzea du helburu, produktu kimiko baten jokabidearen, hau da, erantzun kimikoaren, eta bide esperimentaletatik zein teorikoetatik lor daitezkeen ezaugarri kimiko kuantitatibo multzoen artean. Ikerketaren izena, ereduaren erantzunaren izaeraren araberakoa da, hots, egitura-propietate/aktibitate/toxizitate (EPEK / EAEK / ETEK) erlazioen ikerketak ezagutzen dira nagusiki, non propietate fisiko-kimikoak, aktibitate biologikoak eta datu toxikologikoak erabiltzen diren. Erlazio hori bilatzeko funtsezkoak dira deskriptore molekularrak, egiturak definitzen baitituzte, (1) Ekuazioa [33].

$$\text{Erantzuna} = \int (\text{egitura kimikoa}, \text{propietate fisikokimikoak}) \quad (1)$$

Erantzun zehatz baterako, aktibitate biologiko zehatz baterako adibidez, EAEK ekuazioa honako modu honetan defini daiteke matematikoki, (2) Ekuazioa.

$$Y_i = w_0 + w_1X_1 + w_2X_2 + w_3X_3 \ldots + w_nX_n \quad (2)$$

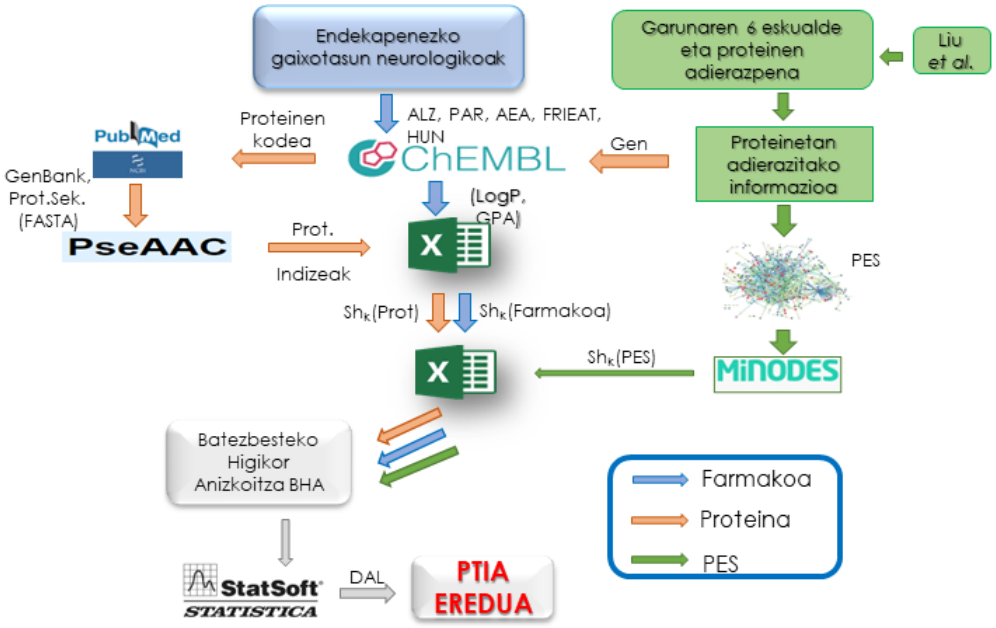
Y_i menpeko aldagaia izango da eta ereduak emango duen erantzuna adierazten du, adibidez, aktibitatea/propietatea/toxizitatea. Aldagai independenteak, X_1, $X_2 \ldots$ X_n dira eta egitura ezaugarriak zein propietate fisiko-kimikoak adierazten dituzte, baina zenbakizko kantitateetan, hau da, deskriptoreetan. Deskriptore bakoitzaren, w_1, $w_2 \ldots$ w_n, banakako ekarpena da erantzunean eta a_0 erroreari dagokion konstantea [33]. Matematikoki w_1, $w_2 \ldots$ w_n, erregresio koefiziente horiek entrenatze-datu sorta batetik kalkulatzen dira, non aldagai independenteak eta menpekoak ezagunak diren. Adierazpen horrek, sistema kimiko/biologiko batean eredu kimioinformatikoaren erantzunak izaera lineala duenean balio du [34].

Oro har, metodo klasikoek ez dute kontuan hartzen entsegu aurre klinikoetan agertzen diren datu sorta handiak edo aldi bereko hainbat propietate biologiko [35–37]. Arazo honi aurre egiteko, ohikoak ez diren metodoak erabili behar dira eta hor, perturbazioaren teoria kontuan hartu behar da [38]. Perturbazio-teoriaren bidez, perturbazio txikiek molekulen egituran eta propietateetan dituzten ondorioak azter daitezke. Beraz, jatorrizko sistemaren antzekoa den beste sistema ezezagun baten propietateak aurresan daitezke. Horrela, PTIA ereduak botika bat esperantzagarria izan daitekeen edo ez adierazten duen probabilitatea aurresan dezake. Hala, konposatu kimiko bat baldintza jakin batzuetan aktiboa izateko probabilitatea neurtzen duen erreferentzia-funtzioa eta PTOak erabiltzen ditu; non,

operadore horiek, baldintza berberetan entseatutako farmakoen populazio batekiko, farmako horren sarrera-aldagaien desbideratzeen (perturbazioen) berri ematen duten [29].

2.1. Datu bilketa, garbiketa eta tratamendua

Ikerkuntza lan honetan, endekapenezko honako gaixotasun neurologikoak aztertu ditugu: Alzheimer, Parkinson, Alboko Esklerosi Amiotrofikoa (AEA), Friedreich ataxia eta Huntington gaixotasunak. Lehenengo, Liu *et al*. artikuluan oinarrituta [39], endekapenezko gaixotasun neurologiko horiekin erlazioa izan dezaketen proteinen sekuentziak eta garunaren eskualde desberdinen (kortex entorrinalaren (KE), hipokanpoaren (HIP), erdiko loki bihurgunearen (ELB), atzeko zingulu kortexaren (AZK), goi bekoki bihurgunearen (GBB) eta kortex bisualaren (KB)) *Protein Interaction Network*, (*PIN)* edo Proteinen Elkarrekintza Sarea (PES) lortu dira [39]. Ondoren, proteina horiek ChEMBLan bilatu dira eta web gune horretatik ere, endekapenezko gaixotasun neurologiko desberdinen inguruan, datu sorta esanguratsua izateko hainbat zientzialarik burututako klinika aurreko esperimentuen datuak lortu dira [40]. Garunaren sei eskualdeak kontuan hartzeko, bildutako kasu bakoitza bider 6 egin da. Behin datu basean, gutxi gorabehera, 790000 sarrera inguruko informazioa bilduta dagoela, aktibitate biologiko desberdin asko agertu dira. Hala nola, IC_{50} (nM) propietateak, inhibitzailearen kontzentrazio maximoaren erdiari egiten dio erreferentzia, hau da, farmako baten kontzentrazioa, inhibizioaren %50 emateko; potentzia (nM) edo efektu zehatz bat lortzeko behar den konposatuaren kontzentrazioa; EC_{50} (nM) edo kasuen %50ak efektuaren eragin maximoa lortzeko behar den kontzentrazioa; aktibitatea (%) edo farmako batek, materia biziaren gainean duen efektu onuragarrien zein txarren neurketa, besteak beste. Azkenik, datu multzoa behin eraikita, botiken, proteina sekuentzien eta proteinen elkarrekintza sarearen egiturazko aldagaiak, eta entseguen baldintza desberdinak kontuan hartu dira eredua garatzeko **3. irudian** ikusten den moduan.



3.irudia. Eredu kimioinformatikoa sortzeko jarraitutako pausuak.

Konposatu kimikoen egitura-deskriptoreak kalkulatzeko, D_k(farmako$_i$), bi deskriptore erabili dira. Alde batetik, D_1= ALogP, *n*-oktanol eta uraren arteko banaketa koefizientearen logaritmoa da eta molekula baten lipofilitatearen deskriptore bezala onartuta dago. Bestalde, *Polar Surface Area* (*PSA*), edo Gainazal Polarraren Azalera Topologikoa (GPAT), D_2= GPAT, nitrogeno eta oxigeno atomoek eta horiei lotutako hidrogeno polarrek betetzen duten azalera gisa definitzen da eta hidrogeno lotura horien kapazitatea eta polaritatea islatu egiten ditu [41] [42].

Proteinen egitura-deskriptoreak D_k(proteina$_i$) lortzeko, FASTA kodeak, testu formatuan aminoazidoen (AA) edo azido nukleikoen sekuentzia gorde dezakeen kodea lortu eta horiek PseAAC, *Generating pseudo Amino Acid Composition* [43][44] [45] [46], Chou *et al*. ikerlariek garatu duten zerbitzarian sartu dira. PseAACzerbitzaria: http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/.

Proteina-proteina elkarrekintza sarearen deskriptoreak lortzeko, D_k(PES$_i$), MI-NODES softwarea erabili da [47]. Horren bidez garunaren eskualde desberdinetan k distantziara dauden proteina desberdinen elkarrekintzak kalkula daitezke.

Behin egitura aldagaien deskriptoreak kalkulatuta daudela, lortutako informazioa eskala berdinera eraman behar da eta horretarako Shannonen entropia, edo informazioaren entropia, erabili dugu. Horrek,

datu estokastikoko iturri batean dagoen informazioaren ziurgabetasuna ematen du, batezbesteko indizea emanez. Datuen balio posible bakoitzaren entropiaren neurketa, balioaren masa-probabilitate funtzioaren logaritmo negatibo bezala definitzen da (3. Ekuazioa) [48].

$$Sh_k(sistema_i) = -[p(D_{ki})] \cdot \log_2 [p(D_{ki})] \qquad (3)$$

Egiturazko aldagaiak finkatu ondoren, zenbakizko balioak ez diren muga baldintzak, $\mathbf{c}_j$, hala nola, aktibitatea, organismoa, entseguko organismoa eta abar, definitu behar dira eta horiei zenbakizko balio bat egokitu. Horretarako, *Moving Average* (*MA*) edo Batezbesteko Higikor (BH) den aldagai berria, sartuko da (4.ekuazioa).

$$\Delta Sh_k\left(sistema_i, \boldsymbol{c}_j\right) = \left(Sh_k(sistema_i)_{berria} - \langle Sh_k(\boldsymbol{c}_j)\rangle\right) \quad (4)$$

$\langle Sh_k(\boldsymbol{c}_j)\rangle$ parametroak, $\mathbf{c}_j$ muga baldintza konkretu batzuetarako, egiturazko aldagaien entropia balioen batezbestekoa ematen du. $Sh_k(sistema_i)_{berria}$ aldez aurretik kalkulatuta dagoen egiturazko aldagaien entropia da. Azkenean, BH-ak adierazten du, egiturazko aldagai baten balioa, $\mathbf{c}_j$ espezifiko batzuetarako zenbat aldentzen den aldagai hori osatzen duten balioen batezbestekotik.

Horretarako, aktibitate biologikoa adierazten duten c_0 baldintzak desiragarritasunaren arabera sailkatuko dira, desiragarri eta ez desiragarri moduan, d(c_0) lortuz. Ondoren, propietate horiei *cut-off* edo mozketa mugak aplikatuko zaizkie. Kontzentrazio unitateak dituzten aktibitateei, K_i eta IC_{50}, adibidez, ez desiragarri edo minimizatu nahi den izena jarriko zaie, izan ere, kontzentrazio txikietan farmakoa aktiboa bada, hobea baita. Ehuneko, eta abiadura unitateak dituzten aktibitateei, aldiz, desiragarri, eraginkortasuna (%) eta farmako baten abiadura handiagoa bada, aktibitate hori positiboagoa (desiragarria) izango delako.

Hurrengo pausoan, $f\left(v_{ij}\right)_{helb}$ funtzioa kalkulatu dugu, helburu-funtzioa izango dena, hots, programa estatistikoari irteerako aldagai bezala ematen zaion funtzioa. Horrela, propietatearen d(c_0) desiragarria bada, eta $v_{ij} > cut\ off$ bada, $f\left(v_{ij}\right)_{helb} = 1$ balioa hartuko du; d(c0) ez desiragarria bada eta $v_{ij} < cut\ off$ bada, $f\left(v_{ij}\right)_{helb} = 1$ izango da; eta aurreko bi baldintzak ez gertatzekotan, $f\left(v_{ij}\right)_{helb} = 0$ izango da. Beste aldetik, probabilitateetan oinarritzen den erreferentziazko funtzioa definitu behar da, $f\left(v_{ij}\right)_{erref}$. Beraz, c_0 propietate eta a azpimaila bakoitzerako dauden kasu guztietatik, aktiboak (1 balioa) izateko probabilitatea adierazten du, (5) eta (6) Ekuazioetan adierazten den bezala.

$$f(v_{ij})_{erref} = p(f(v_{ij} = 1), c_0) \quad (5)$$

$$p(f(v_{ij} = 1), c_0) = n(f(v_{ij} = 1, c_0)/n_{c_0,a} \quad (6)$$

Eredua eraikitzeko, sistemarentzat ezaguna den aldagaia, $f(v_{ij})_{erref}$ eta horri gehituko zaizkion PT operadoreak, ΔSh_k, eta deskriptoreak, D_k erabiliko dira sarrerako aldagai gisa, eta $f(v_{ij})_{helb}$ irteerako aldagai bezala. PTIA eredua sortzeko, *Linear Discriminant Analysis (LDA)*, Diskriminazioan oinarritutako Analisi Lineala (DAL), teknika estatistikoa erabili da STATISTICA [49] softwarea erabiliz.

Ereduak itzuliko digun funtzioa $f(v_{ij})_{kalk}$ izango da eta STATISTICAk Mahalanobisen distantzia aplikatuz, $f(v_{ij})_{aurre}$ probabilitate-funtziora eraldatzen du. Horrela, 0 edo ez aktiboa izateko probabilitatea, eta, 1 edo aktibo izateko probabilitatea, taldeak bereiz daitezke. Horrek, etorkizuneko iragarpenak egiteko aukera ematen digu, edota konposatu berrien diseinuak garatzeko aukera.

3. EMAITZAK ETA EZTABAIDA

Gaur egun ez dago eskuragarri endekapenezko gaixotasun neurologikoetara zuzendutako farmakoen egitura, proteinen sekuentzia, garunaren eskualde desberdineko proteinen elkarrekintza sarea (PES) eta entseguen baldintzak kontuan hartzen dituen eredu kimioinformatikorik. Ikerkuntza lan honetan proposaturiko eredua, honako (7) Ekuazio honetan ikus daiteke. PTIA sailkapen eredua sortzeko, sarrerako aldagai gisa ezaguna den erreferentziazko funtzioa $f(v_{ij})_{erref}$ erabili da eta horri gehitutako PT operadoreak, ΔSh_k, deskriptoreak, Sh_k, eta horiei esleitutako koefizienteak.

$$\begin{aligned} f(v_{ij})_{kalk} = {} & -3{,}29182464 + 9{,}35519136 \cdot f(v_{ij})_{erref} - 0{,}01770167 \cdot \mathrm{Sh}_1(\mathrm{farm}_i) \\ & - 20{,}68241027 \cdot \mathrm{Sh}_2(\mathrm{farm}_i) - 13{,}40396847 \cdot \mathrm{Sh}_5(\mathrm{prot}_j) \\ & + 23{,}34368413 \cdot \mathrm{Sh}_5(\mathrm{PES}_j) + 43{,}91215676 \cdot \Delta\mathrm{Sh}_2(\mathrm{farm}_i, \mathbf{c}_j) \\ & - 3{,}32508118 \cdot \Delta\mathrm{Sh}_5(\mathrm{prot}_j, \mathbf{c}_j) + 2{,}34642134 \cdot \Delta\mathrm{Sh}_5(\mathrm{PES}_j, \mathbf{c}_j) \end{aligned} \quad (7)$$

$$n = 745674 \qquad \chi^2 = 138471{,}8 \qquad p < 0{,}05$$

Ereduak aurresandako $f(v_{ij})_{kalk}$ funtzioan agertzen den lehenengo balioa, errorea da. Ondoren, aktibitate biologikoari dagokion erreferentziazko funtzioa $f(v_{ij})_{erref}$ eta sistemari gehitzen edo kentzen zaizkion deskriptoreen eta perturbazioen efektuak ageri dira. Hortaz, eredu horrek hiru sarrera aldagai desberdin dauzka $f(v_{ij})_{erref}$, deskriptoreak, Sh_k, eta perturbazio operadoreak, ΔSh_k, $\mathbf{c}_j$ muga baldintzen menpe. Alde batetik, $f(v_{ij})_{erref}$ sarrera aldagaia, espero den aktibitate biologikoaren balioa ordezkatzen du c_0 baldintzarako. Bestetik, deskriptoreak daude, farmakoen, proteinen eta PESaren hiru egitura aldagaien informazio kimikoa. Amaitzeko, egitura aldagaien/sistemen BH dago. Adierazpen hori, ΔSh_k, aztertutako aldagai/sistemaren balioa, $Sh_k(sistema_i)_{berria}$, sistema hori osatzen duten balio guztien batezbestekotik, $\langle Sh_k(\boldsymbol{c}_j)\rangle$, zenbat desbideratzen den kalkulatzen du, $\mathbf{c}_j$ muga baldintzetarako. Honako taula honetan (**1. Taulan**) ereduan sartzen diren aldagaiak zehazten dira.

1. Taula Eredua sortzeko erabili diren zortzi sarrera aldagaiak.

Operadorea	**Formula**	**Operadorearen informazioa**
$f(v_{ij})_{erref}$	$n(f(v_{ij}=1,c_0)/n_{c_0,a}$	Konposatu bat aktiboa izateko aldez aurreko probabilitatea da, $p(f(v_{ij})=1,c_0)$, c_0 propietate zehatz baterako.
$\mathrm{Sh}_1(\mathrm{farm}_i)$	$Sh_1(farm_i) = -p(D_{1i}) \cdot \log(p(D_{1i}))$	D_1=ALogP. Lipofilitate terminoetan, konposatu baten egitura kimikoaren informazioaren aldakortasuna azaltzen du. Informazioa Shannon entropiara eraldatuta dago.
$\mathrm{Sh}_2(\mathrm{farm}_i)$	$Sh_2(farm_i) = -p(D_{2i}) \cdot \log(p(D_{2i}))$	$D_{2=}$ GPAT. Konposatu batean azaldutako gainazal azalera terminoetan, egitura kimikoaren informazioaren aldakortasuna azaltzen du. Informazioa, Shannon entropiara eraldatuta dago.
$\mathrm{Sh}_5(\mathrm{prot}_j)$	$Sh_5(\mathrm{prot}_j) = -p(D_{1j}) \cdot \log\left(p(D_{1j})\right)$	Proteina baten AA sekuentziaren ordena kuantifikatuta dago eta informazio hori Shannon entropiara eraldatuta. Sh_5 izena du, informazioa sekuentziaren 5. AA auzokideraino kuantifikatu delako.
$\mathrm{Sh}_5(\mathrm{PES}_j)$	$Sh_k(PES) = -\sum_j^n [p_k(Prot_j)] \cdot \log[p_k(Prot_j)]$	Proteinen elkarrekintza sarean, distantzia topologiko batera dauden 5 proteina hartzen dira eta haien interakzioak kuantifikatzendira.

$\Delta \mathrm{Sh}_2(\mathrm{farm}_i, \mathbf{c}_j)$	$\left(Sh_2(farm_i)_{berria} - \langle Sh_2(\boldsymbol{c}_j)\rangle\right)$	PT operadorea da eta berez, konposatu baten, GPAT ($Sh_2(farmako_i)_{berria}$) balioa, batezbestekotik, $\langle Sh_2(\boldsymbol{c}_j)\rangle$, zenbat desbideratzen den kalkulatzen du, $\mathbf{c}_j$ baldintzetarako.
$\Delta \mathrm{Sh}_5(\mathrm{prot}_j, \mathbf{c}_j)$	$\left(Sh_5(prot_j)_{berria} - \langle Sh_5(\boldsymbol{c}_j)\rangle\right)$	PT operadorea da eta berez, konposatu baten $Sh_5(proteina_i)_{berria}$ balioa, batezbestekotik, $\langle Sh_5(\boldsymbol{c}_j)\rangle$, zenbat desbideratzen den kalkulatzen du, $\mathbf{c}_j$ baldintzetarako.
$\Delta \mathrm{Sh}_5(\mathrm{PES}_j, \mathbf{c}_j)$	$\left(Sh_5(PES_j)_{berria} - \langle Sh_5(\boldsymbol{c}_j)\rangle\right)$	PT operadorea da eta berez, konposatu baten $Sh_5(PES_i)_{berria}$ balioa, batezbestekotik, $\langle Sh_5(\boldsymbol{c}_j)\rangle$, zenbat desbideratzen den kalkulatzen du $\mathbf{c}_j$ baldintzetarako

Aurkeztutako ereduaren doitasun metrikak, honako taula honetan (**2. Taulan)** azaltzen dira.

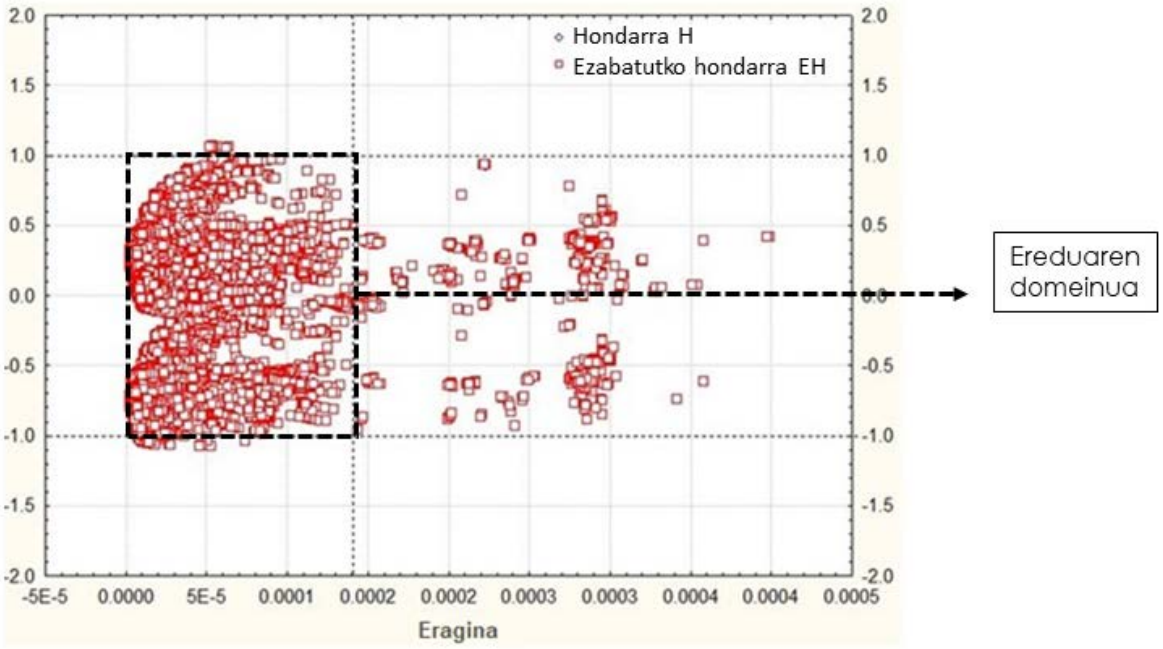
2. Taula Aurkeztutako eredu lineal diskriminantearen sailkapen matrizea.

Ikusitako multzoak	**Doitasun Metrikak[a]**	**Aurres. Param.**	**Aurresandako sorta**		
			n_j	$f(v_{ij})_{aurr} = 0$	$f(v_{ij})_{aurr} = 1$
Entrenamendu sorta					
$f(v_{ij})_{helb} = 0$	Sp (%)	72,69	433922	315407	118515
$f(v_{ij})_{helb} = 1$	Sn (%)	77,76	125357	27874	97483
Guztira	Ac (%)	73,83	559279	343281	215998
Berrespen sorta					
$f(v_{ij})_{helb} = 0$	Sp (%)	72,66	144665	105101	39564
$f(v_{ij})_{helb} = 1$	Sn (%)	77,95	41730	9201	32529

Guztira	Ac (%)	73,84	186395	114302	72093

[a] Sn(%) = Sentikortasuna, Sp(%) = Espezifikotasuna, eta Ac(%) = Zehaztasuna

Artikulu honetan aurkezten den PTIA ereduaren doitasun metrikak onak dira, lortutako sailkapen matrizean ikus daitekeen moduan, bai entrenatze eta berrespen prozesuetan ere. Entrenatze-sortari dagokionez, aurresandako doitasun metrikak honako hauek dira: sentikortasuna, Sn (%) = 77,76, hau da, multzoaren barruan zenbat kasu aurresaten diren positibo edo 1 gisa, zati, hasierako datu sortan, 1 ziren kasu totalak; espezifikotasuna, Sp (%) = 72,69; hots, 0 multzoaren barruan, negatibo edo 0 gisa aurresan direnak, zati, hasierako datu sortan 0 ziren kasu totalak eta zehaztasuna, Ac (%) = 73,83, asmatutako iragarpen kasuak zati datu base osoaren kasuak. Garrantzitsua da bai entrenamenduaren eta berrespen multzoekin lortutako doitasun metriken balioak antzekoak izatea. Honekin, ereduak aurresateko duen eta aurresandako balioak kalkulatzeko gaitasuna mantentzen dela ziurtatzeko. Datu horiek egokiak dira guztiak %70 baliotik gora baitaude, beraz, taldeak behar bezain banatuta daude. Hala ere, eredua balioztatzeko kanpo berrespena egin behar da. Berrespenerako prozesuan metrikak ere, %70 baliotik gora agertzen dira. Hala, Sp (%) = 72,66, Sn (%) = 77,95 eta Ac (%)= 73,84 balioak lortu dira hurrenez hurren. Eredu hau egiteko, datu sortaren %93.9a erabili da; beraz, *outliers* edo bazterrako balioak, puntu susmagarriak izan daitezkeenak, eliminatu dira. Jarritako lehen probabilitateak, 0 izateko, $p(f(v_{ij})=0)$ = 0,55 eta 1 izateko, $p(f(v_{ij})=1)$ = 0,45 izan dira. Horretaz gain, puntu susmagarriak eliminatzeko ($f(v_{ij})_{helb} = 0$ eta $p(f(v_{ij})= 1)_{kalk} > 0,9$, goi-p) edo ($f(v_{ij})_{helb} = 1$ eta $p(f(v_{ij})= 1)_{kalk} < 0,45$, behe-p) baldintzak kontuan izan dira. Hala, ereduaren eremua definitu da **4. irudian** ikus daitekeen moduan.



4. irudia. Ereduak duen domeinuaren egokigarritasun mapa.

EAEK eredu baten domeinuaren egokigarritasuna (DE) garatu dugun ereduaren entrenamendu sortaren gainean egindako espazio fisiko kimikoa ezagutzeko egiten da. Hala, lortzen den eredu kimioinformatikoa baliagarria da konposatu berri desberdinen iragarpenak egiteko aplikagarritasun eremua zein den definitzeko. Prozesu hori puntu susmagarrien ezabaketan edo identifikazioan ere aplika daiteke [50]. EAEK ereduen DE definitzeko *leverage* edo eragin parametroa kontuan hartzen dituzten ikuspegiak erabiltzen dira [51].

Y ardatzaren ezkerraldean, hondarraren (Hren) ardatza dago eta horri dagozkion balioak laburpen-mapan, kolore urdinarekin azalduta daude. Hondarrak, balio esperimentalaren eta aurresandakoaren arteko diferentzia adierazten du [52]. Y ardatzaren eskuinaldean ezabatutako hondarraren (EHren) ardatza dago eta, berez, hondarraren balioak adierazten ditu, baina *cross validation* edo balioztatze gurutzatua berrespen teknika aplikatu ostean; bere balioak, laburpen-mapan, kolore gorriarekin agertzen dira [53]. Beste aldetik, X ardatzean, *leverage* edo eragin parametroa dago [54].

Datuen murrizketa hori egin baino lehen, mozketa balioak aldatu izan dira, baina, aldaketek eredua hobetzen ez dutenez, doitasun metriken ehuneko altuena duen eredua aukeratu da eta horri puntu susmagarrien froga aplikatu zaio. Ezarritako mozketa balioak honako hauek izan dira: kontzentrazio

unitateei 75, ehunekotan dauden propietateei (%) eta abiadura unitateei 90, $v_{ij} > 1000$ kasuetan 100, eta baldintza hauek ez betetzekotan, $< v_{ij} >$ balioa jarri da.

4. ONDORIOAK

Endekapenezko gaixotasun neurologikoak gizartearen populazio handiari eragiten die eta azken urteotan horri buruzko ikerkuntzak garrantzi handia hartu du. Gaixotasun hauek, eragile anitzeko gaitz ahulgarriak dira, hau da, faktore desberdinek hauen sorrera eragin dezakete. Gaur egun, neurozientzialariek, arrakastarik gabeko mekanismo basikoetatik ateratako datuak ustiatzen ari dira, mekanismo berriak proposatzeko. Mekanismo berri hauek, gaixotasun hauen detekzio goiztiarrerako farmako baliagarrien identifikazioa lortu nahi dute. Zentzu honetan, PTIAko ereduak erreminta oso eraginkorrak izan daitezke gaixotasun horien kontrako sendagaiak aurkitzeko datu sorta handiak analizatzeko gai baitira. Izan ere, konposatu bat endekapenezko gaixotasun neurologikoetan aktiboa izateko edo ez izateko aukera iragar dezakete ereduek. Beraz, eredu hauek, gaixotasun hauen atzean dauden mekanismoak ulertzeko erreminta oso baliagarriak direla frogatu da eta horri esker, etorkizun handiko bide terapeutikoak zabaltzen dira.

5. ESKER ONAK

Artikulu hau UPV/EHUk eta Eusko Jaurlaritzak finantzatutako IT1558–22, PID2022–137365NB-100 eta KK-2022/00032 proiektuei esker idatzi izan da.

6. BIBLIOGRAFIA

[1] RUSSELL, S. J., eta NORVIG, P. 2022. *Artificial Intelligence A Modern Approach*. Pearson Education, United Kingdom.

[2] COPELAND, J. 2004. *The Essential Turing: The Ideas That Gave Birth to the Computer Age* . Clarendon Press, Oxford.

[3] MITCHELL, T. M. 1997. *Machine Learning*. McGraw Hil.

[4] DUDA, R., HART, P., eta STORK, D. 2001. *Pattern Classification.* Wiley, New York.

[5] MACKAY, D. 2005. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

[6] BISHOP, C. 2006. *Pattern Recognition and Machine Learning*. Springer, Berlin.

[7] HASTIE, T., TIBSHIRANI, R., eta FRIEDMAN, J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York.

[8] NEEL S MADHUKAR, P. K. 2019. «A Bayesian machine learning approach for drug target identification using diverse data types». *Nature Communications* 1–14.

[9] VAPNIK, V. 2006. *Estimation of Dependences Based on Empirical Data*. Springer Science eta Business Media.

[10] SCHÖLKOPF, B., TSUDA, K. eta VERT, J.-P. 2004. *Kernel Methods in Computational Biology*. MIT press.

[11] CAMPBELL, C. eta YING. Y. 2011. *Learning with Support Vector Machines. Synthesis Lectures on Artificial Intelligence and Machine Learning 5.1*. Morgan & Claypool Publishers series.

[12] NELLO CRISTIANINI, J. S.-T. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge university press, Cambridge.

[13] BREIMAN, L. 2001. «Random forests». *Machine learning*, **45**, 5–32.

[14] HSU, C. Y. 2021. «Intrusion detection by machine learning for multimedia platform». *Multimedia tools and applications,* **80**, 29643–29656.

[15] GIL-MARTÍN, M., VILLA-MONEDERO, M., POMIRSKI, A., SÁEZ-TRIGUEROS, D., eta SAN-SEGUNDO, R. 2023. «Sign Language Motion Generation from Sign Characteristics». *Sensors (Basel, Switzerland)*, **23**, 9365.

[16] KAUSHIK, K. B. 2022. «Multinomial Naive Bayesian Classifier Framework for Systematic Analysis of Smart IoT Devices». *Sensors (Basel, Switzerland),* **22**, 7318.

[17] HU, Q. M. 2022. «FROST: Fallback Voice Apps Recommendation for Unhandled Voice Commands in Intelligent Personal Assistants». *Frontiers in big data*, **5**, 867251.

[18] SANTANA, R., ZULUAGA, R., GAÑAN, P., ARRASATE S., ONIEVA, E., GONZÁLEZ-DÍAZ H. 2020. «Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models». *Nanoscale*, **12**, 13471–13483.

[19] CARRACEDO-REBOREDO, P. ARANZAMENDI, E., HE, S., ARRASATE, S., MUNTEANU, C. R., FERNANDEZ-LOZANO, C., SOTOMAYOR, N., LETE, E. eta GONZÁLEZ-DÍAZ, H. 2024. «MATEO: intermolecular α-amidoalkylation theoretical enantioselectivity optimization. Online tool for selection and design of chiral catalyst and products». *Journal of Cheminformatics*, **16**, 9.

[20] ENGEL, T. 2006. «Basic Overview of Chemoinformatics». *Journal Chemical Information and Modelling*, **46**, 2267–2277.
[21] LIU, S. 2024. «Harvesting Chemical Understanding with Machine Learning and Quantum Computers». *ACS physical chemistry Au*, **4**, 135–142.
[22] PARRILL, A. L., eta LIPKOWITZ, K. B. 2018. *Reviews in Computational Chemistry,* Wiley.
[23] CENTURY, N. R. 2003. *Chemical Theory and Computer Modeling: From Computational Chemistry to Process Systems Engineering, Beyond the Molecular Frontier: Challenges for Chemis.* National Academies Press (US), Washington.
[24] KOROBOV, V. I., eta KARR, J.-P. 2021. «Rovibrational spin-averaged transitions in the hydrogen molecular ions». *Physical Review A*, **104**, 032806.
[25] NOZIÈRES, P. 1997. *Theory of Interacting Fermi Systems. Advanced Book Classics*. Mass: Perseus Publishing, Cambridge.
[26] WILLEMS, H., DE CESCO, S., eta SVENSSON, F. 2020. «Computational Chemistry on a Budget: Supporting Drug Discovery with Limited Resources: Miniperspective». *Journal of Medicinal Chemistry*, **63**, 10158–10169.
[27] RESTREPO, G. 2016. *Mathematical Chemistry, a New Discipline. In Essays in the Philosophy of Chemistry*. Oxford University Press, New York, UK.
[28] IVAN GUTMAN, O. E. 1988. *A Review of the Book by Ivan "Mathematical Concepts in Organic Chemistry" in SIAM Review 30, 2, 1988*. Springer-Verlag, Berlin.
[29] GONZÁLEZ-DÍAZ, H. ARRASATE, S., GÓMEZ-SANJUAN, A., SOTOMAYOR, N., LETE, E., BESADA-PORTO, L. eta RUSO, J. M. 2013. «General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry». *Current Topics in Medicinal Chemistry*, **13**, 1713–1741.
[30] QUEVEDO-TUMAILLI, V. , ORTEGA-TENEZACA, B., GONZÁLEZ-DÍAZ, H. 2021. «IFPTML Mapping of Drug Graphs with Protein and Chromosome Structural Networks vs. Pre-Clinical Assay Information for Discovery of Antimalarial Compounds». *Int. J. Mol. Sci.*, **22**, 13066.
[31] BALTASAR-MARCHUETA, M., LLONA, L., M-ALICANTE, S., BARBOLLA, I., IBARLUZEA, M. G., RAMIS, R., SALOMON, A. M., FUNDORA, B., ARAUJO, A., MUGURUZA-MONTERO, A., NUÑEZ, E., PÉREZ-OLEA, S., VILLANUEVA, C., LEONARDO, A., ARRASATE, S., SOTOMAYOR, N., VILLARROEL, A., BERGARA, A., LETE, E. eta GONZÁLEZ-DÍAZ, H. 2024. «Identification of Riluzole derivatives as novel calmodulin inhibitors with neuroprotective activity by ajoint synthesis, biosensor, and computational guided strategy.» *Biomedicine and Pharmacotherapy* ,**174**, 116602.
[32] VELÁSQUEZ-LÓPEZ, Y., RUIZ-ESCUDERO, A., ARRASATE, S., eta GONZÁLEZ-DÍAZ, H. 2024. «Implementation of IFPTML Computational Models in Drug Discovery Against Flaviviridae Family». *Journal of Chemical Information and Modeling* ,**64**, 1841–1852.
[33] ROY, K. , SUPRATIK, K., eta NARAYAN, D. 2015. *A Primer on QSAR/QSPR Modeling Fundamental Concepts* . Springer, New York.

[34] NANTASENAMAT, C., ISARANKURA-NA-AYUDHYA, C., NAENNA, T., eta PRACHAYASITTIKUL, V. 2009. «A practical overview of quantitative structure-activity relationship». *EXCLI Journal* ,**8**, 74–88.

[35] VILAR, S. , SANTANA, L. eta URIARTE, E. 2006. «Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action». *J.Med.Chem.* , **49**, 1118–1124.

[36] SANTANA, L. , URIARTE, E., GONZALEZ-DIAZ, H., ZAGOTTO, G., SOTO-OTERO, R., eta MENDEZ-ALVAREZ, E. 2006. «A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins». *J. Med. Chem.*, **49**, 1149–1156.

[37] SANTANA, L. , GONZALEZ-DIAZ, H., QUEZADA, E., URIARTE, E., YANEZ, M., VINA, D. eta ORALLO, F. 2008. «Quantitative structure-activity relationship and complex network approach to monoamine oxidase A and B inhibitors». *J. Med. Chem.* ,**51**, 6740–6751.

[38] LANEY, D. 2001. «3D data management: Controlling data volume, velocity and variety». *META Group Research Note* ,**6**.

[39] LIU, Z., YONG, W., XIANG-SUN, Z., WEIMING, X., eta LUONAN, C. 2011. «Detecting and analyzing differentially activated pathways in brain regions of Alzheimer's disease patients». *The Royal Society of Chemistry* ,**7**, 1441–1452.

[40] GAULTON, A., HERSEY, A., NOWOTKA, M., BENTO, A.P., CHAMBERS, J., MENDEZ, D., MUTOWO, P., ATKINSON, F., BELLIS, L.J., CIBRIÁN-UHALTE, E., DAVIES, M., DEDMAN, N., KARLSSON, A., MAGARIÑOS, M.P., OVERINGTON, J.P., PAPADATOS, G., SMIT, I., LEACH, A.R. 2017. «The ChEMBL database in 2017». *Nucleic Acids Research*, **45**, 945–954.

[41] SUN, H. 2004. «A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption». *Journal of Chemical Information and Modeling*, **44**, 748–757.

[42] PAJOUHESH, H., eta LENZ, G. 2005. «Medicinal Chemical Properties of Successful Central Nervous System Drugs». *The Journal of The American Society for Experimental NeuroTherapeutics*, **2**, 541–553.

[43] SHEN, H., eta CHOU, K. 2008. «PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition». *Analytical Biochemistry*, **373**, 386–388.

[44] CHOU, K. 2001. «Prediction of protein cellular attributes using pseudo-amino-acid-composition». *PROTEINS: Structure, Function, and Genetics*, **43**, 246–255.

[45] CHOU, K. 2005. «Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes». *Bioinformatics*, **21**, 10–19.

[46] CHOU, K. 2000. «Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect». *Biochemical and Biophysical Research Communications,* **278**, 477–483.

[47] DUARDO-SANCHEZ, A., GONZÁLEZ-DÍAZ, H., eta PAZO, A. 2014. «MI-NODES Multiscale Models of Metabolic Reactions, Brain Connectome, Ecological, Epidemic, World Trade, and Legal-Social Networks». *Current Bioinformatics*, **9**.

[48] PATHRIA, R. K., eta BEALE, P. 2011. *Statistical Mechanics*. Academic Press, USA.

[49] 2001. *StatSoft, Inc. STATISTICA (Data Analysis Software System), Version 6. (2001).* Retrieved from www.statsoft.com.

[50] JAWORSKA, J., NIKOLOVA-JELIAZKOVA, N., eta ALDENBERG, T. 2005. «QSAR applicability domain estimation by projection of the training set descriptor space: A review». *Altern. Lab. Anim*, **33**, 445–459.

[51] TROPSHA A., GRAMATICA P., GOMBAR V. 2003. «The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models». *QSAR Comb. Sci.,* ,**22**, 69–77.

[52] COOK, R., eta WEISBERG, S. 1982. *Residuals and Influence in Regression*. Chapman and Hall, New York.

[53] HU, M., ZHANG, G., JIANG, C., eta PATUWO, B. 1999. «A Cross-Validation Analysis of Neural Network Out-of-Sample Performance in Exchange Rate Forecasting». *Decision Sciences*, **30**.

[54] NETZEVA, T.I., WORTH A.; ALDENBERG, T.; BENIGNI, R.; CRONIN, M.T.D.; GRAMATICA, P.; JAWORSKA, J.S.; KAHN, S.; KLOPMAN, G.; MARCHANT, C.A.; et. al. 2005. «Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. Thereport and recommendations of ECVAM Workshop 52». *Alternatives to laboratory animals*, **33**, 155–173.