Universidad del País Vasco Euskal Herriko Unibertsitatea
BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

TELEKOMUNIKAZIO INGENIARITZA UNIBERTSITATE MASTERRA

MASTER AMAIERAKO LANA

NIR espektroskopia eta aldagai anitzeko anilisiaren bitarteko behi esnearen parametro fisiko-kimiko desberdinen modelizazioa

Ikaslea: Zenarrutzabeitia Velez, Sabin

Zuzendaria: ***Ayesta Ereño, Igor***

Ikasturtea: 2023-2024

Data: Durango, 2024ko ekainaren 6a

LABURPENA

Esne-lantegiek, produzitutako behi-esnearen segurtasun eta kalitate-kontrola egin behar dute bere produktua kontsumorako egokia dela eta kalitate optimoa duela ziurtatzeko. Tradizionalki, kontrol hau analisi kimikoen bidez egin da, behi-esnearen laginak laborategian aztertuz. Esne-lantegiek baina, laborategira eraman beharraren gastuak eta denbora aurreztu nahi dituzte.

Proiektu honen helburua, NIR (Near-Infrared) espektroskopiaren eta datuen aldagai anitzeko analisiaren bidez, behi esnearen zenbait parametro fisiko-kimiko modelatzea eta aurreikustea da, produktuaren kalitate eta segurtasunean esanahi handia dutenak. Parametro fisiko-kimiko horiek, zehazki, honako hauek dira: Laktosa, grasa, proteina, zelula somatikoak eta gihar estraktu lehorra.

Helburu hau lortzeko, NIR espektroskopiaren bidez lortutako behi-esne laginen espektroarekin eta laborategitik lortutako laginen analisi kimikoarekin, Simca software-aren bitartez datuen aldagai anitzeko analisia egingo da. Parametro bakoitzerako modelo desberdinak probatu eta ikertuko dira, guztietatik parametroa zeinek modelatzen eta aurreikusten duen hobeto aztertuz.

Beraz, parametro bakoitzerako modelo egokiena bilatuko da eta honen analisi sakon bat jorratuko da bere ezaugarriak deskribatuz.

Gako hitzak: NIR, espektroskopia, datuen aldagai anitzeko analisia, MVDA, behi-esnea, iragarpena, Simca, esne-lantegia.

RESUMEN

Las fábricas de lácteos deben realizar un control de seguridad y calidad de la leche de vaca producida para asegurar que su producto es apto para el consumo y tiene una calidad óptima. Tradicionalmente, este control se ha realizado mediante análisis químicos, analizando las muestras de leche de vaca en un laboratorio. Sin embargo, las fábricas de lácteos quieren ahorrar tiempo y gastos en llevarlos al laboratorio.

El objetivo de este proyecto, mediante el análisis de datos multivariado y espectroscopia NIR (Near-Infrared), es modelar y predecir ciertos parámetros físico-químicos de la leche de vaca, que tienen una gran importancia en la calidad y seguridad del producto. Estos parámetros físico-químicos son: lactosa, grasa, proteína, células somáticas y extracto seco magro.

Para conseguir este objetivo, se realizará un análisis de datos multivariado mediante el software Simca a partir del espectro de muestras de leche de vaca obtenidas mediante espectroscopia NIR y el análisis químico de muestras obtenidas del laboratorio. Se probarán e investigarán diferentes modelos para cada parámetro, analizando mejor cuál de ellos modela y prevé el parámetro.

Por lo tanto, se buscará el modelo más adecuado para cada parámetro y se abordará un análisis en profundidad de éste describiendo sus características.

Palabras clave: NIR, espectroscopia, análisis de datos multivariado, MVDA, leche de vaca, predicción, Simca, fábrica de lácteos.

ABSTRACT

Dairy factorys must carry out a safety and quality control of the cow's milk produced to ensure that their product is fit for consumption and of optimum quality. Traditionally, this control has been done by chemical analysis, testing cow's milk samples in a laboratory. However, dairy factorys want to save time and expense in taking them to the laboratory.

The objective of this project, using multivariate data analysis and NIR (Near-Infrared) spectroscopy, is to model and predict certain physicochemical parameters of cow's milk, which are of great importance in the quality and safety of the product. These physicochemical parameters are lactose, fat, protein, somatic cells and dry weight extract.

To achieve this objective, a multivariate data analysis will be performed using Simca software from the spectrum of cow's milk samples obtained by NIR spectroscopy and the chemical analysis of samples obtained from the laboratory. Different models for each parameter will be tested and investigated, analysing better which one models and predicts the parameter.

Therefore, the most suitable model for each parameter will be searched for and an in-depth analysis of it will be approached describing its characteristics.

Keywords: NIR, spectroscopy, multivariate data analysis, MVDA, cow's milk, prediction, Simca, dairy factory.

AURKIBIDEA

IRUDIEN AURKIBIDEA

TAULEN AURKIBIDEA

AKRONIMOAK

MVDA: Multivariate Data Analysis

PCA: Principal Components Analysis

PLS-R: Partial Least Squares – Regression

OPLS: Orthogonal Partial Least Squares

LDA: Linear Discriminant Analysis

RMSEcv: Root Mean Squared Error Cross Validation

RMSEP: Root Mean Squared Error Prediction

ESM: Extracto Seco Magro

IR: Infrared

NIR: Near-Infrared

MIR: Medium-Infrared

MEMORIA

1.SARRERA ETA TESTUINGURUA

Master amaierako lan honen helburu nagusia, NIR (Near-Infrared) espektroskopiaren eta datuen aldagai anitzeko analisiaren bidez, behi esnearen zenbait parametro fisiko-kimiko modelatzea eta aurreikustea da.

Analisi hauek esne-lantegi batek emandako behi-esne laginekin egin da. Normalean, esne-lantegiek, euren produktuaren kalitatea eta egoera ziurtatzeko, esne laginak laborategira bidali izan dituzte bertan ikertuak izateko. Bertan analisi kimiko desberdinen bitartez esnearen kalitatea eta egoera zehazten da. Hau beharrezkoa da behiak osasuntsu daudenaren eta esneak beharrezko konposizio eta balio nutrizionala dutenaren bermea izateko. Honek, baina, gastu ekonomiko gehigarri bat suposatzen du esne ekoizleentzako. Honen ondorioz, kalitate eta segurtasun-kontrol hau bermatzen duten teknika berriak bilatu dira, eta NIR espektroskopia eta datuen aldagai anitzen analisiaren konbinaketa, aukera errentagarrienetako bat da; honen bidez laginak laborategira eramatea ekiditen daitekeelako. Espektrometro baten bidez espektro elektromagnetikoko infragorri hurbileko esparruko espektroa lortzen da eta ondoren datuen aldagai anitzeko analisiaren bidez, bertatik informazio baliagarria atera eta iragarpen modeloak sortzen dira, jakin nahi den parametroaren balioa lortzeko.

Ekoizpenaren kalitatea, segurtasuna eta eraginkortasuna handia izatea beharrezkoa da gaur egungo esne-industria modernoan. Kalitate handiko esnekien eskaerak eta elikagaien segurtasuna bermatzeko beharrak esnea aztertzeko eta kontrolatzeko teknologia aurreratuak erabiltzea bultzatzen dute. Teknologia hauen artean NIR espektroskopia eta datuen aldagai anitzeko

analisia erreminta boteretsuak bihurtu dira esparru honetan, euren lan egiteko modu azkar, ez-suntsitzaile eta zehatzagatik.

Lan honetan konkretuki, behi-esnearen parametro fisiko-kimiko hauek modelizatu eta aurreikusi nahi dira: laktosa, grasa, proteina, zelula somatikoak eta gihar estraktu lehorra. Parametro bakoitzerako modelorik egokiena bilatu, aztertu eta deskribatuko da. Analisi guzti hau SIMCA programarekin gauzatuko da.

2.LANAREN HELBURUAK ETA IRISMENA

Helburu nagusia aipatutako parametro fisiko-kimikoetara hobeto egokitzen diren modeloak aurkitzea da. Helburu honetara heltzeko zenbait pausu edo azpi-helburu jorratu beharko dira.

Lehendabizi, espektrometrotik ateratako espektroa analizatu beharko da SIMCA programaren bitartez, honen izaera eta berezitasunak ezagutzeko. NIR esparruan zehar esne-laginek duten forma aztertuko da, normaz kanpoko portaerak identifikatzeko.

Ondoren, iragarri nahi diren parametro bakoitzerako, modelo ezberdinak doituko dira. Prozesua errepikatuko da lagin multzo desberdinekin eta haietatik lortutako datuak kontrastatuko dira.

Modelo guztietatik, parametro bakoitzerako, parametrora hobetoen egokitzen den eta hobetoen aurreikusten duen modeloa aukeratuko da.

Azkenik, aukeratutako modelo hauetan sakonduko da eta hauen bidez lortzen diren emaitzak analizatu eta deskribatuko dira.

3. GAIAREN EGUNGO EGOERA

Atal honetan proiektu honi dagokien gai garrantzitsuenak azaldu eta deskribatuko dira. Kimiometria zer den azalduko da, datuen aldagai anitzeko analisia egiteko dauden teknika desberdinen laburpen bat egingo da, NIR espektroskopiari buruz hitz egingo da eta hauek esne-industrian duten garrantzia azpimarratuko da.

3.1 Kimiometria

Kimiometria diziplina kimiko bat da, metodo matematikoak eta estatistikoak erabiltzen dituena neurketa-prozedura eta esperimentu optimoak diseinatzeko edo hautatzeko, eta datu kimikoen analisiaren bidez ahalik eta informazio kimiko handiena emateko [1].

3.1.1 Kimiometriaren aplikazio-eremuak

- **Aldagai anitzeko optimizazioa: Analisi kimiko bat egiteko baldintza optimoak zehazten laguntzen du, emaitzen zehaztasuna hobetuz.**
- **Aldagai anitzeko analisia: Osagai Nagusien Analisia (PCA) bezalako metodoak datu konplexuetan patroiak eta erlazioak identifikatzeko erabiltzen dira.**
- **Aldagai anitzeko kalibrazioa: Laginetan osagai kimikoen kontzentrazioak iragartzeko "Partial Least Square (PLS)" bezalako metodoak erabiltzen dira.**
- **Kalitate-kontrola eta kalitate-bermea: Prozesu kimikoak monitorizatu eta kontrolatzeko eta zehaztutako mugen barruan mantentzen direla ziurtatzeko erabiltzen da [1].**
- **Metodoak balioztatzea: Metodo analitikoen fidagarritasuna eta erreproduzigarritasuna balioztatzeko eta egiaztatzeko estatistika-teknikak aplikatzea.**
- **Modelatzea eta simulazioa: Eredu matematikoak sortzea portaera kimikoak simulatzeko eta konposatu kimikoen propietateak iragartzeko [1].**

Aplikazio praktikoen adibideak:

- **Industria farmazeutikoa: Sendagaietan konposatu aktiboak identifikatzeko eta kuantifikatzeko metodoak garatzea eta balioztatzea [1].**
- **Elikagaien industria: Elikagaien kalitate-kontrola eta autentifikazioa, baita adulteratzaileen detekzioa ere [1].**
- **Ingurumen-ikerketa: kutsatzaileen monitorizazioa eta analisia ingurumen-laginetan [1].**

3.2 Datuen aldagai anitzeko analisia

Datuen aldagai anitzeko analisia (MVDA – Multivariate Data Analysis) estatistika-teknika bat da, aldagai anitz dituzten behaketak dituzten datu-multzoak aztertzeko erabiltzen dena [2].

Aldagai bakarreko analisiarekin konparatuta, MVDA-k aldagai anitzen arteko elkarrekintza aztertzen du aldi berean. Metodologia hau bereziki praktikoa da aldagaiak euren artean erlazionatuta dauden datu multzo konplexuekin lan egiten denean [3]. Elikagaien zientzia, Kimika, Biologia, Ingeniaritza eta Gizarte zientziak bezalako esparruetan aplikatzen da MVDA.

MVDA-k datuen egitura eta erlazioak sakonago ulertzea ahalbidetzen du, eta hori oso baliagarria da datuak esploratzeko, bistaratzeko, dimentsionaltasuna murrizteko, predikziorako, modelizaziorako eta beste aplikazio analitiko batzuetarako [3].

MVDA teknika ohikoenetako batzuk hauek dira: PCA (Principal Components Analysis), PLS-R (Partial Least Squares – Regression) eta OPLS (Orthogonal Partial Least Squares).

Jarraian hiru teknika hauek deskribatuko dira.

3.2.1 MVDA teknikak

3.2.1.1 PCA (Principal Components Analysis)

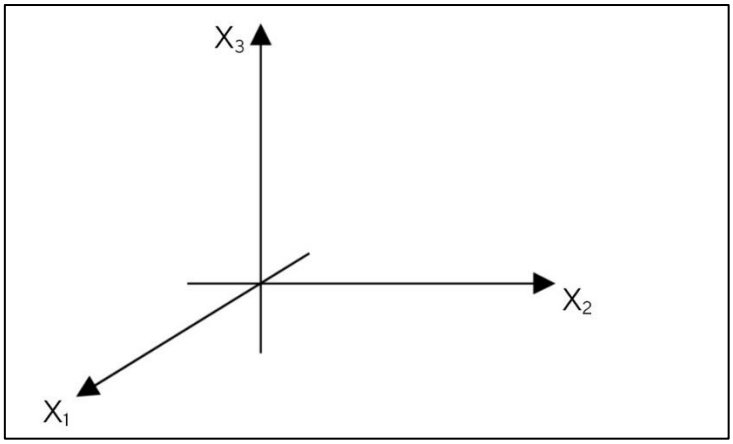
Osagai Nagusien Analisia (PCA, ingelesezko sigletan) dimentsionaltasuna murrizteko teknika bat da, eta asko erabiltzen da datuen aldagai anitzeko analisian. Haren helburu nagusia, datu-multzo batean dagoen aldakortasun gehiena azaltzen duten osagai nagusiak identifikatzea da [3].

PCA-k proiekzio-metodoetan oinarritutako aldagai anitzeko datuen analisiaren oinarria osatzen du. PCA-ren erabilerarik garrantzitsuena, aldagai anitzeko datuen taula bat, aldagai multzo txikiago baten gisa irudikatzea da, joerak, jauziak, konglomeratuak eta balio atipikoak aztertzeko [4]. Ikuspegi orokor horrek behaketen eta aldagaien arteko erlazioak eta aldagaien arteko erlazioak adieraz ditzake.

PCA oso tresna malgua da eta multikolinealtasuna, falta diren balioak, datu kategorikoak eta neurketa zehaztugabeak dituzten datu-multzoak aztertzeko aukera ematen du [4]. Helburua da datuetatik informazio garrantzitsua ateratzea eta informazio hori osagai nagusi gisa adieraztea.

Estatistikoki, PCA-k lerroak, planoak eta hiperplanoak aurkitzen ditu K-dimentsioko espazioan, datuetara ahalik eta hobeen hurbiltzen direnak minimo karratuen zentzuan [4].

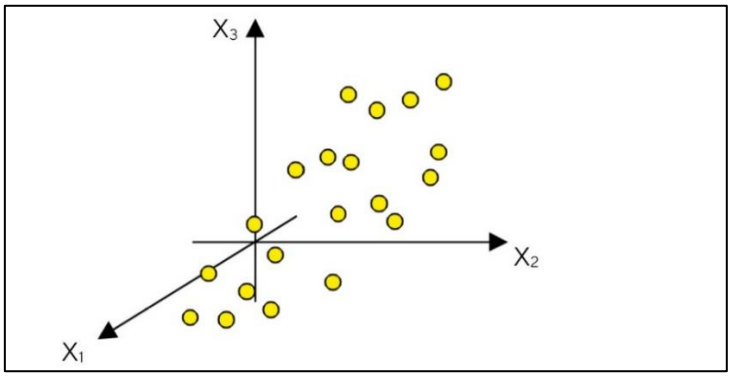
Kontsidera dezagun X matrize bat, N lerrorekin ("behaketa" ere esaten zaio) eta K zutaberekin ("aldagai" ere esaten zaio). Matrize horretarako, aldagaiak adina dimentsioko espazio bat eraikitzen da, hurrengo irudian ikus daitekeen moduan.



***1**. Irudia: 3 dimentsioko espazioa [4]*

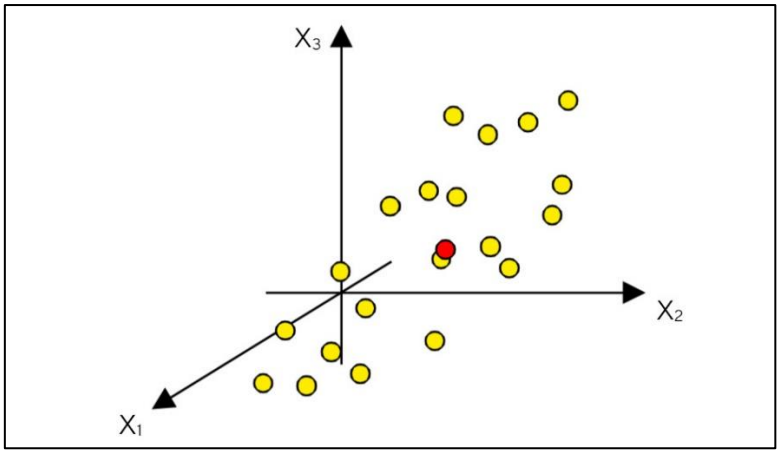
Aldagai bakoitzak koordenatu-ardatz bat irudikatzen du. Aldagai bakoitzerako, luzera eskala-irizpide baten arabera estandarizatu da, normalean bariantza unitariora eskalatuz [4].

Ondoren, X matrizeko behaketa edo ilara bakoitza K dimentsioko espazioan jartzen da. Ondorioz, datu-taulako lerroek puntu-multzo bat osatzen dute espazio horretan.



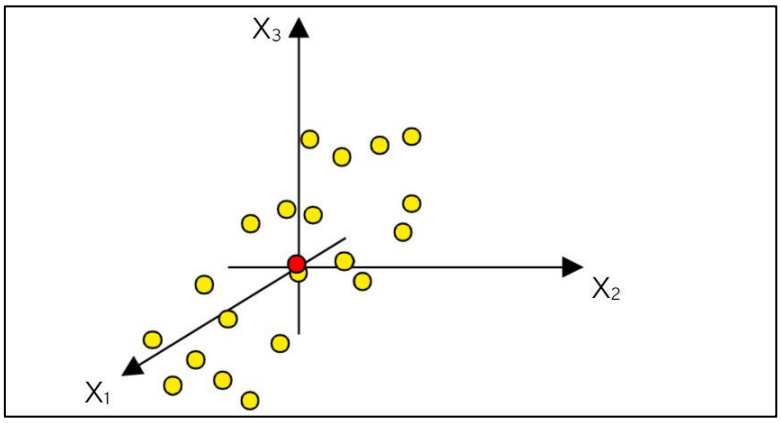
***2**. Irudia: X matrizeko behaketak 3 dimentsioko espazio baten kokatuta [4]*

Ondoren, batezbestekoan zentratzen da, honek datuen aldagaien batez bestekoak kentzea dakar. Lehenik, aldagaien batez bestekoak kalkulatzen dira. Batez bestekoen bektore hori puntu bat (beheko irudian gorriz ageri dena) bezala interpreta daiteke espazioan eta puntu-sortaren erdian dago [4].



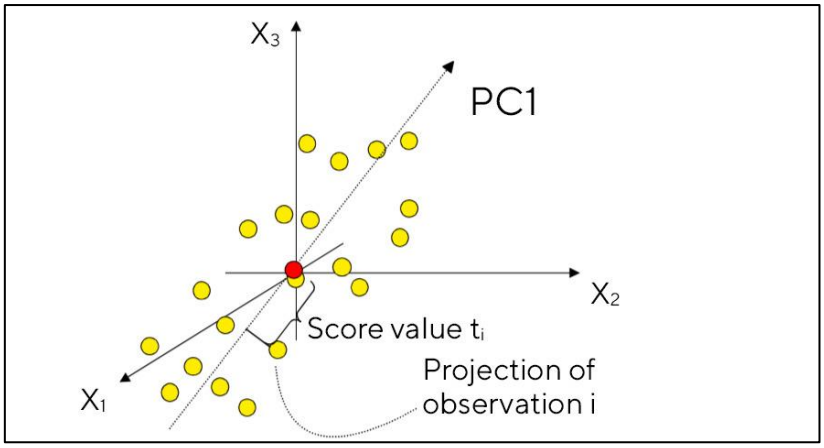
***3**. Irudia: Batezbestekoen bektorea gorriz puntu-sortaren erdian [4]*

Datuen batez bestekoen kenketak koordenatu-sistemaren birposizionamendu bat ekartzen du; beraz, orain batez besteko puntua jatorria da:



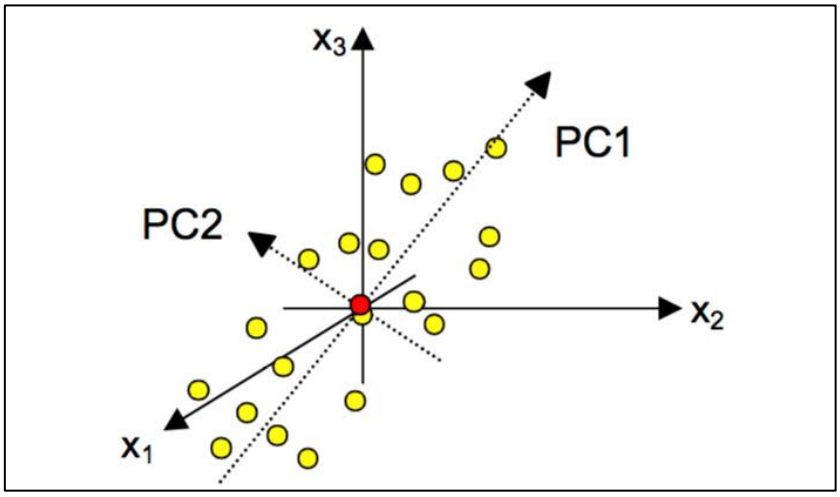
***4**. Irudia: Koordenatu-sistemaren birposizionamendua jatorrira [4]*

Orain, datu-multzoa prest dago lehen osagai nagusia (PC1) kalkulatzeko. Lehenengo osagai nagusia puntu-sortaren forma hobekien adierazten duen lerroa da [4]. Datuen bariantza maximoaren norabidea adierazten du. Behaketa bakoitza (puntu horiak) lerro honetan proiekta daiteke, osagai nagusiaren lerroaren zehar koordenatu-balio bat lortzeko. Balio horri "Score" esaten zaio.



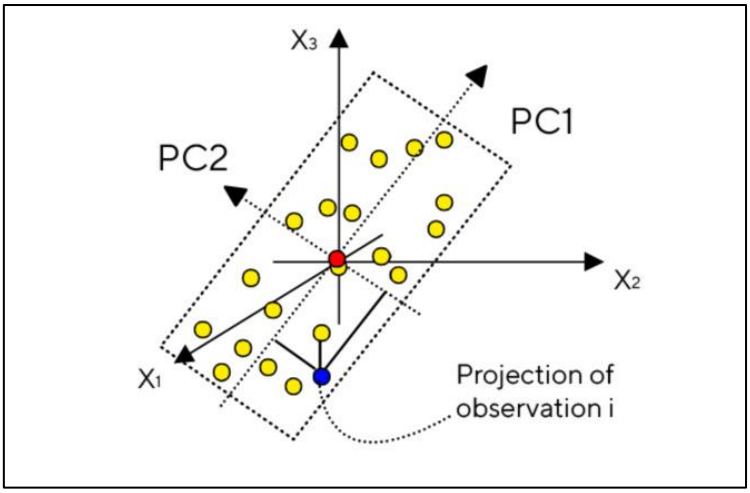
***5**. Irudia: Puntu-sortaren lehen osagai nagusia (PC1)[4]*

Normalean, osagai nagusi bat ez da nahikoa datu-multzo baten aldakortasuna modelatzeko. Beraz, bigarren osagai nagusi bat kalkulatzen da (PC2). Bigarren osagai nagusia datuen bigarren aldakuntza-iturri handiena islatzeko eta, aldi berean, lehen osagai nagusiarekiko ortogonala izateko moduan orientatuta dago. PC2 ere erdiko puntutik igarotzen da [4].



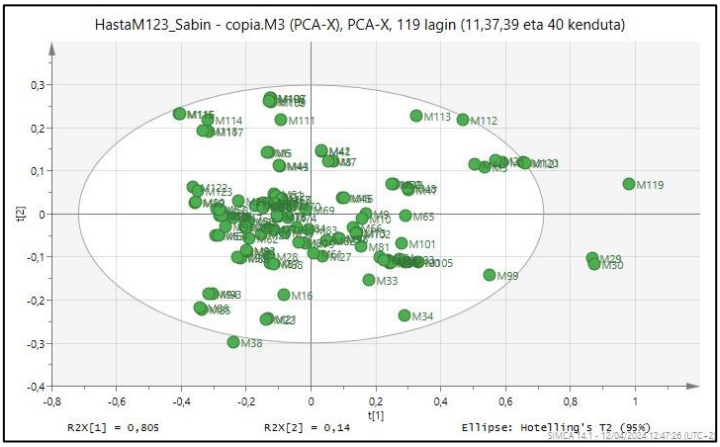
***6**. Irudia: Puntu sortaren bigarren osagai nagusia (PC2)[4]*

Bi osagai nagusik plano bat osatzen dute. Behaketa bakoitza plano honetan proiekta daiteke, bakoitzari "score" bat emanez. Proiektatutako konfigurazio horren irudikapena "score" grafiko gisa ezagutzen da.



***7**. Irudia: Bi osagai nagusik osatutako planoa [4]*

Honen bidez beheko irudiko moduko "Score" grafikoak lortzen dira.



***8**. Irudia: "Score" grafiko baten adibidea*

3.2.1.2 PLS (Partial Least Squares)

Partial Least Squares (PLS) aldagai anitzeko modelatze-metodo bat da, bi aldagai multzo erlazionatzeko erabiltzen dena: aldagai iragarleen multzo bat (X) eta erantzun-aldagaien multzo bat (Y). Bereziki erabilgarria da dimentsionaltasun handiko datuekin lan egiten denean edo aldagai iragarleen artean multikolinealtasuna dagoenean.

PLS-k aldagai iragarleen konbinazio linealak diren hainbat osagai aurkitzen ditu. Osagai horiek aldagai iragarleetan ahalik eta

aldakortasun gehien azaltzeko eta, aldi berean, erantzun-aldagaiekin korrelazio handia izateko hautatzen dira.

PLS-ren modeloa eraikitzeko, osagai horiek modu iteratiboan prozesatzen dira, non urrats bakoitzean X eta Y-ren arteko kobariantza maximizatzen duten norabideak identifikatzen diren.

PLS-k aldagai iragarleetatik abiatuta erantzun-aldagaiak aurreikusteko erabil daitezkeen osagai-multzo bat lortzen du emaitza modura [5]. Osagai horiek aldagaien arteko erlazioei buruzko informazioa ere eman dezakete.

Datu konplexuekin eta korrelazio handia dutenekin lan egiteko erabiltzen da bereziki.

Tresna indartsua da aldagai anitzeko datuak modelatzeko eta aldagai multzoen arteko erlazioak esploratzeko.

3.2.1.3 OPLS (Orthogonal Partial Least Squares)

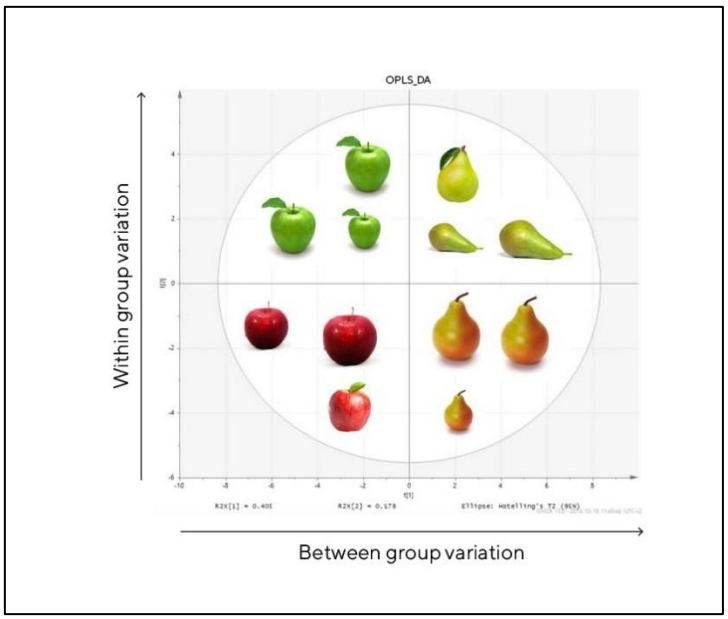
OPLS, PLS metodoaren hedapen bat da, aldagai iragarle multzo bat (X) eta erantzun-aldagai multzo bat (Y) erlazionatzeko erabiltzen dena. PLS-ren antzera, OPLS-ren helburu nagusia bi aldagai-multzo horien arteko erlazioa modelatzea da, baina berezitasun batekin: iragarleen aldakuntza bi zatitan bereizten da: zati bat zuzenean lotuta dago erantzun-aldagaiarekin (osagai prediktiboak) eta beste zati bat ortogonala da erantzun-aldagaiarekin (osagai ortogonalak) [6].

Osagai prediktiboek erantzun-aldagaiarekin zuzenean lotuta dagoen iragarleen aldakuntza adierazten dute; osagai ortogonalek, berriz, erantzun-aldagaiarekin korrelaziorik ez duten iragarleen aldakuntza. Horri esker, argi eta garbi bereiz daiteke erantzunarekin zerikusirik ez duen informazioa eta iragartzeko garrantzitsua den informazioa.

Honek ereduaren interpretagarritasuna eta erlazio esanguratsuak identifikatzeko gaitasuna hobetzen ditu.

OPLS bereziki erabilgarria da aldagai iragarleen artean korrelazio handia dagoen egoeretan, aurreikuspenerako garrantzitsua den informazioa erantzunarekin zerikusirik ez duen informaziotik argi eta garbi bereizteko aukera ematen baitu [6].

Hurrengo irudian modu oso argian ikusten da OPLS-ren funtzionamendua adibide erreal baten bidez.



***9**. Irudia: OPLS modelo baten "score" grafiko adibidea [6]*

Bertan, laginak sagarrak edo madariak diren iragarri nahi da. "Score" grafika honetan, ardatz horizontala osagai prediktiboari dagokio eta bertikala ortogonalari. Ikusten da ardatz horizontalaren norabidean, ezkerretik eskumara, sagarretatik madarietara igarotzen garela; hau da, talde batekoa den edo beste taldekoa den. Ardatz bertikalaren norabidean aldiz, ez da sagarra edo madaria den desberdintzen, baizik eta kolore berdeagoa edo gorriagoa duten; hau da, aldakortasun horrek ez du zerikusirik sagarra edo madaria den iragartzearekin, horregatik hautematen du osagai ortogonalak, kolore desberdintasunak ez duelako erlaziorik fruta bat edo beste den iragartzeko.

3.3 Espektroskopia

Espektroskopia teknika zientifiko bat da, erradiazio elektromagnetikoak materiarekin duen interakzioa aztertzeko erabiltzen dena. Teknika hori funtsezkoa da diziplina zientifiko eta teknologiko askotan, substantzien konposizioari, egiturari eta propietateei buruzko informazio zehatza emateko gaitasuna baitu [7].

Espektroskopia mota desberdinak daude, lan egiten duten espektro elektromagnetikoaren eskualdearen eta elkarrekintza-mekanismoen arabera. Hauek dira ohikoenak:

***Xurgapen-espektroskopia*: lagin batek uhin-luzera desberdinetan xurgatzen duen argi kantitatea neurtzeaz arduratzen da, eta horrek substantzien kontzentrazioa eta konposizioa zehaztea ahalbidetzen du [8].**

***Emisio-espektroskopia*: Energia gutxiagoko egoerara itzultzean atomoek edo molekulek igortzen duten argia aztertzen du. Elementuak eta konposatu zehatz batzuk aztertzeko erabiltzen da [9].**

***Fluoreszentzia espektroskopia*: argia edo erradiazio elektromagnetikoa xurgatu duen substantzia batek ondoren igortzen duen argia aztertzen du. Asko erabiltzen da biokimikan eta biologia molekularrean [10].**

***Espektroskopia Infragorria (IR (Infrared)):* Bibrazio molekularrak eta lotura kimikoak aztertzen ditu eta oso erabilgarria da konposatu organikoetan dauden talde funtzionalak identifikatzeko [6]. Honen barruan bi mota daude lan egiten duten uhin luzeraren arabera:**

- **NIR (Near-Infrared): 780 – 2500 nm**
- **MIR (Medium-Infrared): 2500 – 25000 nm**

Raman espektroskopia: **IR espektroskopiaren antzekoa da, baina argiaren dispertsio ez-elastikoan oinarritzen da. Bibrazio molekularrei buruzko datuak lortzen ditu eta baliotsua da materialen analisian [11].**

Erresonantzia Magnetiko Nuklearreko espektroskopia: **eremu magnetikoak eta irrati-uhinak erabiltzen ditu molekula organiko zein ez-organikoen egitura ikertzeko. Funtsezkoa da egitura molekularrak zehazteko prozesuan [12].**

Lan honetan NIR (Near-Infrared) espesktroskopiarekin lan egingo da, beraz honetan sakonduko da jarraian.

3.3.1 NIR espektroskopia

Gutxi gorabehera 780 nm-tik 2500 nm-ra bitarteko espektro elektromagnetikoa erabiltzen duen teknika espektroskopikoa da. NIR espektroskopian molekulek argia xurgatzen dute infragorri gertuko eremuan, bibrazio-mailen arteko trantsizioen ondorioz. Xurgapen horiek MIR-arenak baino ahulagoak izaten diren arren, NIR-ek abantaila bat du: argia sakonago sartzen da laginetan, eta lodiera edo dentsitate handiagoko materialak analizatzea ahalbidetzen du [13].

Erabilera asko ditu, laginak modu ez suntsitzailean, azkarrean eta gutxieneko prestakuntzarekin aztertzeko duen gaitasunari esker [14].

Honako aplikazio hauek nabarmentzen dira [14]:

-Elikagaien industrian: elikagaien konposizioaren, hezetasun-edukiaren, koipearen, proteinaren, laktosaren eta beste nutriente batzuen analisia.

- Nekazaritzan: elikagaien kalitatea eta edukia ebaluatzea aleetan.

-Farmazia-industrian: sendagaien kalitatea eta autentikotasuna kontrolatzea eta osagai aktiboen edukia zehaztea.

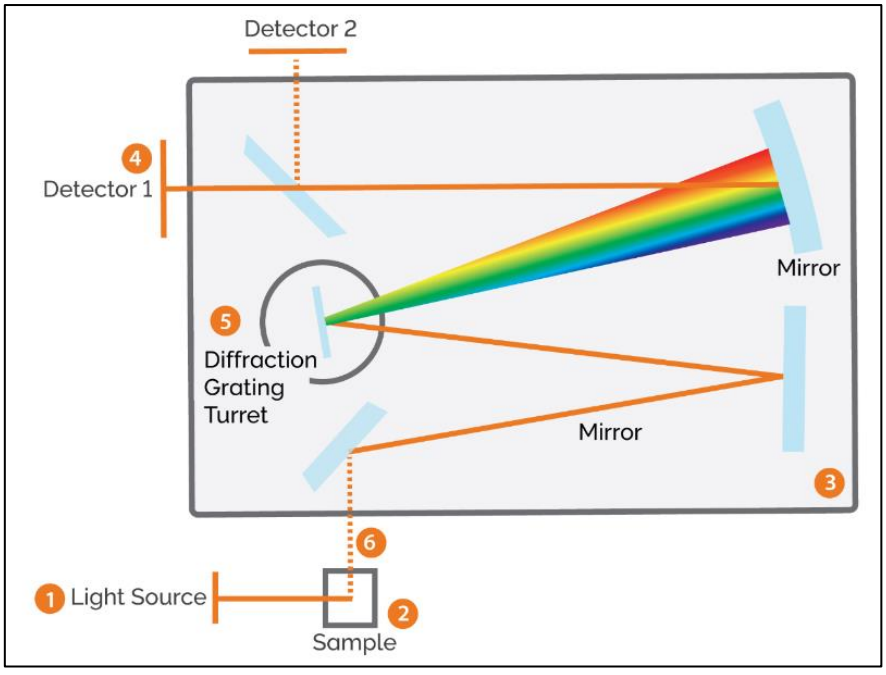
-Medikuntzan: diagnostiko ez-inbaditzailea eta ehunen jarraipena, gorputz-konposizioaren analisia.

- Polimero eta plastikoetan: material polimerikoen konposizioaren eta propietateen analisia.

-Ingurumenaren arloan: kutsatzaileen eta uraren kalitatea monitorizatzea.

Instrumentazioari dagokionez, espektometroetan konfigurazio eta diseinu optiko desberdinak erabiltzen dira, aplikazioaren, beharrizanen, tamainuaren eta kostuaren arabera. Konfigurazio ohikoenetariko bat Czerny-Turner konfigurazioa da, bereizmen handia emateko duen gaitasunarengatik eta uhin-luzerak hautatzeko duen malgutasunarengatik.

Diseinu honek bi ispilu ahur ditu, eta difrakzio lauko sareta bat erdian. Lehen ispiluak argi-iturria kolimatzen du; bigarrenak, berriz, saretaren argia detektagailura bideratzen du [15].



***10**. Irudia: NIR espektometro baten ispilu sistema ahur baten adibidea [15]*

Espektrometro horiek, normalean, hurrengo osagaiekin batera erabiltzen dira:

-<u>Argi iturria:</u>

NIR espektrometroek argi-iturri bizi eta egonkorrarekin erabiltzen dira, eta infragorri hurbilean igortzen dute [15]. Argi iturriaren hautaketa aplikazioaren araberakoa da baina NIR espektrometroetan honakoak erabil daitezke: Tungsteno-Halogeno lanparak, Xenon lanparak, Merkurio lanparak edo LED diodoak.

-<u>Lagina:</u>

Espektrometro mota honetan, lagina espektrometrotik kanpo jartzen da, sarrerako zirrikituaren aurretik. Adibidez, elikagaien ekoizpenean kalitate-probak egiteko, lagina prezipitatuen edalontzi batean egon daiteke, haren bidez zuzendutako argi-iturriarekin. Transmisio-, islapen- edo fluoreszentzia-neurketak espektrometroaren sarrerako zirrikitutik argi emergentea zuzenduz egiten dira [15].

-<u>Detektagailua:</u>

Detektagailu egokia aukeratzea funtsezkoa da NIR espektrometrian. NIR eskualdean oso sentikorrak diren detektagailuak izan behar dira. Detektagailu batek seinale-zarata erlazio onak lortzen baditu, NIR neurketen zehaztasuna hobetzen du [15].

-<u>Espektometro-sareta:</u>

Saretak bere osagai espektraletan sartzen den NIR argia sakabanatzen du. Zehaztasunez diseinatzen dira, dispertsio-propietate bikainak lortzeko eta espektro-analisi zehatza egin ahal izateko [15].

-<u>Zuntz optikoa:</u>

Zuntz optikoak berebiziko garrantzia du NIR espektrometrian, laginaren argia espektrometrora bideratzen du. NIR argia modu eraginkorrean transmititzea bilatzen da, seinale-galera minimizatuz.

Laburbilduz, NIR espektroskopia teknika moldakor eta ahaltsua da, eta aukera ematen du lagin ugari modu azkar eta ez-suntsitzaile baten aztertzeko. Materialen konposizioari eta propietateei buruzko informazio zehatza emateko duen gaitasuna dela eta, balio handikoa da arlo industrial eta zientifiko askotan.

3.5 NIR espektroskopia eta aldagai anitzeko analisia esne-industrian

NIR espektroskopia eta aldagai anitzeko analisia esne-industria guztiz aldatu duten tresna boteretsu eta osagarriak dira. Teknika horiei esker, esnearen eta esnekien hainbat osagai eta propietate modu azkar, ez-suntsitzaile eta zehatz batean azter daitezke.

Hauek esne-industrian dituzten aplikazioen artean honakoak aurki ditzakegu:

-<u>Esnearen propietateen iragarpena</u>:

- **Fenotipoak iragartzea, gantz-azidoen edukia eta abereen osasuna.**
- **Abereen nutrizio-egoera eta egoera metabolikoa monitorizatzea.**

-<u>Kalitatea eta segurtasun kontrolak</u>:

- **Esnearen eta produktu eratorrien etengabeko kalitatea ziurtatzeko eredu iragarleak inplementatzea.**
- **Kalitate- eta segurtasun-arazoak goiz detektatzea.**
- **Esnean kutsatzaileak edo adulteratzaileak identifikatzea.**
- **Esnearen eta esnekien freskotasuna eta egonkortasuna monitorizatzea biltegiratzean.**

-<u>Esnearen konposizioa determinatzea</u>:

- **Esnearen koipe-edukia zehaztea.**
- **Proteina totalen eta espezifikoen analisia.**
- **Laktosaren, mineralen eta uraren neurketa esnean.**

NIR espektroak esnearen osagaiekin korrelatzen dira PCA, PLS edo OPLS bezalako metodo estatistikoen bidez. Teknika hauek datu espektral konplexuak aztertzen dituzte, dimentsionaltasuna murrizten dute eta iragartzeko ereduak eraikitzen dituzte.

3.5 Iragarri nahi diren behi-esnearen parametro fisiko-kimikoak

Lan honetan behi-esnearen bost parametro fisiko-kimiko iragarri nahi dira NIR espektroskopia eta aldagai anitzeko analisiaren bidez. Jarraian parametro hauetako bakoitzak duen garrantzia deskribatuko da.

1.Proteina:

Esnearen proteinak, batez ere kaseinak eta serumaren proteinak (laktoalbumina eta laktoglobulina), funtsezkoak dira esnekiak fabrikatzeko, hala nola gazta eta jogurta.

Proteinaren kantitateak eta kalitateak eragina dute esnearen propietate funtzionaletan eta nutrizionaletan.

2.Grasa:

Esnearen grasa energia-iturri esanguratsua da, eta gantz-azido esentzialak ditu. Esnekien testuran, zaporean eta egonkortasunean eragiten du.

3. Laktosa:

Laktosa esnearen karbohidrato nagusia da, eta energia-iturri garrantzitsua. Esnearen gozotasunari eta propietate osmotikoei eragiten die.

4. Zelula somatikoak:

Zelula somatikoak behien bular-guruinaren osasunaren eta esnearen kalitatearen adierazle da. Zelula somatiko kontzentrazio altuak mastitisa adieraz dezake, errapearen inflamazioa.

5. Gihar estraktu lehorra:

Gihar estraktu lehorrak esnearen osagai solido guztiak barne hartzen ditu, grasa izan ezik: proteinak, laktosa, mineralak eta bitaminak.

Funtsezkoa da esnekiak estandarizatzeko, hala nola esne gaingabetua eta esne-hautsa.

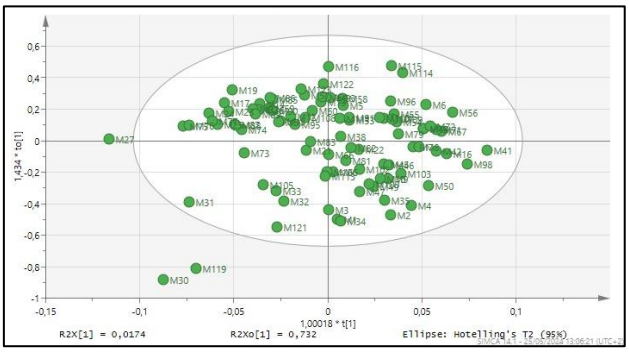
3.6 Datuen aldagai anitzeko analisian erabiltzen diren neurri eta erremintak

Datuen aldagai anitzeko analisia egiteko, erreminta eta neurri desberdinak erabiltzen dira. Hauek datuetatik informazio erabilgarria ateratzen laguntzen dute, haietatik baliozko interpretazioak egitea lortzeko eta ebaluatzeko.

Jarraian erreminta eta neurri garrantzitsuenak deskribatuko dira.

1.<u>"Score" grafikoa:</u>

"Score" grafikoak, egitura eta erlazioak dimentsionaltasun handiko datuetan bistaratzea ahalbidetzen du, interpretazioa eta erabaki-hartzea erraztuz.



***11**. Irudia: "Score" grafikoa*

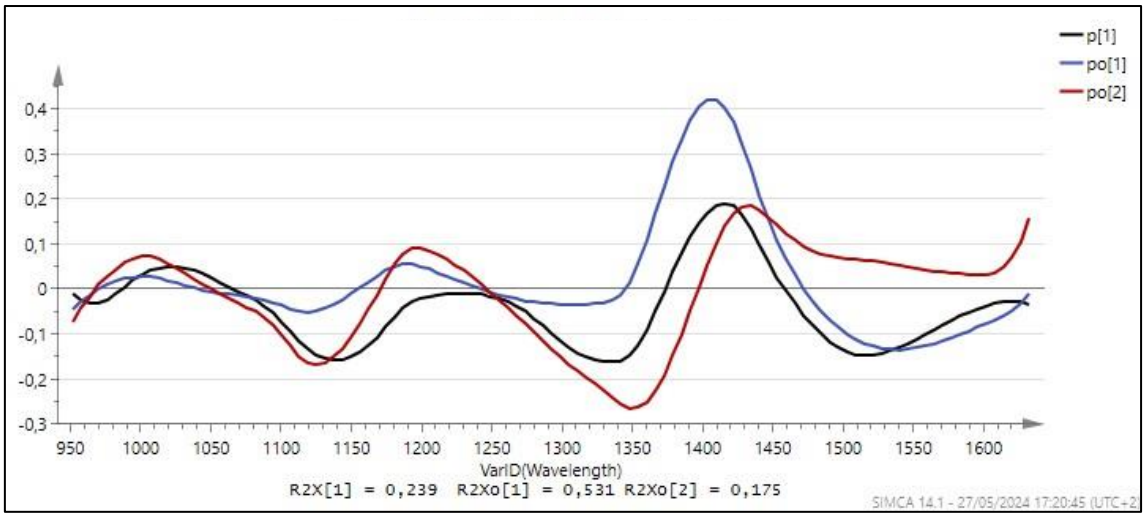
"Score" grafikoaren ardatzek aldagai anitzeko analisiak sortutako osagai nagusiak adierazten dituzte. Grafikoko puntu bakoitzak jatorrizko datu-multzoaren behaketa edo lagin bat adierazten du.

Grafikoko puntu bakoitzaren posizioak, osagai nagusien espazioan duen kokapen erlatiboa adierazten du. Grafikoan hurbilen dauden puntuek datu-profil antzekoagoak dituzte jatorrizko aldagaiei dagokienez.

2. <u>"Loading" grafikoa:</u>

"Loading" grafikoak, aldagaiek osagai nagusiei egiten dieten ekarpenari buruzko informazioa ematen du, datuetan oinarritutako interpretazioa erraztuz.

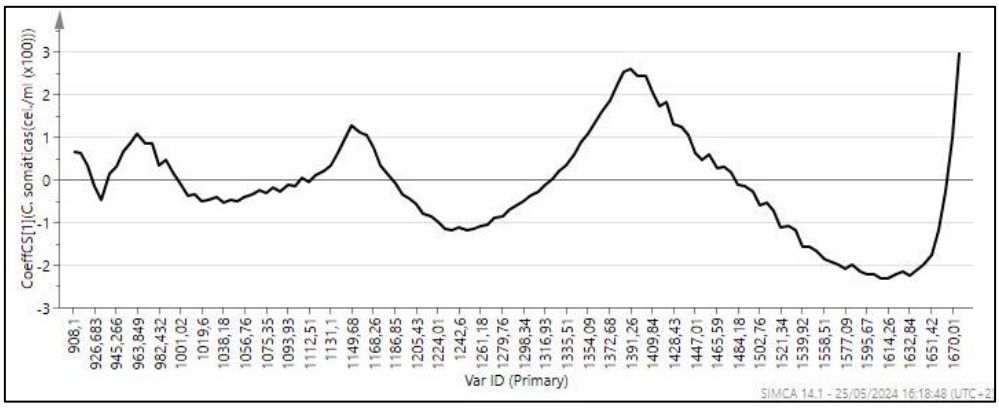
Adibide honetan esaterako, 3 osagai daude, eta hauek aldagai prediktiboari egiten dioten ekarpena espektro osoan zehar ikus daiteke:



***12**. Irudia: "Loading" grafiko baten adibidea*

3. Koefizienteen grafikoa:

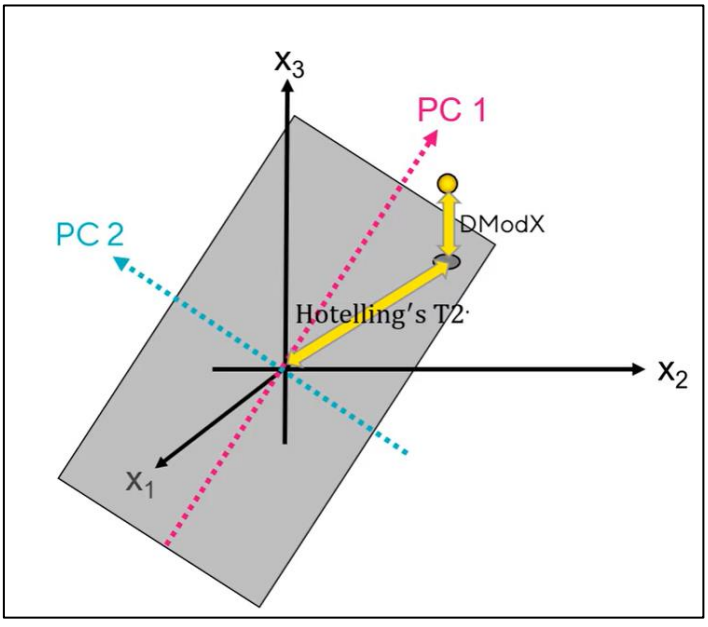
Ereduaren emaitzei eragiten dieten aldagai garrantzitsuenak identifikatzen laguntzen du. Bertan modeloaren osagai edo erregresio-koefiziente guztiek, hautatutako Y aldagairako sortzen duten iragarpen-bektorea irudikatzen da.



***13**. Irudia: Koefizienteen grafiko baten adibidea*

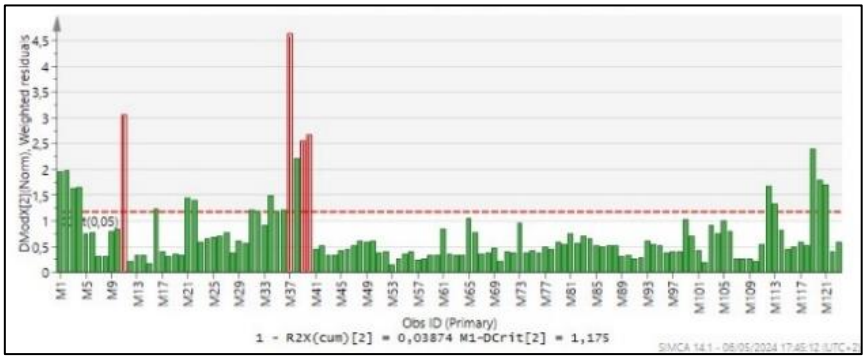
4. DMODX eta Hotelling's T2 neurriak:

Kalitate-kontrolean eta anomalien detekzioan erabiltzen diren bi estatistika garrantzitsu dira. Estatistika hauen grafikoek ikuspegi argi bat ematen dute outlier-ak identifikatzeko eta ereduaren osotasuna ziurtatzeko.



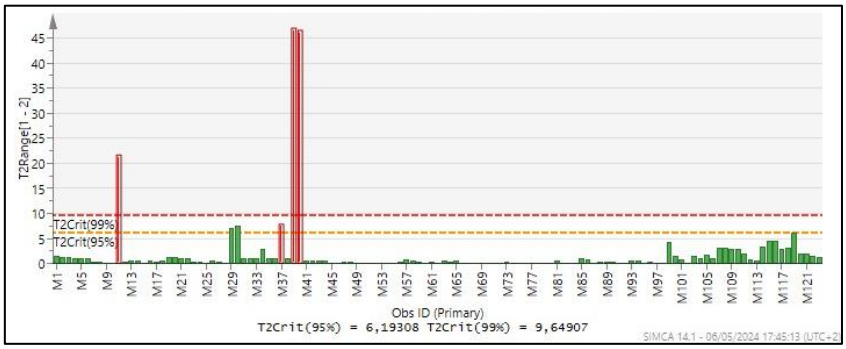
***14**. Irudia: DMODX eta Hotelling's T2 neurriak*

DMODX neurriak aldagai prediktiboen espazioan, behaketa espezifiko batetik eredura dagoen distantzia neurtzen du. Behaketa bat ereduari zein ondo egokitzen zaion ebaluatzen du.



***15**. Irudia: DMODX grafiko baten adibidea*

Hotelling's T2 neurriak osagai nagusien espazioan, behaketa batetik ereduaren zentrorako distantzia neurtzen du. Osagai nagusien bariantza eta kobariantza kontuan hartzen ditu.



***16**. Irudia: Hotelling's T2 grafiko baten adibidea*

Bi neurrietan, konfiantza-maila batean oinarritutako atalasea ezartzen da (normalean, %95) behaketa atipikoak identifikatzeko.

5. <u>Behatutako Vs Iragarritako grafikoa:</u>

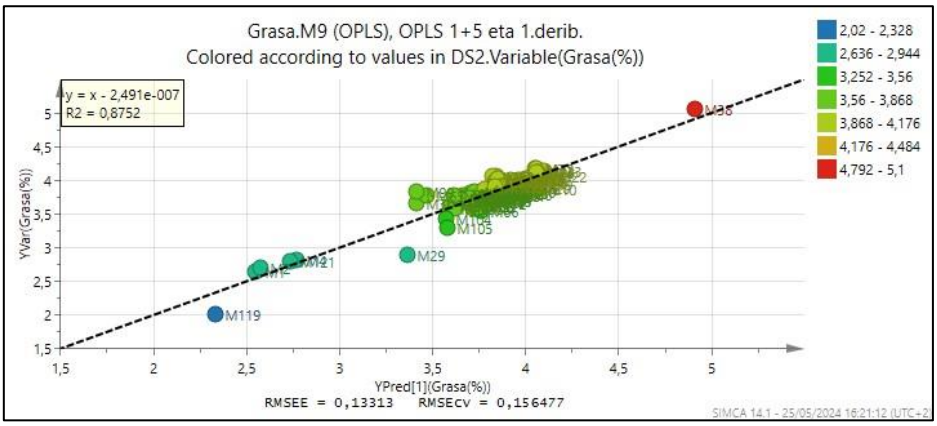
Modeloen zehaztasuna eta kalitatea ebaluatzeko tresna baliotsua da eta funtsezkoa da hainbat aplikaziotan eredu prediktiboen sendotasuna balioztatzeko eta hobetzeko.

Ardatz bertikalak, edo ordenatu-ardatzak, hautatutako aldagaiaren neurtutako (benetako) balioak adierazten ditu eta

ardatz horizontalak, edo koordenatu-ardatzak, modeloak aldagai horretarako iragarritako balioak adierazten ditu.

Iragarpena egiteko erabiltzen diren laginen arabera, bi grafiko desberdin daude:

1- **Entrenamendurako erabilitako laginen bidez eta balidazio-gurutzatuaren bidez egindako iragarpenak.**
2- **Modeloarentzako ezezagunak diren lagin berrien bidez egindako iragarpenak.**



***17**. Irudia: Behatutakoa Vs Iragarritako grafiko baten adibidea*

6. <u>R2X eta R2Y</u>:

R2X, ereduaren osagaiek azaltzen duten, aldagai iragarleen (X) aldakuntza-kantitatearen neurria da. Sarrera-datuetako aldakortasun osoaren zer ehuneko azal daitekeen ereduan hautatutako osagaiak erabiliz, adierazten du. R2X-k, modeloko osagaiek azaltzen duten X aldagaien aldakuntza-kantitatea adierazten du.

R2Y, ereduaren osagaiek azaltzen duten, erantzun-aldagaien (Y) aldakuntza-kantitatearen neurria da. Horrek erakusten du ereduak zein ongi iragar ditzakeen Y aldagaiak X aldagaietan oinarrituta. R2Y-k X aldagaietatik eratorritako osagaiek (X) azaltzen duten Y aldagaien bariantza-proportzioa adierazten du.

R2X balio altu batek, ereduko osagaiek, X aldagaien aldakuntzaren zati handi bat azaltzen dutela adierazten dute. Horrek esan nahi du eredua ondo atzematen ari dela X aldagaien egitura.

R2Y balio altu batek, ereduko osagaiek erantzun-aldagaien (Y) aldaketaren zati handi bat azaltzen dutela adierazten dute, eta horrek iradokitzen du ereduak aurreikusteko gaitasun ona duela.

7. Q2:

Q2 funtsezkoa da eredu baten sendotasuna eta datu berrietara duen gaitasuna balioztatzeko. R2X-k eta R2Y-k modeloak entrenamendu-datuen bariantza ondo azaltzen duela neurtzen dute; Q2-k, berriz, inoiz ikusi gabeko datuei buruzko jarduera prediktiboaren ebaluazio zuzena ematen du.

8. RMSEcv eta RMSEP:

RMSEcv-k, modeloaren aurreikusteko gaitasuna ebaluatzen du, balidazio gurutzatuaren bidez, entrenamendu-datuen multzoaren barruan.

RMSEP-k, modeloak inoiz ikusi gabeko datu berriak eta entrenamenduan eta balidazio gurutzatuan erabili ez direnak, aurreikusteko duen gaitasuna neurtzen du.

Bi metrikak erabakigarriak dira iragarpen-modelo baten kalitatea eta sendotasuna ebaluatzeko; entrenamendu-datuetara ondo egokitzeaz gain, ikusi gabeko datuetara ere behar bezala orokortzen dela ziurtatuz.

4.ONURAK

Atal honetan, proiektu honek eskaintzen dituen onura tekniko, sozial eta ekonomikoak aurkeztuko dira.

4.1 Onura teknikoak

Onura teknikoen artean, analisi hau egiteko erabiltzen diren teknikak ez-suntsitzaileak direla dago. Beste metodo batzuekin konparatuz, NIR espektroskopiak ez ditu laginak suntsitzen, lagin berdinarekin proba bat baino gehiago egitea ahalbidetzen duena adibidez.

Beste onura bat neurketak egiteko bizkortasuna eta efizientzia da, laginen preparazio konplexuaren beharrik gabe.

Beste alde batetik, neurketa bakarrarekin, hainbat parametro fisiko-kimiko aldi berean analizatzeko gai da, ez dira neurketa desberdinak egin behar.

4.2 Onura sozialak

Onura sozialei dagokienez, elikagaien kalitatea eta segurtasuna bermatzen laguntzen duenak garrantzi handia du. Kritikoak diren parametroen balioak bizkor jakiteak, esnekiak kalitatezkoak eta seguruak izatea bermatzen laguntzen du.

Beste onura bat osasun publikoari egiten dion ekarpena da. Batez ere, zelula somatikoen kontzentrazio altuak dituzten laginak identifikatuz, gaixotasunak ekar ditzakeenak, biztanleriaren osasun orokorrari laguntzen zaio.

4.3 Onura ekonomikoak

Onura ekonomikoetatik garrantzitsuena, metodo tradizionalekin konparatuta gastuak txikiagotzen dituela da. Metodo honekin momentuan jakin ahalko da parametroren bat segurtasun atalase batetik kanpo dagoen. Kasu horretan produkzioa gelditzeko aukera

egongo da, irtenbide bat bilatzeko, lote oso bat alperrik galtzea ekidituz.

5.ALTERNATIBEN ANALISIA

Jarraian, proiektua aurrera eramateko egon diren alternatiba ezberdinak analizatuko dira.

5.1 Aldagai anitzeko datuen analisia egiteko software-a

5.1.1 Simca

Simca, Sartorius Stedim Data Analytics-ek garatutako software espezializatu bat da, aldagai anitzeko analisietan eta esperimentuen diseinuan espezializatua [16].

Bere abantailen artean honako hauek ditugu:

- **Interfaze erabilterraza:**
  Simca-k oso intuitiboa den interfaze grafikoa du, eta horrek bere erabilera asko errazten du. Ez da beharrezkoa programazioan edo estatistikan esperientzia handia edukitzea [16].

- **Kimiometria Espezializazioa:**

  Kimiometria aplikazioetarako espreski diseinatua dago. Hori dela eta, egokia da elikagaien eta farmazien industrian analisiak egiteko.

- **Datuen integrazioa:**
  Hainbat iturritako datuak integratzeko eta datu-bolumen handiak eraginkortasunez erabiltzeko gaitasuna du [16].

Desabantailei dagokienez:

- **Malgutasun mugatua:**
  Beste programa batzuekin alderatuta malgutasun gutxiago du analisi espezifikoak pertsonalizatzeko edo funtzionalitate berriak garatzeko.

- **Interfazearen mendekotasuna:**
  Erabiltzaile aurreratuek mugak aurki ditzakete interfaze grafikoan, beste programa batzuek eskaintzen dituzten programazio-gaitasunekin alderatuta.

- **Kostua:**

  Simca-ren lizentzia garestia da, 5000€ inguruko kostua du.

5.1.2 Matlab-en PLS_Toolbox

Matlab, kalkuluak egiteko diseinatutako lengoaia da. Kalkulua, bisualizazioa eta programazioa modu praktiko eta erraz batean uztartzen dituena. PLS_Toolbox, Eigenvector Research-ek garatutako aldagai anitzeko analisirako tresna multzoa da, eta Matlabekin integratzen da [17].

Bere abantaila anitzen artean:

- **Malgutasuna eta Pertsonalizazioa:**
  Matlabek oso ingurune malgua eskaintzen du, pertsonalizazioa eta algoritmo espezifikoen garapena ahalbidetuz, erabiltzailearen beharren arabera.

- **Script eta Automatizazioa:**

  Programazioa nahiago duten erabiltzaileentzat ezin hobea da, prozesuak automatizatzeko eta analisi konplexuetarako script pertsonalizatuak garatzeko aukera ematen baitu.

- **Komunitatea eta baliabideak:**
  Dokumentazio ugari, tutorialak, eta arazoak konpontzen eta konponbideak partekatzen lagun dezaketen erabiltzaileen komunitate zabala du.

- **Beste produktu batzuekin integratzea:**
  Matlaben beste produktu eta toolboxekin erraz integratu daiteke, aztertzeko eta modelatzeko aukerak zabalduz.

  Desabantailei dagokienez:

- **Ikaskuntza-kurba:**
  Matlab-ek programazio ezagutza minimo bat eskatzen du, eta hori oztopo izan daiteke aldez aurretik esperientziarik ez duten erabiltzaileentzat.

- **Datuen aldagai anitzeko analisia egiteko interfazea:**
  Toolbox gehigarriak erabili beharretik aparte, ez dauka datuen aldagai anitzeko analisia jorratzen duen interfaze espezializatu bat, honen garapena eta ikerketa errazten dituena.

- **Kostua:**
  Matlab lizentzia batek 900€ inguruko kostua du.

5.3 Alternatiben analisiaren eraginkortasun taulak

Irizpideak	Simca	Matlab-en PLS_Toolbox
Interfazea(%15)	13	10
Erabilgarritasuna(%10)	9	7
Analisia(%45)	45	35
Funtzionaltasuna(%15)	9	14
Kostua(%10)	2	5
Komunitatea(%5)	3	5
Puntuazio totala (%100)	81	76

***1**. Taula: Aldagai anitzeko datuen analisia egiteko software-n eraginkortasun taula*

5.4 Hautatutako aukeraren arrazoiketa

Datuen aldagai anitzeko analisia egiteko duen interfaze espezializatu eta erabilterrazagatik, Simca hautatu da analisia egiteko software moduan.

Beraz, honen bidez egingo da NIR espektrometrotik lortutako laginen analisia eta haren bidez jorratuko da iragarri nahi diren behi-esnearen parametro fisiko-kimikoen modelizazioa.

6.PROPOSATUTAKO IRTENBIDEAREN DESKRIBAPENA

Esne-lantegi batetik jasotako behi-esne laginak neurtuko dira, NIR espektrometro baten bidez eta baita laborategi baten analisi kimiko baten bidez.

NIR espektrometroaren bidez lagin hauen infragorri hurbileko espektroa lortuko da. Laborategiko analisi kimikoaren bidez aldiz, iragarri nahi diren parametro fisiko-kimikoen balioak lortuko dira, hauek modeloen entrenamendurako erabiliko baitira.

Ondoren bi neurketa hauek uztartzen dituen dataset bat sortuko da, hau gero Simca programaren bidez erabilia eta prozesatua izateko. Espektrometrotik lortutako laginen espektro-balioak X aldagai modura definituko dira eta analisi kimikotik lortutako parametroen balioak Y aldagai modura.

Hemendik aurrera, lagin hauen datuen aldagai anitzeko analisiari ekingo zaio Simca programaren bitartez, behi esnearen hurrengo parametro fisiko-kimikoak modelatzeko eta aurreikusteko: Laktosa, grasa, proteina, zelula somatikoak eta gihar estraktu lehorra.

Datuen aldagai anitzeko analisian, Y aldagai bakoitzerako modelo eta osagai kopuru desberdinak probatuko dira. Modelo hauek ikertuko dira eta neurri ezberdinen bitartez baloratuko dira. Azkenean, aldagai bakoitzerako hobetoen egokitzen den eta iragarpen onena gauzatzen duen modeloa hautatuko da eta bere ezaugarriak deskribatuko dira.

7.ARRISKUEN ANALISIA

Arriskuen analisiaren bidez, proiektuan eragina izan ditzaketen mehatxuak eta aurreikusi ez diren gertaerak identifikatu eta aztertu egiten dira, baita horiek gertatzeko dagoen probabilitatea eta gertatzekotan eragin ditzaketen kalteak eta ondorioak ere.

7.1 IDENTIFIKATUTAKO ARRISKUAK

7.1.1 Epeen ez betetzea (A)

Planifikazio txar baten edota aurreikusita ez dauden eragozpenen ondorioz, batzuetan finkatutako epeak ez betetzea gerta daiteke. Hau ekiditeko, beharrezkoa da planifikazioa egiterako orduan beharrezkoak diren marjinak uztea; horrela, egon ahal diren atzerapenek ahalik eta arrisku gutxien edukitzeko.

7.1.2 Laginetako akatsak (B)

Proiektu honetan erabiltzen diren laginak, behi-esne laginak dira. Hauek produktu biologikoak izatean, kontu handiz zaindu behar dira bai neurtzen diren denborak eta baita mantentze lanak ere. Hauen parametro fisiko-kimikoak neurtzen dira; beraz, egoera estandarretik kanpo dauden laginek parametro hauen balioak akastu ditzakete.

7.2 ARRISKUEN EBALUAZIOA

Arriskuen ebaluazioa probabilitate-eragin matrizearen bidez egin da.

		Eragina		
		Baxua 0,2	**Ertaina 0,5**	**Altua 0,8**
Probabilitatea	**Baxua 0,2**	**0,04**	**0,1**	**0,16 (A)**
	Ertaina 0,5	**0,1**	**0,25 (B)**	**0,4**
	Altua 0,8	**0,16**	**0,4**	**0,64**

***2**. Taula: Arriskuen probabilitate-eragin matrizea*

LANERAKO ERABILITAKO METODOLOGIA

1.EGINBEHARREKOAREN DESKRIBAPENA, FASEAK ETA PROZEDURA

Ahalik eta emaitza hoberenak lortzeko xedearekin, lanerako metodologia egoki eta eraginkor bat diseinatu da. Ikerketaren garapenaren ikuspegi sistematiko eta sekuentzial bat emango da, fase sekuentzial batzuen finkapenaren bidez. Honi esker, ordena logiko eta zentzuzko bat jarraitu ahal izango da prozesu osoan zehar. Ikerketa honen garapenerako, hurrengo fase hauek finkatu dira:

1. **Sartorius-ek erraztutako tutorialen bisualizazioa, Simca programa erabiltzen ikasteko eta datuen aldagai anitzeko analisiak zertan datzan ikasteko.**

2. **Simca-ren adibide praktikoen jorrapena, programarekin trebatzeko eta hauen txostenen burutzea, lengoaiarekin ohitzeko eta kontzeptuak finkatzeko.**

3. **NIR espektroskopia eta datuen aldagai anitzeko analisiari buruzko bilaketa bibliografikoa.**

4. **Laktosa aldagaiaren modelizazioa.**

5. **Grasa aldagaiaren modelizazioa.**

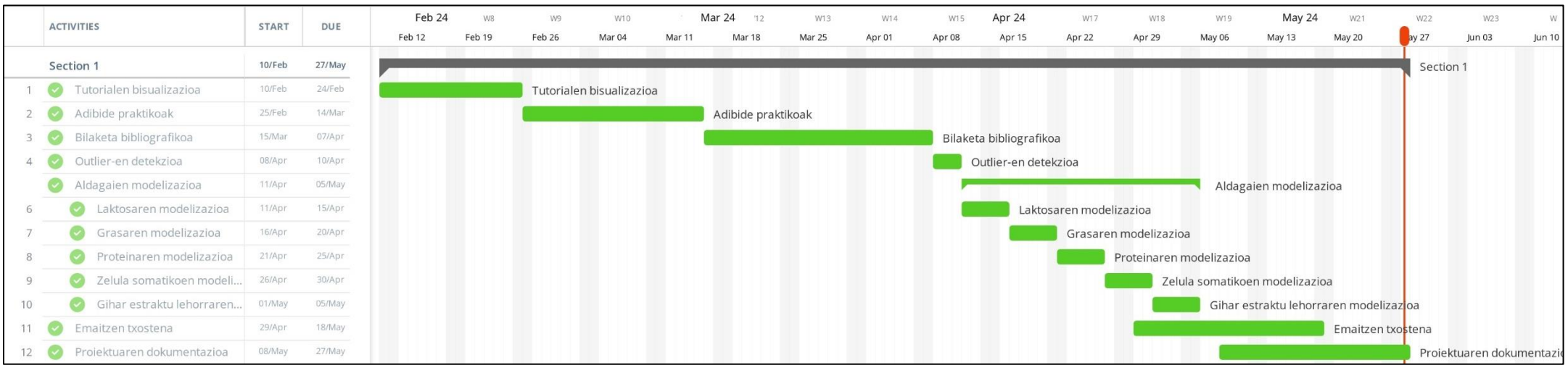
6. **Proteina aldagaiaren modelizazioa.**

7. **Zelula somatiko aldagaiaren modelizazioa.**

8. **Gihar estraktu lehorra aldagaiaren modelizazioa.**

9. **Emaitzen txostenaren burutzea.**

2.GANTT-EN DIAGRAMA/KRONOGRAMA



18. Irudia: Gantt-en diagrama

3.HARDWARE ETA SOFTWARE BALIABIDEAK

Jarraian erabilitako software eta hardware baliabideak zerrendatuko dira.

Materiala	Kopurua
Lenovo ordenagailu eramangarria	1
Simca lizentzia	1
Microsoft Office 365 lizentzia	1

3. Taula: Erabilitako software eta hardware baliabideak

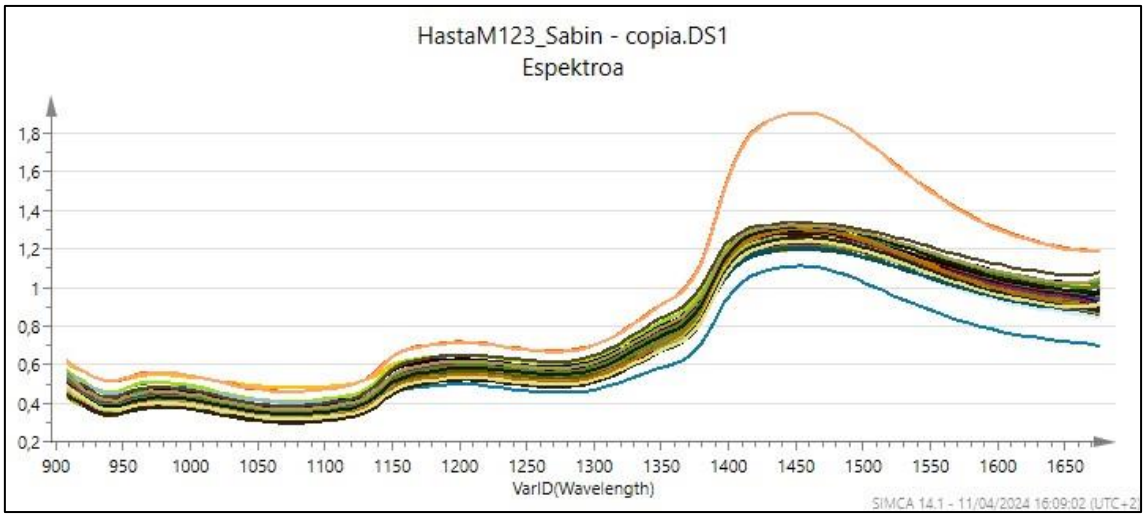
4.EMAITZEN DESKRIBAPENA

4.1 Espektro analisia – X aldagaiei dagokien outlierren detekzioa

Lehendabizi, Y aldagaien analisian sartu baino lehen, X aldagaiak aztertu dira; hau da, lagin bakoitzak espektro osoan dituen balioak. Espektroa 908,1 – 1676,2 nm artekoa da eta 6,2 nm bakoitzeko balio bat neurtzen da, hauetako balio bakoitza X aldagai bat izanik.

Erabilitako dataset-a 123 esne lagin desberdinez osatua dago, eta lagin bakoitzeko espektro osoko balioak eta bost Y aldagai fisiko-kimikoren balioak neurtu dira: Proteina, grasa, laktosa, zelula somatikoak eta ESM("Extracto Seco Magro").

Hau da espektroaren itxura:



***19**. Irudia: Behi-esne laginen NIR espektroa*

Ikus daitekeenez, badaude lagin gutxi batzuk gainontzekoetatik bereizten direnak, portaera desberdin bat erakusten dute. Hauek, arriskutsuak izan daitezke modeloak sortzerako orduan, berauek distortsionatu baititzakete. Horregatik hauen izaera aztertzea oso garrantzitsua da, kasu batzuetan outlierrak izan ahal baitira.

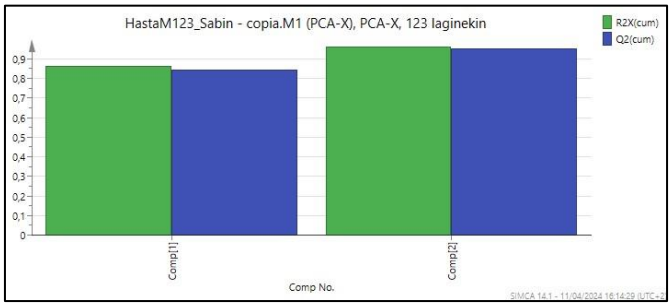
X aldagaiei dagozkien outlierren detekziorako osagai nagusien analisia edo Principal Components Analysis (PCA-X) modelo bat sortu da, 123 esne laginei dagozkien X aldagaiez osatua, Y

aldagaiak bazter batera utziz. PCA oso erabilia da datu multzoen bistaraketarako, haietatik ezaugarriak ateratzeko eta ikuspegi zabal bat lortzeko, aldagaien arteko erlazio linealen erabileraz.

PCA-n metodorik erabiliena lehen bi osagai nagusiena da. Osagai hauek datuetan ahalik eta aldakortasun handiena antzematen dute, datuak bi dimentsioetako espazio batean bistaratzea ahalbidetuz. Hau oso erabilgarri da dimentsionaltasun handiko datuak bistatzerako orduan, irudikapen bidimentsionalaren bitartez datuak bistaratzea eta hauen ezaugarriak ulertzea errazagoa baita.

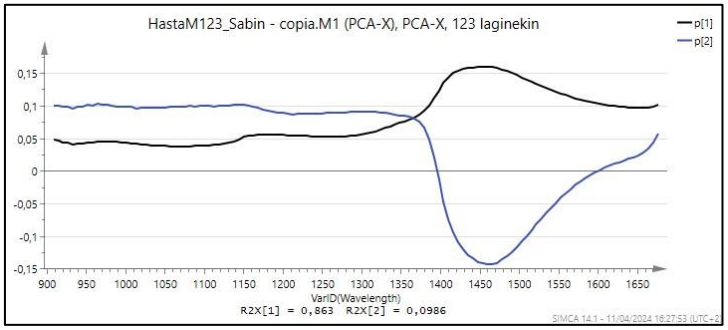
Modelo honekin lortu ditugun R2X eta Q2 balioak, 0,961 eta 0,954 dira, hurrenez hurren.

Beheko irudian ikus daitekeen moduan, balio hauen pisu handiena lehenengo osagai nagusiak darama.



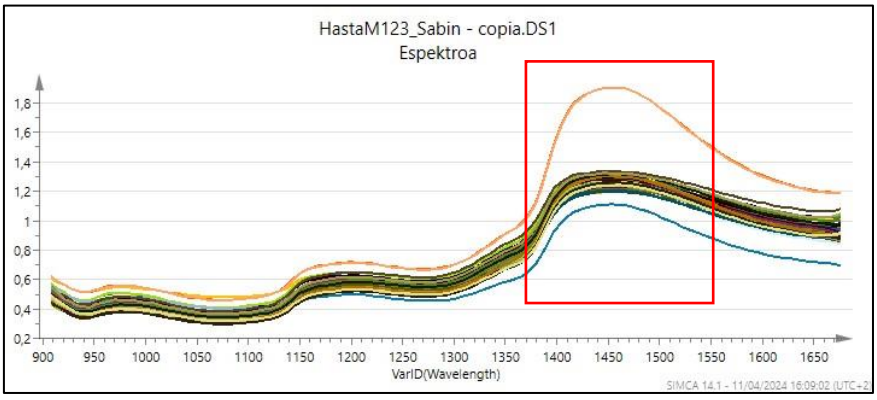
***20**. Irudia: Lehenengo bi osagai nagusien R2X eta Q2*

Osagai nagusiek espektroaren zein zatitan duten influentzia gehiago ikustea ere oso baliagarria da. Loading grafikoaren bitartez lehenengo bi osagai nagusiak irudikatu dira:



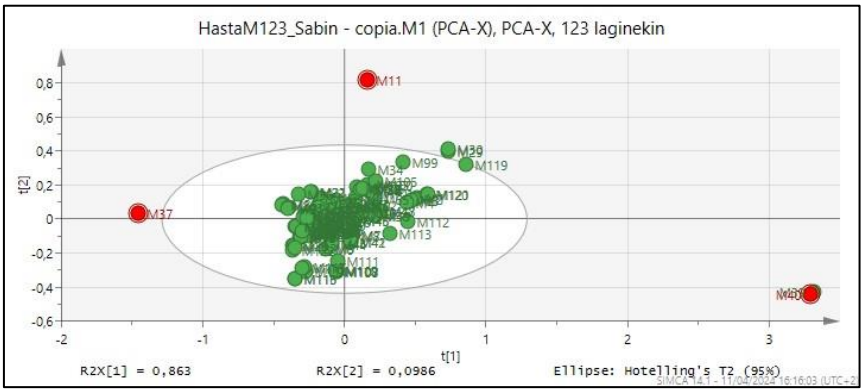
***21**. Irudia: Lehenengo bi osagai nagusien Loading grafikoa*

Lehen esan den bezala, lehenengo osagaia da pisu handiena duena eta bertan ikus daiteke honek influentzia handiena duen gunea 1400-1500 nm arteko bandan dela. Honek esan nahi du, lehenengo osagai nagusiak datuen banda horretako aldakortasuna antzematen duela. Lagin guztien espektroari begiratuz, ikus daiteke banda hau bat datorrela laginen arteko desberdintasun handiena duen gunearekin.



***22**. Irudia: Laginen NIR espektroa, koadro gorrian desberdintasun gehien dagoen espektro zatia*

"Score" grafikoaren bitartez laginen ikuspegi bidimentsional bat lortzen da:

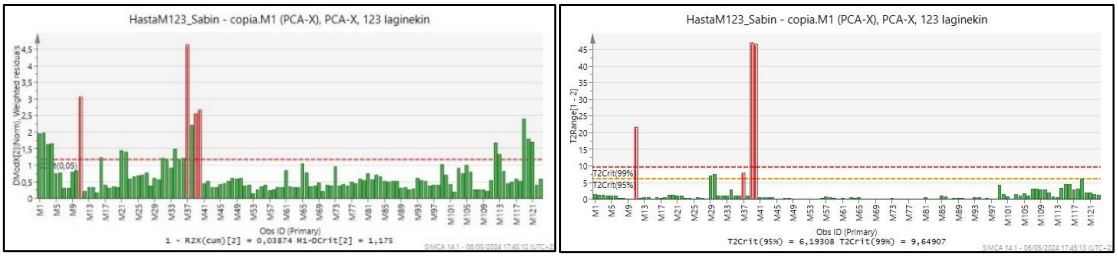


***23**. Irudia: "Score" grafikoa*

Bertan argi ikus daiteke badaudela zenbait lagin bai tolerantzia-elipsetik oso urrun daudenak (11,39 eta 40 laginak) edo gainontzeko laginengatik oso aparte daudenak (37 lagina).

"DMODX" eta "Hotelling's T2" grafiken bitartez, oso baliagarriak diren neurri estatistikoak lortzen dira, outlierren detekzioan lagungarriak direnak.

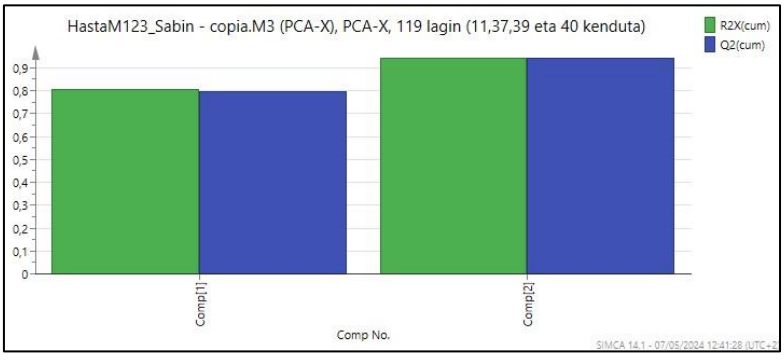
Beheko bi grafiketan ikus daitekeenez, aipatutako laginek balio oso altuak dituzte bi neurrietan. Lagin batek bai neurri batean zein bestean %99-ko konfiantza-maila gainditzeak, outlierra izateko aukera oso handiak dituela esan nahi du. Eskumako grafikan ikus daiteke 37 laginak ez duela "Hotelling's T2" neurriaren %99-ko konfiantza-maila gainditzen (gertu geratzen da); hala ere, hau nahikoa ez balitz, outlierra delakoaren susmoa handiagoa da, honi lehen ikusitako Score grafikoan duen posizioa gehitzen badiogu, non gainontzeko laginetatik urrun eta isolatuta ageri baitzen eta tolerantzia-elipsetik kanpo.



24. Irudia: DMODX eta Hotelling's T2 neurriak

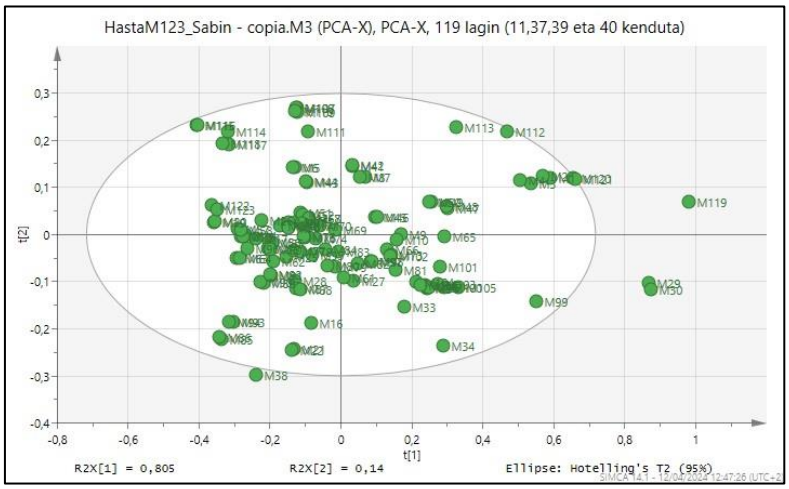
Beraz, analisi honen bitartez, dataset honetan 11,37,39 eta 40 balioak outliertzat hartuko dira eta modelo guztietatik kanpo geratuko dira.

Jarraian beste PCA-X modelo bat sortuko da aipatutako lau outlierrak kenduta:



***25**. Irudia: Lehenengo bi osagaien R2X eta Q2 balioak outlier-ak kenduta*

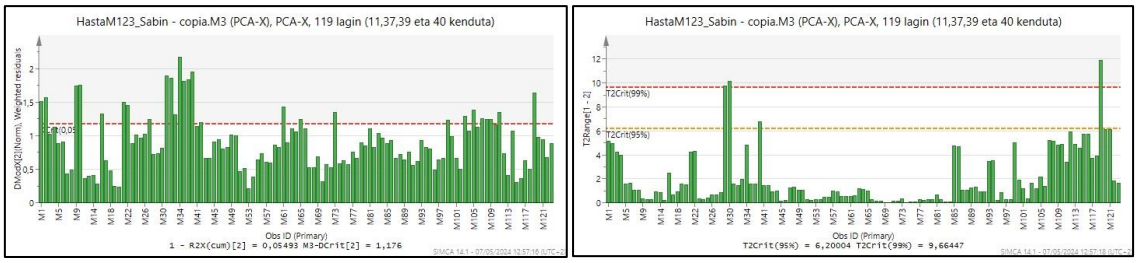
Score grafikoa irudikatuz, lagin hauen distribuzio bidimentsionala aztertzen da:



***26**. Irudia: "Score" grafikoa outlier-ak kenduta*

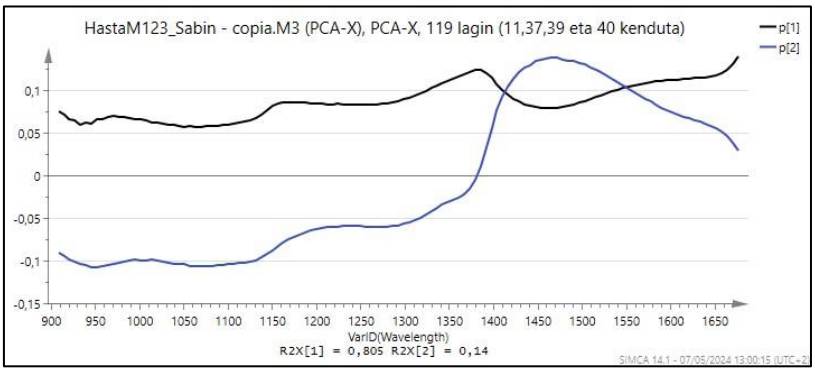
Ikus dezakegu orain aurrekoan baino distribuzio uniformeagoa dagoela, eta nahiz eta zenbait lagin tolerantzia-elipsetik kanpo egon, esan daiteke distantzia ez dela gehiegizkoa eta asumitu daiteke ez duela modeloa kaltetuko.

"DMODX" eta "Hotelling's T2" neurriei begiratuz, ikus daiteke 119 laginak bietan balio altuak dituela, baina ez da gehiegizkoa gainontzekoekin konparatuz eta bertan uztea erabaki da.



27. Irudia: DMODX eta Hotelling's T2 neurriak outlier-ak kenduta

Lehenengo bi osagai nagusiei begiratuz, aldaketaren bat ikus daiteke aurrekoarekin konparatuz.



***28**. Irudia: Lehenengo bi osagai nagusien Loading grafikoa outlier-ak kenduta*

Oraingoan bigarren osagaiak garrantzi handia hartu du 1400-1550 nm bandan, eta honen aldakuntza antzemateaz arduratzen da.

Beraz, PCA-X analisitik, 11,37,39 eta 40 laginak outlierrak direla ondorioztatu da eta gure modeloak ez kaltetzeko, hoberena hauek modeloetatik kanporatzea dela erabaki da.

Hala ere, hauek ez dira zertan outlier bakarrak izan behar, hauen X aldagaiei dagozkienak baino ez baitira. Jarraian, Y aldagai bakoitzerako, modeloak sortu baino lehen, Y aldagai horri dagokion outlierrik badagoen aztertu beharko da eta beharrezkoa denean modeloetatik kendu.

4.2 Y aldagaien analisia eta iragarpen modeloak

Jarraian Y aldagai hauen analisia jorratuko da: Laktosa, grasa, proteina, zelula somatikoak eta gihar estraktu lehorra (ESM gaztelerako sigletan, "Extracto Seco Magro").

Y aldagai bakoitza indibidualki aztertuko da X aldagaiekin batera eta modelo eta osagai kopuru desberdinak probatuko dira. Modeloak aldagai eta laginetara egokitzeko kapazitatea eta lagin berriak iragartzeko ahalmena ebaluatuko dira, horrela aldagai bakoitzerako modelo egokiena aukeratuz.

Iragarpenak egiteko eta modelo bakoitza balidatzeko, modeloak ikusi ez dituen laginak behar ditugu; beraz, gure dataset-eko laginetatik, %80-a entrenamendurako erabiliko dira eta %20-a iragarpenak egiteko. Lagin hauek ausaz hautatuko dira eta Y aldagai bakoitzerako zenbait aldiz errepikatuko da prozesua, lagin multzo desberdinekin kontrastatzeko.

4.2.1 Laktosa

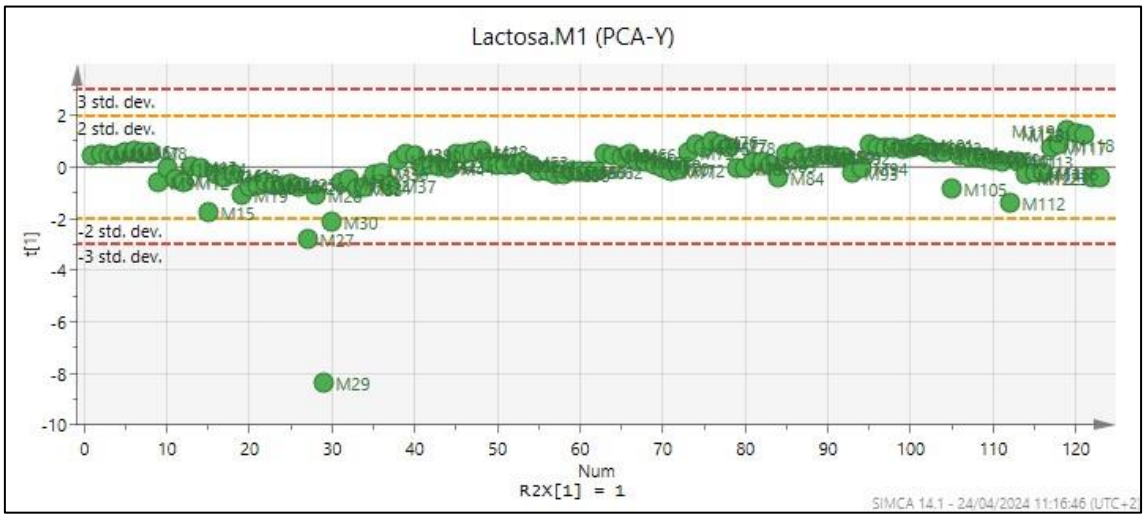
Lehendabizi laktosa aldagaia aztertuko da.

Normalean laktosaren kontzentrazioa behi-esnean %4,7-4,8 ingurukoa da. Erabilitako dataset-ean lagin gehienen laktosa balioak %4,6-4,9 tartean daude, nahiz eta badauden zenbait lagin balio hauetatik nahikoa desbideratzen diren.

Proba hauetarako gainontzeko Y aldagai guztiak kendu dira Laktosa aldagaia izan ezik.

Lehenengo eta behin PCA-Y modelo bat sortu da lagin guztiekin, ea aldagai honi dagokion beste outlierrik dagoen ikusteko.

Ondorengo Score grafikoa lortu da:



***29**. Irudia: Laktosaren PCA-Y modeloaren "Score" grafikoa*

Bertan argi ikus daiteke 29 lagina gainontzekoetatik oso desbideratuta dagoela eta tolerantzia lerrotik oso urrun dagoela. Beraz, argitasun osoz esan daiteke lagin hau outlier bat dela.

Modu honetan, X aldagaiei dagokien outlierrei(11,37,39 eta 40 laginak), azken hau gehitzen zaie eta laktosaren analisiari dagokien modeloetatik kenduko da ere.

Hainbat modelo aztertu dira, aldagai kopuru ezberdinekin, non Q2 neurria maximizatzea eta RMSEP neurria minimizatzea bilatu den, beti ere zarata modelizatzea ekidituz.

Laktosarentzako, balio onenak lortu dituen modeloa hau izan da:

OPLS modelo bat osagai prediktibo batekin eta hiru osagai ortogonalekin, bigarren deribatua egiten duen iragazki batekin.

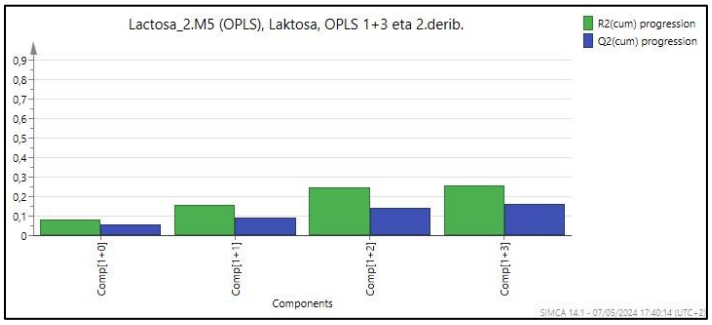
Hurrengo balioak lortu dira modelo honetarako:

R2X: 0,971 ; R2Y: 0,258 ; Q2: 0,159

Type: OPLS Observations (N)=94, variables (K)=112 (X=111, Y=1)

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,971			0,258			0,159		1		
Predictive		0,14			0,258			0,159		1		
P1	0,14	0,14	13,1	0,258	0,258	0,159	0,01	0,159	1	1	1	R1
Orthogonal in X(OPL...		0,831			0							
O1	0,592	0,592	55,6	0	0							R1
O2	0,135	0,727	12,7	0	0							R1
O3	0,104	0,831	9,78	0	0							R1

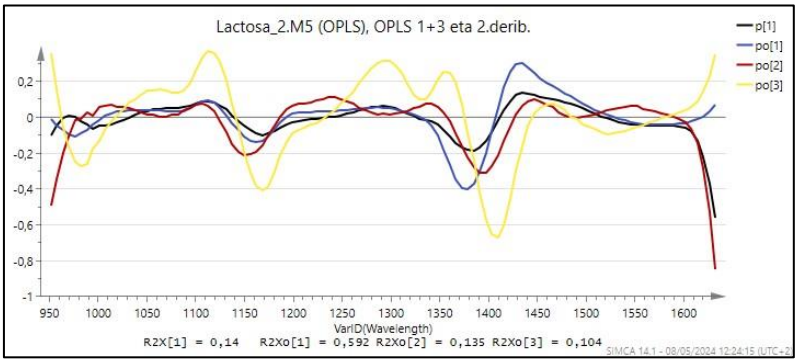
***30**. Irudia: Laktosaren OPLS modeloaren R2X, R2Y eta Q2 balioak*



***31**. Irudia: Laktosaren OPLS modeloaren osagaiak*

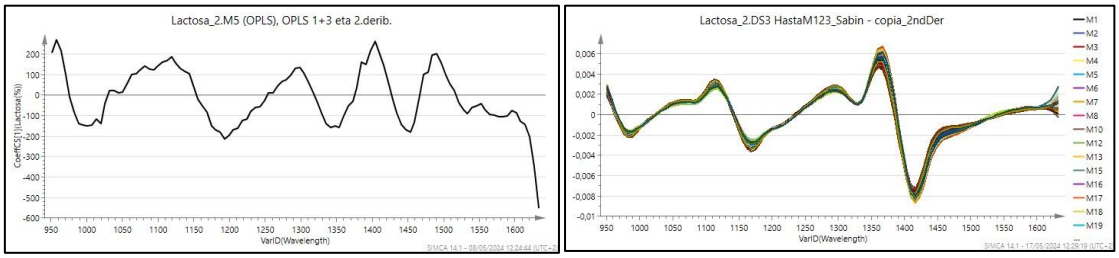
Aldagai gehiago ez erabiltzearen arrazoia, aldagai gehiagorekin modeloa ez hobetzeaz gain, okertuz doalako eta gainera konplexutasuna gehitzen diolako da. Aldi berean ere, modeloa zarata modelizatzen hasiko litzateke osagai gehiago gehitzen goazen heinean.

Beraz, esan bezala, aukeratutako modeloak 4 osagai ditu, prediktibo bat eta hiru ortogonal. Loading grafikoaren bitartez osagai hauen forma indibidualki ikus dezakegu espektro osoan:



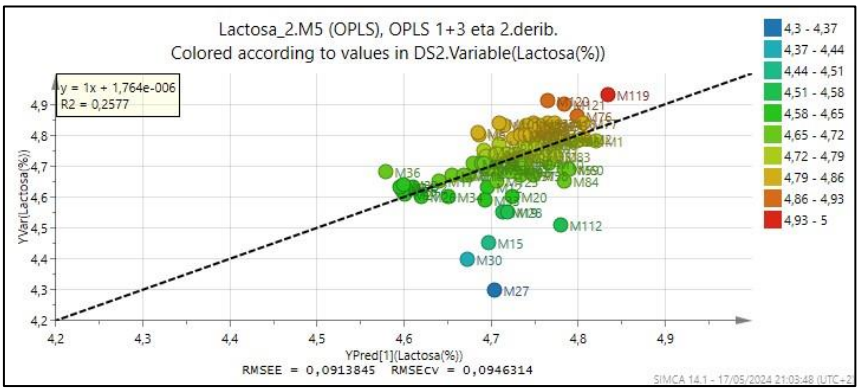
***32**. Irudia: Laktosaren OPLS modeloaren osagaien Loading grafikoa*

Beheko ezkerreko irudian, osagai guzti hauek batera laktosa aldagairako sortzen duten iragarpen bektorea ikus dezakegu. Honek forma hau izatearen arrazoia, 2.deribatuak espektroa aldatzen duelako da, eskumako irudian ikus daitekeen moduan.



33. Irudia: Laktosaren OPLS modeloaren iragarpen bektorea (ezkerra) eta espektroa bigarren deribatuarekin (eskuma)

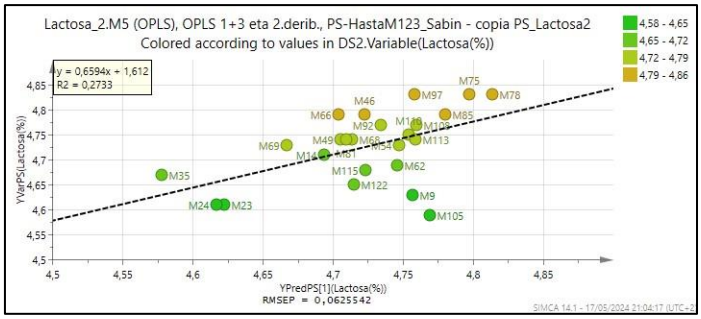
Behatutakoa Vs Iragarritakoa grafikoa, entrenamenduko laginen balidazio gurutzatuaren bidez:



***34**. Irudia: Laktosaren OPLS modeloaren Behatutako Vs Iragarritakoa grafikoa entrenamenduko laginekin*

Esan daiteke lagin gehienak erregresio lerroarekin bat datozela, nahiz eta ikus daitekeen badaudela zenbait lagin nahikoa desbideratzen direnak, errorea handituz. Balidazio gurutzatuaren bidez lortutako iragarpenen errorea (RMSEcv – Root Mean Squared Error Cross Validation) %0,0946 da. Kontuan izanda dataset-eko laginen balio tartea %0,6-koa dela, RMSEcv balio honek %15,77-ko errorea dakar balio tarte honekiko.

Honen ostean, modeloak ikusi ez dituen lagin berriak sartzen dira, modeloak haien laktosa kontzentrazioa aurreikusteko duen kapazitatea ikusteko. Guztira 24 lagin berriren iragarpena egin da eta honako hau izan da emaitza:



***35**. Irudia: Laktosaren OPLS modeloaren Behatutako Vs Iragarritakoa grafikoa lagin berriekin*

Ikus daitekeenez, laginak nahiko ondo egokitzen dira erregresio lerrora. Iragarpen errorea (RMSEP – Root Mean Squared Error Prediction) %0,0625-ekoa da, balidazio gurutzatuarekin lortutako

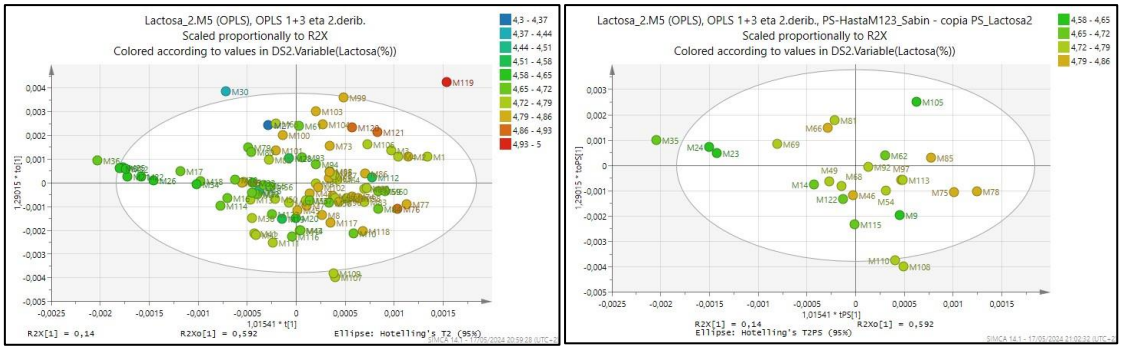
errorea baino txikiagoa. Kontuan izanda iragarpenerako erabilitako laginen balio tartea %0,25-koa dela, RMSEP balio honek %25-eko errorea dakar balio tarte honekiko.

Hurrengo taulan modeloak iragarri dituen balioak eta benetakoak dituen zerrenda bat ikus daiteke lagin bakoitzerako:

	1	2	3
1	Obs ID (Primary)	M5.YVarPS(Lactosa(%))	M5.YPredPS[1](Lactosa(%))
2	M9	4,63	4,75679
3	M14	4,71	4,69346
4	M23	4,61	4,62228
5	M24	4,61	4,61652
6	M35	4,67	4,57781
7	M46	4,79	4,72254
8	M49	4,74	4,70503
9	M54	4,73	4,74681
10	M62	4,69	4,7458
11	M66	4,79	4,70405
12	M68	4,74	4,71366
13	M69	4,73	4,66681
14	M75	4,83	4,7967
15	M78	4,83	4,81358
16	M81	4,74	4,70915
17	M85	4,79	4,77971
18	M92	4,77	4,73384
19	M97	4,83	4,75765
20	M105	4,59	4,7691
21	M108	4,77	4,75935
22	M110	4,75	4,75357
23	M113	4,74	4,75873
24	M115	4,68	4,72328
25	M122	4,65	4,71498

***36**. Irudia: Laktosaren OPLS modeloak iragarri dituen balioak eta benetakoak*

Jarraian, lagin berri hauek modeloan duten distribuzio bidimentsionala aztertuko da. Beheko irudian entrenamenduko laginen distribuzioa ikusten da ezkerrean eta iragarpenerako erabilitako laginena eskuman. Laginak laktosa balioen arabera koloreztatuta daude.



37. Irudia: Laktosaren OPLS modeloaren entrenamendurako (ezkerra) eta iragarpenerako (eskuma) erabilitako laginen distribuzio bidimentsionalak

Ikusten da nola ia espazio osoan ez dagoen ordena zehatzik. Laktosa kontzentrazioa ez da osagai prediktiboaren norabidean,

ardatz horizontalean, ondo banatuta ikusten. Beraz, esan daiteke osagai prediktiboa ez dela gai laktosa aldagaiaren aldakortasuna ondo hautemateko. Ikusten da lagin berrien distribuzioa entrenamendu laginen antzekoa dela; beraz, bere egiturarekin bat datorrela esan daiteke.

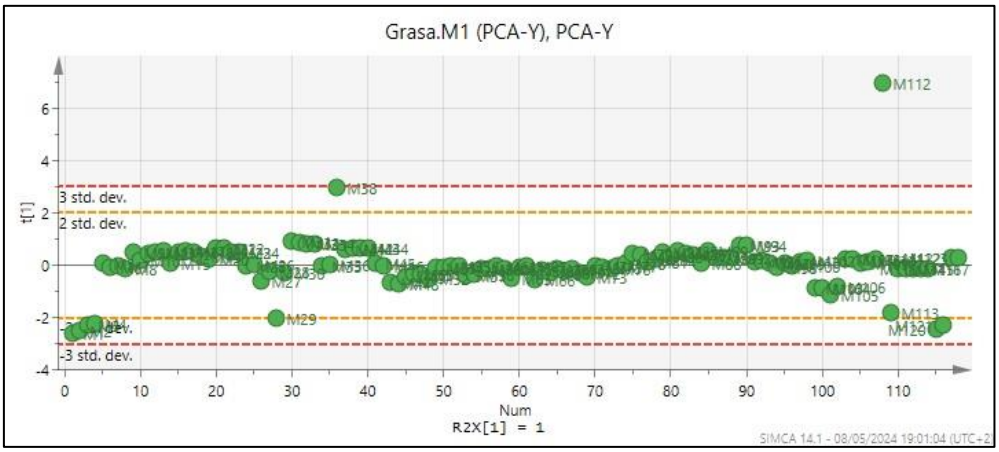
Modelo hau aldagai honetara ez dela oso ondo egokitzen esan daiteke batez ere bere errore balio altuagatik; hala ere, laktosaren kontzentrazioa nahiko ondo aurreikusten duela frogatu da.

4.2.2 Grasa

Normalean behi-esnearen grasa kontzentrazioa %3,5 ingurukoa izaten da. Erabilitako dataset-ean lagin gehienak %3-4 artean daude, nahiz eta zenbait lagin balio hauetatik asko desbideratzen diren.

Lehendabizi, PCA-Y modelo bat sortu da ea aldagai honi dagokion outlierrik dagoen aztertzeko.

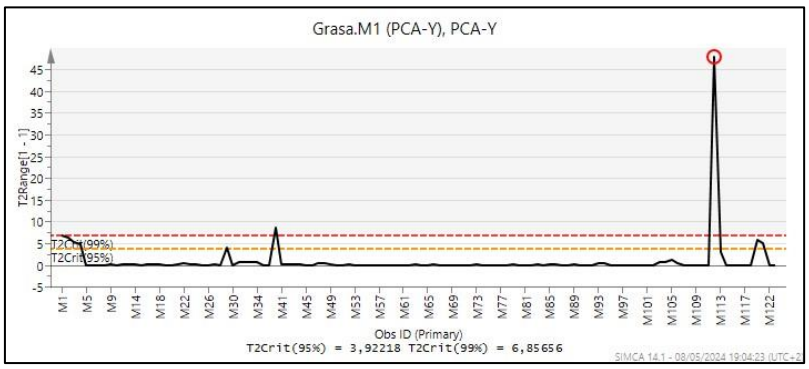
Ondorengo Score grafikoa lortu da:



***38**. Irudia: Grasaren PCA-Y modeloaren "Score" grafikoa*

Ikus daitekeenez 112 lagina gainontzekoetatik oso desbideratuta dago eta tolerantzia lerrotik oso urrun.

"Hotelling's T2" neurriari begiratuz ikus dezakegu lagin honen balioa oso altua dela:



***39**. Irudia: Grasaren PCA-Y modeloaren "Hotelling's T2" neurria*

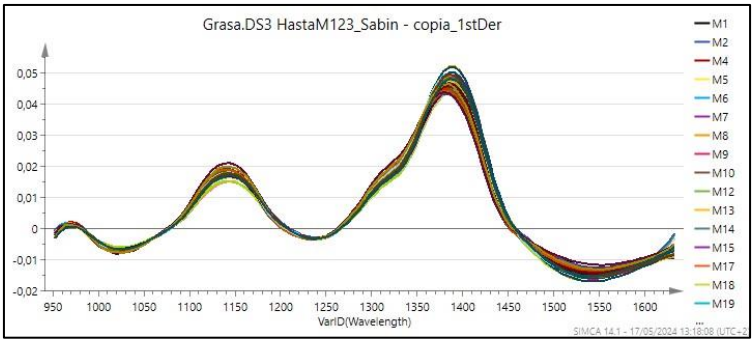
Beraz, 112 lagina outlier bat dela esan daiteke eta hurrengo modeloetatik ezabatuko da X aldagaiei dagozkien outlierrekin batera.

Hainbat modelo aztertu dira, aldagai kopuru ezberdinekin, non Q2 neurria maximizatzea eta RMSEP neurria minimizatzea bilatu den, beti ere zarata modelizatzea ekidituz.

Grasarentzako, balio onenak lortu dituen modeloa hau izan da:

OPLS modelo bat osagai prediktibo batekin eta bost osagai ortogonalekin, lehenengo deribatua egiten duen iragazki batekin.

1.deribatua egitearen ondorioz espektroaren forma aldatzen da:



***40**. Irudia: 1.deribatua egiteagatik geratzen den espektro forma*

Ikus daitekeenez honi esker lehen batez ere 1350-1450 nm artean zeuden offset-ak asko txikiagotu dira.

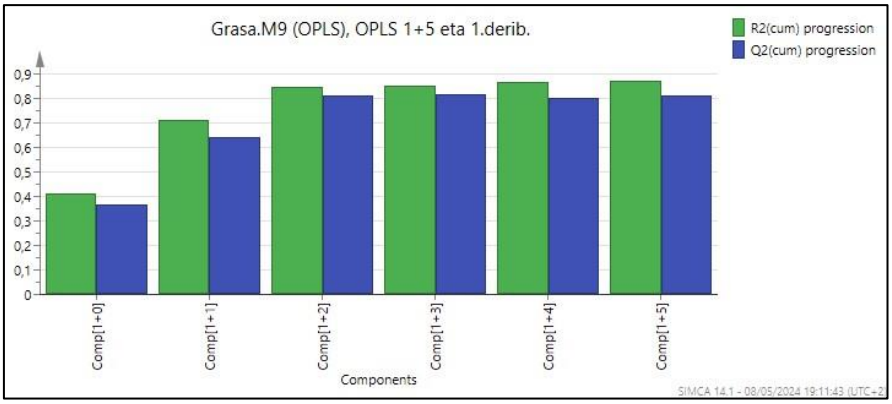
Hurrengo balioak lortu dira modelo honetarako:

R2X: 0,996 ; R2Y: 0,875 ; Q2: 0,814

Type: OPLS Observations (N)=94, variables (K)=112 (X=111, Y=1)

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,996			0,875			0,814		1		
Predictive		0,234			0,875			0,814		1		
P1	0,234	0,234	22	0,875	0,875	0,814	0,01	0,814	1	1	1	R1
Orthogonal in X(OPL...		0,761			0							
O1	0,531	0,531	50	0	0							R1
O2	0,175	0,706	16,4	0	0							R1
O3	0,04...	0,75	4,08	0	0							NS
O4	0,00...	0,756	0,596	0	0							NS
O5	0,00...	0,761	0,487	0	0							R1

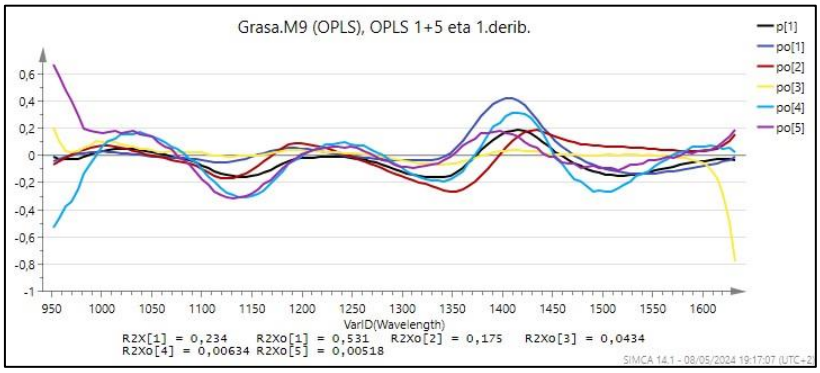
***41**. Irudia: Grasaren OPLS modeloaren R2X, R2Y eta Q2 balioak*



***42**. Irudia: Grasaren OPLS modeloaren osagaiak*

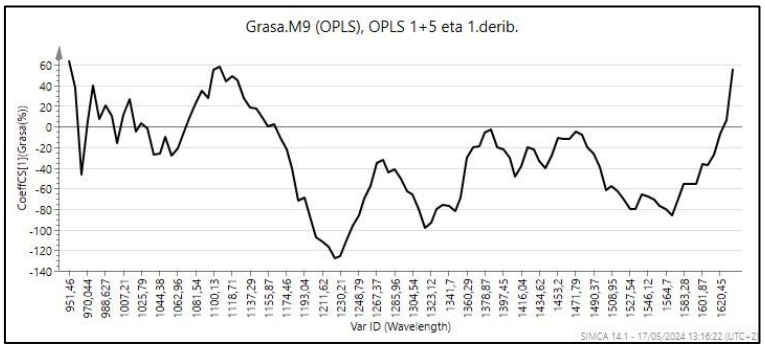
Nahiz eta ikus daitekeen azken osagaiekin modeloa berez ez dela ia ezer ez hobetzen, proba ezberdinen bidez ikusi da osagai gehikuntza honekin RMSEP balioa hobetzen dela. Gainera osagai gehikuntza hauek ez dutela ia zaratarik modelizatzen bermatu da.

Bertan ikus daiteke osagai honetako bakoitzak duen itxura espektroan zehar:



***43**. Irudia: Grasaren OPLS modeloaren "Loading" grafikoa*

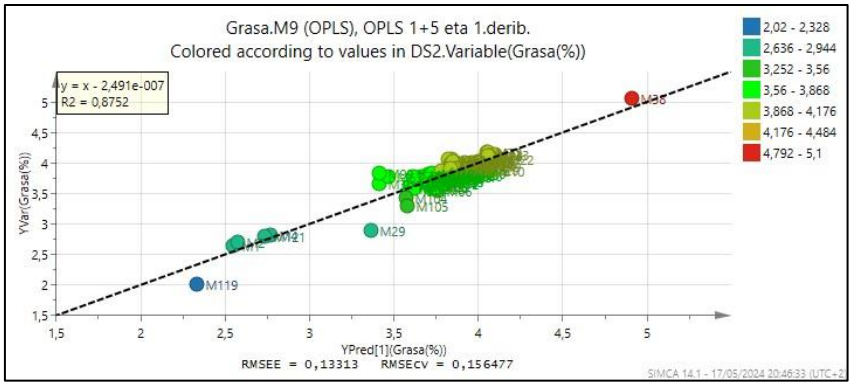
Hurrengo irudian, osagai guzti hauek batera grasa aldagairako sortzen duten iragarpen bektorea ikus dezakegu.



***44**. Irudia: Grasaren OPLS modeloaren iragarpen bektorea*

Ikus daiteke banaketa nahiko uniformea dela orokorrean nahiz eta desbiderapen handia duten zenbait lagin dauden.

Behatutakoa Vs Iragarritakoa grafikoa, entrenamenduko laginen balidazio gurutzatuaren bidez:

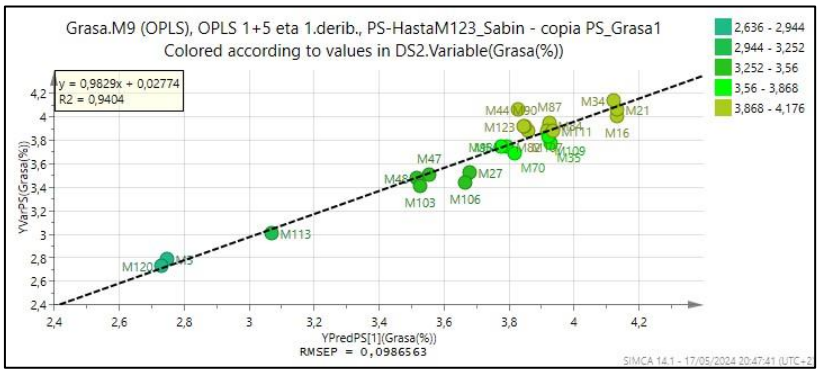


***45**. Irudia: Grasaren OPLS modeloaren entrenamendurako laginen Behatutakoa Vs Iragarritakoa grafikoa*

Ikus daiteke oso lagin gutxi direla erregresio lerroaren norabidea jarraitzen ez dutenak, gehiengoa lerroaren gain-gainean daude. Balidazio gurutzatuaren bidez lortutako iragarpenen errorea %0,1565 da. Kontuan izanda dataset-eko laginen balio tartea %3-ekoa dela, RMSEcv balio honek %5,21-ko errorea dakar balio tarte honekiko.

Honen ostean, modeloak ikusi ez dituen lagin berriak sartzen dira, modeloak haien grasa kontzentrazioa aurreikusteko duen

kapazitatea ikusteko. Guztira 24 lagin berriren iragarpena egin da eta honako hau izan da emaitza:



***46**. Irudia: : Grasaren OPLS modeloaren Behatutako Vs Iragarritakoa grafikoa lagin berriekin*

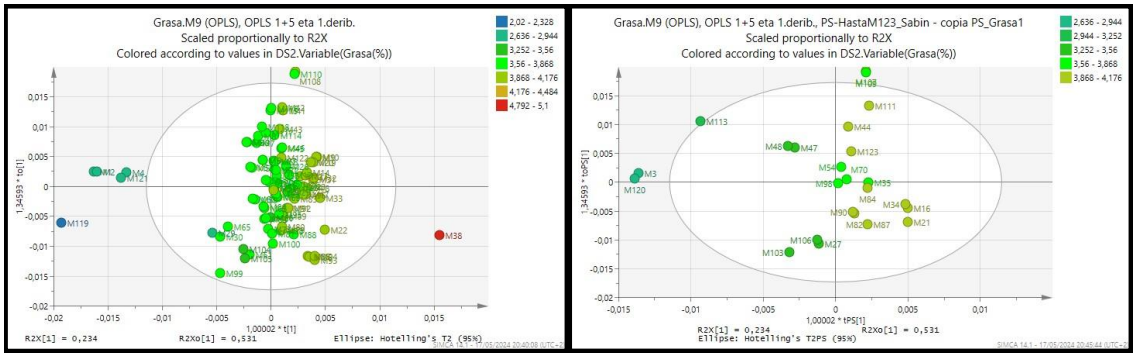
Erregresio lerroarekiko oso desbiderapen gutxirekin ikusten dira lagin berrien iragarpenak. Iragarpen errorea %0,0986-ekoa da eta kontuan izanda iragarpenerako erabilitako laginen balio tartea %1,5-ekoa dela, RMSEP balio honek %6,6-ko errorea dakar balio tarte honekiko.

Hurrengo taulan modeloak iragarri dituen balioak eta benetakoak dituen zerrenda bat ikus daiteke lagin bakoitzerako:

	1	2	3
1	Obs ID (Primary)	M9.YVarPS(Grasa(%))	M9.YPredPS[1](Grasa(%))
2	M3	2,79	2,74815
3	M16	4,01	4,13127
4	M21	4,06	4,13354
5	M27	3,53	3,67726
6	M34	4,14	4,12265
7	M35	3,78	3,92928
8	M44	4,06	3,82883
9	M47	3,51	3,55228
10	M48	3,48	3,51632
11	M54	3,75	3,79098
12	M70	3,69	3,81791
13	M82	3,88	3,85701
14	M84	3,93	3,92575
15	M87	3,95	3,9255
16	M90	3,92	3,84931
17	M98	3,75	3,77408
18	M103	3,41	3,52605
19	M106	3,44	3,66625
20	M107	3,88	3,91742
21	M109	3,83	3,91908
22	M111	3,88	3,93425
23	M113	3,01	3,06972
24	M120	2,73	2,72872
25	M123	3,92	3,8449

***47**. Irudia: Grasaren OPLS modeloak iragarri dituen balioak eta benetakoak*

Jarraian, lagin berri hauek modeloan duten distribuzio bidimentsionala aztertuko da. Beheko irudian entrenamenduko laginen distribuzioa ikusten da ezkerrean eta iragarpenerako erabilitako laginena eskuman. Laginak grasa kontzentrazioaren arabera koloreztatuta daude.



48. Irudia: Grasaren OPLS modeloaren entrenamendurako (ezkerra) eta iragarpenerako (eskuma) erabilitako laginen distribuzio bidimentsionalak

Argi ikusten da kontzentrazio txikiena dutenak ezkerreko aldean daudela eta gehien dutenak eskuman. Bereizketa hau osagai prediktiboari dagokiona da eta laginak ordenatzen ditu grasa kontzentrazioaren arabera ezkerretik eskumara. Lehen esan den bezala lagin gehienek oso balio estandarrak dituzte eta horregatik ageri dira hainbeste lagin erdiko gunean. Bestalde, ardatz bertikala osagai ortogonalei dagokie. Hauek, Y aldagaiarekin, grasarekin, korrelaziorik ez duten X aldagaien (espektroa) aldakortasuna atzematen dute. Erraz ikusten da osagai ortogonalek eragin handia dutela kasu honetan distribuzio bidimentsionalean.

Orokorrean esan daiteke lagin berrien distribuzioa entrenamendu laginen antzekoa dela.

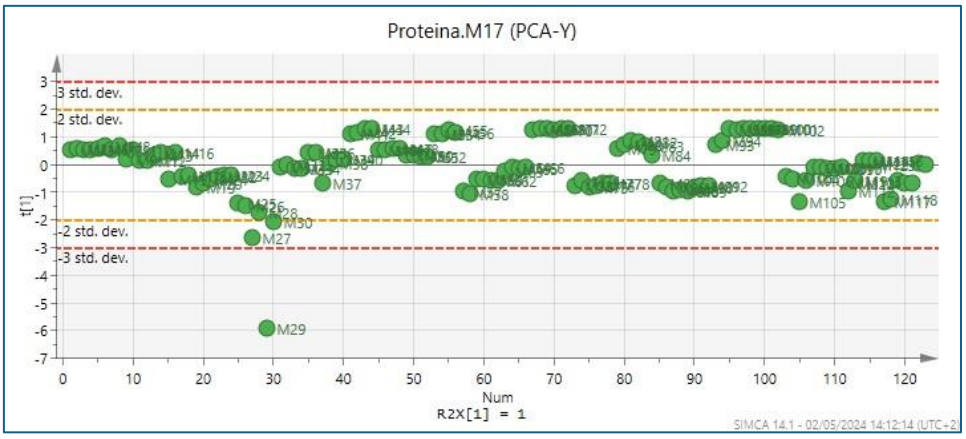
Modelo hau aldagai honetara oso ondo egokitzen da eta grasa kontzentrazioa aurreikusteko indar handia du; batez ere, RMSEP balio oso ona lortzeaz gain Q2 eta R2Y oso onak lortu direlako ere, modeloak etorkizuneko balioetara egokitzeko eta aurreikusteko duen boterea erakusten duena.

4.2.3 Proteina

Normalean behi-esnearen proteina kontzentrazioa %3,2 ingurukoa izaten da. Erabilitako dataset-ean lagin gehienak %3-3,5 artean daude, nahiz eta zenbait lagin balio hauetatik asko desbideratzen diren.

Lehendabizi, PCA-Y modelo bat sortu da ea aldagai honi dagokion outlierrik dagoen aztertzeko.

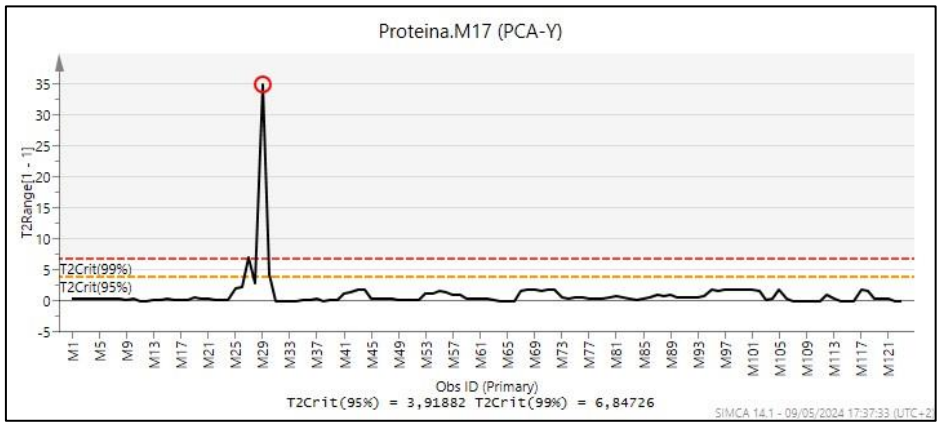
Ondorengo Score grafikoa lortu da:



***49**. Irudia: Proteinaren PCA-Y modeloaren "Score" grafikoa*

Ikus daitekeenez 29 lagina gainontzekoetatik oso desbideratuta dago eta tolerantzia lerrotik urrun.

"Hotelling's T2" neurriari begiratuz ikus dezakegu lagin honen balioa oso altua dela:



***50**. Irudia: : Proteinaren PCA-Y modeloaren "Hotelling's T2" neurria*

Beraz, 29 lagina outlier bat dela esan daiteke eta hurrengo modeloetatik ezabatuko da X aldagaiei dagozkien outlierrekin batera.

Hainbat modelo aztertu dira, aldagai kopuru ezberdinekin, non Q2 neurria maximizatzea eta RMSEP neurria minimizatzea bilatu den, beti ere zarata modelizatzea ekidituz.

Proteinarentzako, balio onenak lortu dituen modeloa hau izan da:

PLS modelo bat bederatzi osagairekin.

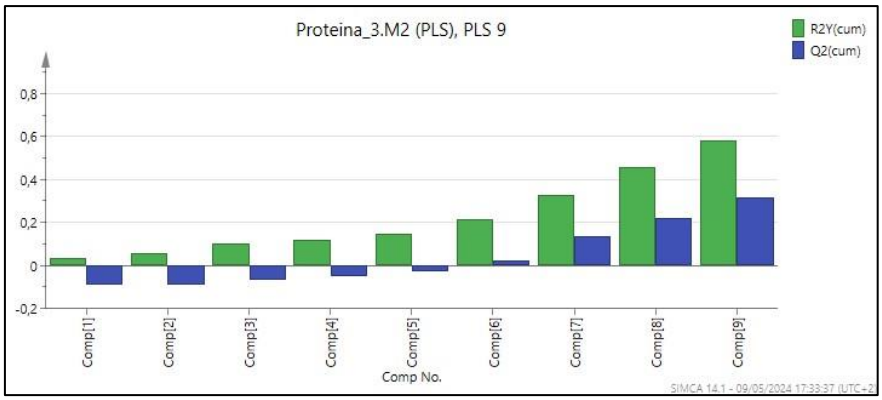
Hurrengo balioak lortu dira modelo honetarako:

R2X: 0,421 ; R2Y: 0,58 ; Q2: 0,316

Type: PLS Observations (N)=94, variables (K)=126 (X=125, Y=1)

Component	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.									
1	0,421	0,421	39,6	0,0342	0,0342	-0,0909	0,05	-0,0909	NS	1
2	0,498	0,919	46,8	0,0239	0,0581	0,00201	0,05	-0,0887	NS	1
3	0,0549	0,974	5,16	0,0428	0,101	0,0217	0,05	-0,0651	NS	1
4	0,0174	0,991	1,64	0,016	0,117	0,0127	0,05	-0,0515	NS	1
5	0,005...	0,997	0,563	0,0336	0,15	0,0211	0,05	-0,0294	NS	1
6	0,001...	0,999	0,178	0,0643	0,215	0,0518	0,05	0,024	R1	1
7	0,000...	0,999	0,0308	0,116	0,331	0,113	0,05	0,134	R1	1
8	0,000...	1	0,0147	0,125	0,456	0,0977	0,05	0,219	R1	1
9	8,02e...	1	0,00754	0,124	0,58	0,125	0,05	0,316	R1	1

***51**. Irudia: Proteinaren PLS modeloaren R2X, R2Y eta Q2 balioak*

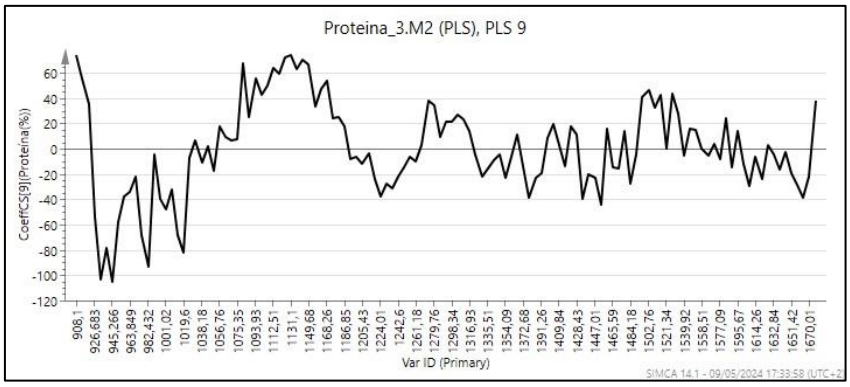


***52**. Irudia: Proteinaren PLS modeloaren osagaiak*

Hainbeste aldagai sartzea beharrezkoa izan da, Q2 balioa oso baxua delako aldagai gutxiagorekin, negatiboa 6. osagairarte gainera. Hala ere, hainbeste aldagai sartzeak bere kalteak dakartza; izan ere,

honek modeloaren konplexutasuna handitzen du eta gainera zarata modelizatzen du.

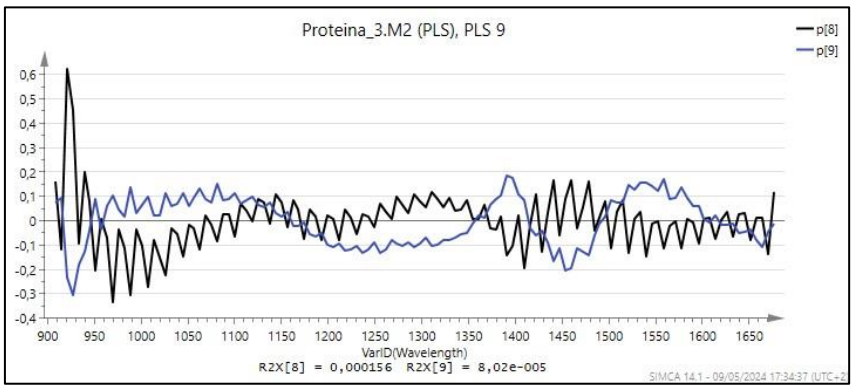
Hurrengo irudian, osagai guzti hauek batera proteina aldagairako sortzen duten iragarpen bektorea ikus dezakegu.



***53**. Irudia: Proteinaren PLS modeloaren iragarpen bektorea*

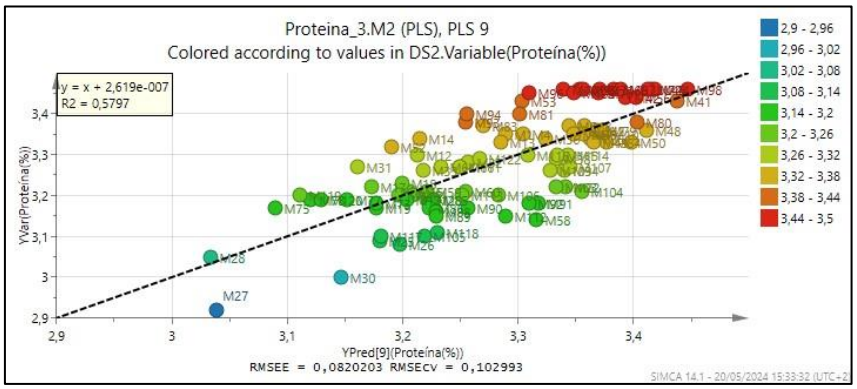
Argi eta garbi ikusten da zarata modelizatzen ari dela.

Zortzigarren eta bederatzigarren osagaien itxurari erreparatuz, ikus daiteke zaratatsuak direla:



***54**. Irudia: : Proteinaren PLS modeloaren zortzigarren eta bederitzagarren osagaien "Loading" grafikoa*

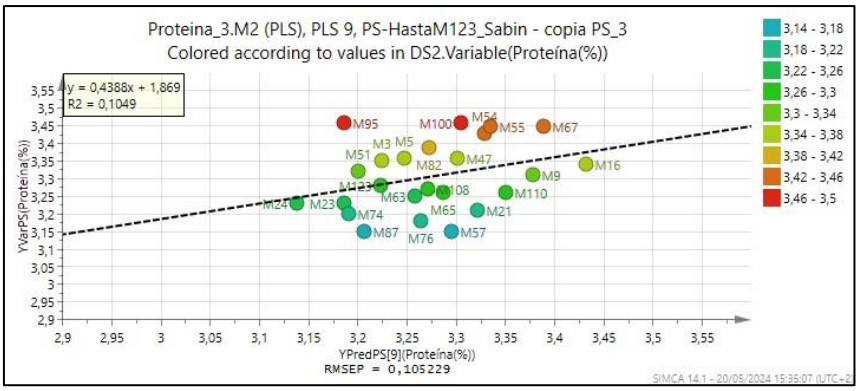
Behatutakoa Vs Iragarritakoa grafikoa, entrenamenduko laginen balidazio gurutzatuaren bidez:



***55**. Irudia: Proteinaren PLS modeloaren entrenamendurako laginen Behatutakoa Vs Iragarritakoa grafikoa*

Balidazio gurutzatuaren bidez lortutako iragarpenen errorea %0,102 da. Kontuan izanda dataset-eko laginen balio tartea %0,6-ekoa dela, RMSEcv balio honek %17-ko errorea dakar balio tarte honekiko.

Honen ostean, modeloak ikusi ez dituen lagin berriak sartzen dira, modeloak haien proteina kontzentrazioa aurreikusteko duen kapazitatea ikusteko. Guztira 24 lagin berriren iragarpena egin da eta honako hau izan da emaitza:



***56**. Irudia: Proteinaren PLS modeloaren Behatutako Vs Iragarritakoa grafikoa lagin berriekin*

Erregresio lerroarekiko desbiderapen nabaria ikusten da lagin berrien iragarpenean. Gainera erregresio zuzenaren maldari erreperatuz, honek 0,4388-ko balioa du; honen balioa 1etik ahalik eta gertuena izan behar da. Erregresioaren R2-a ere oso txarra da.

Honek guztiak esan nahi du modeloa ez dela ondo egokitzen aldagai honen balioak aurreikusteko.

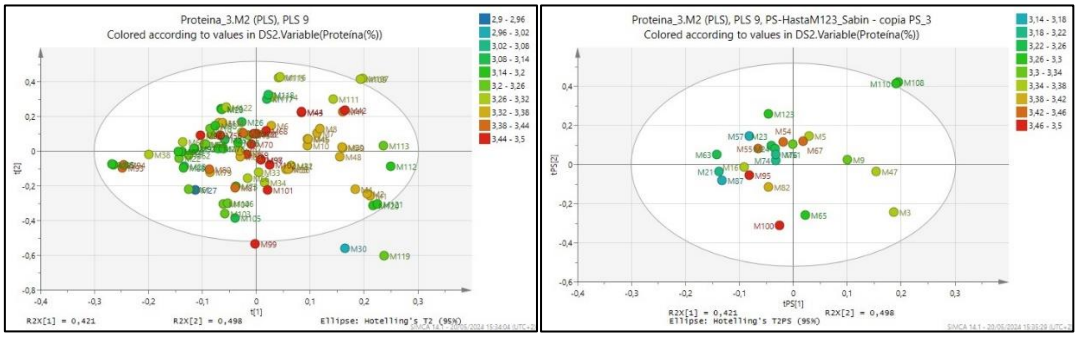
Iragarpen errorea %0,105-ekoa da eta kontuan izanda iragarpenerako erabilitako laginen balio tartea %0,3-ekoa dela, RMSEP balio honek %35-eko errorea dakar balio tarte honekiko.

Hurrengo taulan modeloak iragarri dituen balioak eta benetakoak dituen zerrenda bat ikus daiteke lagin bakoitzerako:

	1	2	3
1	Obs ID (Primary)	M2.YVarPS(Proteína(%))	M2.YPredPS[9](Proteína(%))
2	M3	3,35	3,22459
3	M5	3,36	3,24746
4	M9	3,31	3,37813
5	M16	3,34	3,43211
6	M21	3,21	3,3215
7	M23	3,23	3,18559
8	M24	3,23	3,13807
9	M47	3,36	3,30043
10	M51	3,32	3,19968
11	M54	3,43	3,32845
12	M55	3,45	3,33492
13	M57	3,15	3,29485
14	M63	3,25	3,25728
15	M65	3,26	3,28682
16	M67	3,45	3,38873
17	M74	3,2	3,19052
18	M76	3,18	3,26333
19	M82	3,39	3,27271
20	M87	3,15	3,20561
21	M95	3,46	3,18615
22	M100	3,46	3,30496
23	M108	3,27	3,27054
24	M110	3,26	3,35046
25	M123	3,28	3,2235

***57**. Irudia: Proteinaren PLS modeloak iragarri dituen balioak eta benetakoak*

Jarraian, lagin berri hauek modeloan duten distribuzio bidimentsionala aztertuko da. Beheko irudian entrenamenduko laginen distribuzioa ikusten da ezkerrean eta iragarpenerako erabilitako laginena eskuman. Laginak proteina kontzentrazioaren arabera koloreztatuta daude.



58. Irudia: Proteinaren PLS modeloaren entrenamendurako (ezkerra) eta iragarpenerako (eskuma) erabilitako laginen distribuzio bidimentsionalak

Ikusten da nola ez dagoen inongo ordenarik espazio bidimentsional osoan zehar. Proteina kontzentrazio desberdineko laginak nahastuta ageri dira bata bestearekin. Lehen bi osagai nagusiek ez dute kontzentrazioaren araberako inongo banaketarik egiten; honek esan nahi du osagai hauek ez direla gai proteinaren aldakortasuna hautemateko. Ikusten da lagin berrien distribuzioa entrenamendu laginen antzekoa dela; beraz, bere egiturarekin bat datorrela esan daiteke.

Nahiz eta modelo hau aldagai honetara hobetoen egokitzen den modeloa den, ikus daiteke ez duela proteina kontzentrazioa zehaztasun handiarekin aurreikusteko indarra; batez ere, erregresio zuzenaren malda eta R2 balio baxuak ditu, ondo egokitzen ez denaren seinale. Egia da, zenbait lagin multzorekin hobeto aurreikusten eta egokitzen dela modeloa, baina aldi berean beste lagin multzo askorekin, erabilitakoa adibidez, ez du emaitza oso onik lortzen.

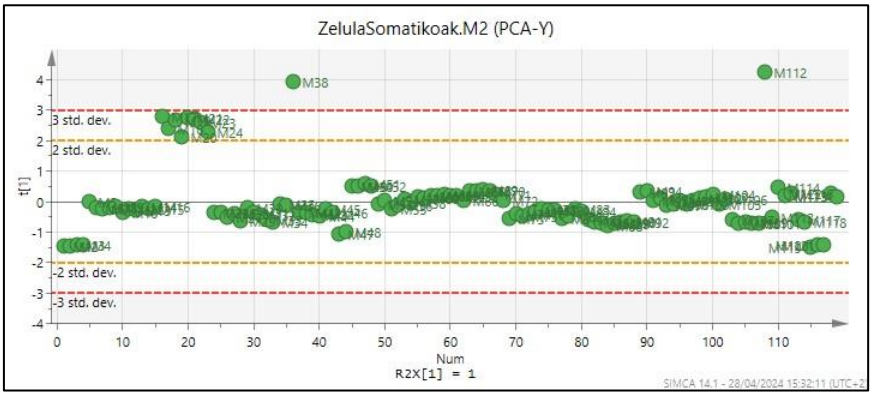
4.2.4 Zelula Somatikoak

Behi-esnearen zelula somatikoen kopurua hainbat faktoreren menpe dago, behiaren osasun-egoera, esnealdi-etapa, genetika, maneiua eta ingurumen-baldintzak. Oro har, zelula somatikoen kontzentrazio batetik gorako kantitate batek (normalean 200.000 zelula/ml inguru) errapeko osasun-arazoak adieraz ditzake.

Normalean behi osasuntsu batek 50.000 zelula/ml-ko zelula somatiko izaten ditu; nahiz eta kopuru hauek 5.000-200.000 zelula/ml artean mugitu ahal diren. Erabilitako dataset-ean lagin gehienak 100.000-220.000 zelula/ml artean daude, nahiz eta lagin asko balio hauetatik asko desbideratzen diren.

Lehendabizi, PCA-Y modelo bat sortu da ea aldagai honi dagokion outlierrik dagoen aztertzeko.

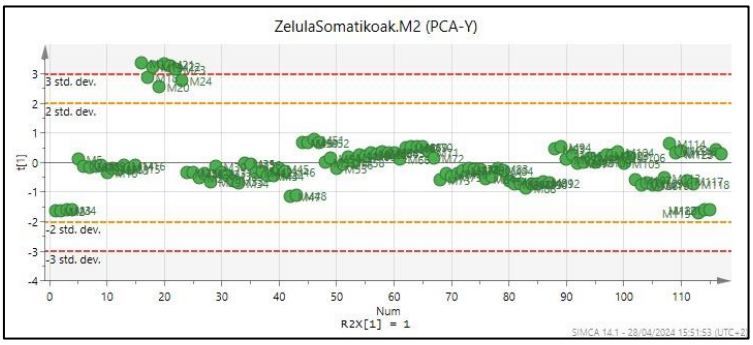
Ondorengo Score grafikoa lortu da:



***59**. Irudia: Zelula somatikoen PCA-Y modeloaren "Score" grafikoa*

Ikus daitekeenez 38 eta 112 laginak gainontzekoetatik desbideratuta daude eta tolerantzia lerrotik urrun. Beraz, lagin hauek outliertzat hartu ditugu eta hurrengo modeloetatik ezabatu dira.

Lagin hauek kenduta ondorengo Score grafikoa lortu da:



***60**. Irudia: Zelula somatikoen PCA-Y modeloaren "Score" grafikoa outlier-ak kenduta*

Ikus daiteke badaudela zenbait lagin goiko partean ez dutela gainontzeko laginen egitura jarraitzen eta nahiko desbideratuta daudela. Ikertu da lagin hauek modeloa izorratzen dutela; beraz, modelotik kentzea hautatu da. Lagin hauek 17-tik 24-ra doazenak dira. Dataset-ari begiratuz, ikusi da lagin hauek 500.000 zelula/ml inguruko kontzentrazioa dutela, gainontzekoetatik oso aldenduta.

Ondoren hainbat modelo aztertu dira, aldagai kopuru ezberdinekin, non Q2 neurria maximizatzea eta RMSEP neurria minimizatzea bilatu den, beti ere zarata modelizatzea ekidituz.

Zelula somatikoentzako, balio onenak lortu dituen modeloa hau izan da:

OPLS modelo bat osagai prediktibo batekin eta hiru osagai ortogonalekin.

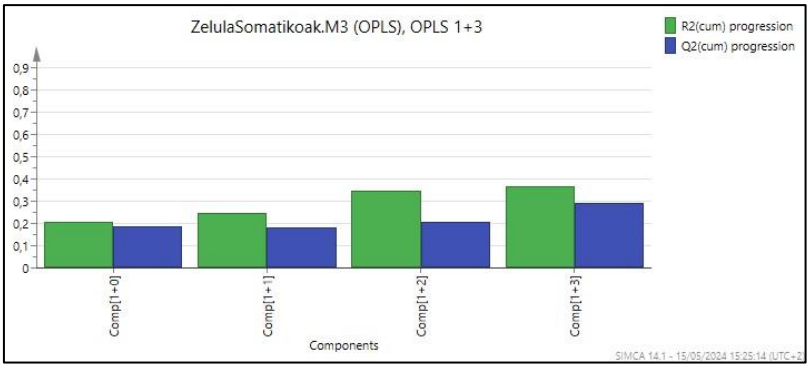
Hurrengo balioak lortu dira modelo honetarako:

R2X: 0,976; R2Y: 0,367; Q2: 0,292

Type: OPLS Observations (N)=88, variables (K)=126 (X=125, Y=1)

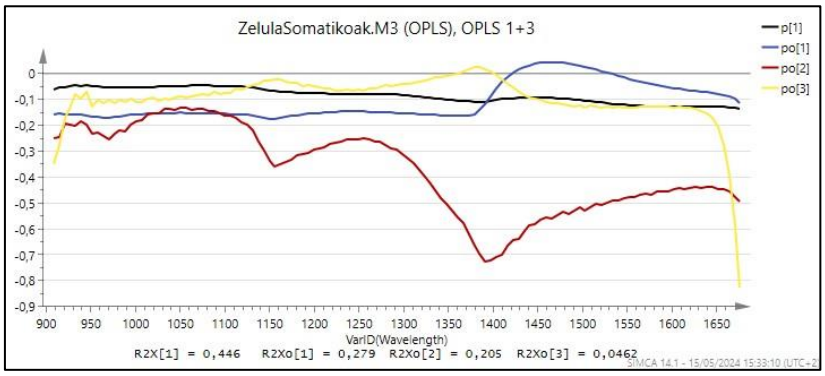
Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,976			0,367			0,292		1		
Predictive		0,446			0,367			0,292		1		
P1	0,446	0,446	39,2	0,367	0,367	0,292	0,01	0,292	1	1	1	R1
Orthogonal in X(OPL...		0,53			0							
O1	0,279	0,279	24,6	0	0							NS
O2	0,205	0,484	18	0	0							R1
O3	0,04...	0,53	4,07	0	0							R1

***61**. Irudia: Zelula somatikoen OPLS modeloaren R2X, R2Y eta Q2 balioak*



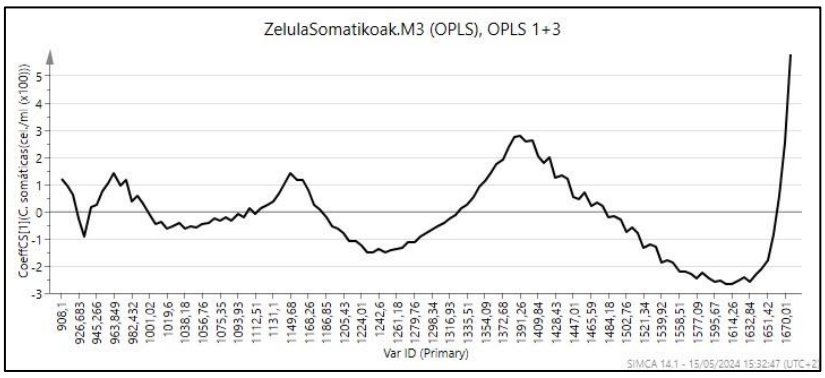
***62**. Irudia: Zelula somatikoen OPLS modeloaren osagaiak*

Bertan ikus daiteke osagai honetako bakoitzak duen itxura espektroan zehar:



***63**. Irudia: Zelula somatikoen OPLS modeloaren "Loading" grafikoa*

Hurrengo irudian berriz, osagai guzti hauek batera sortzen duten iragarpen bektorea ikus dezakegu.

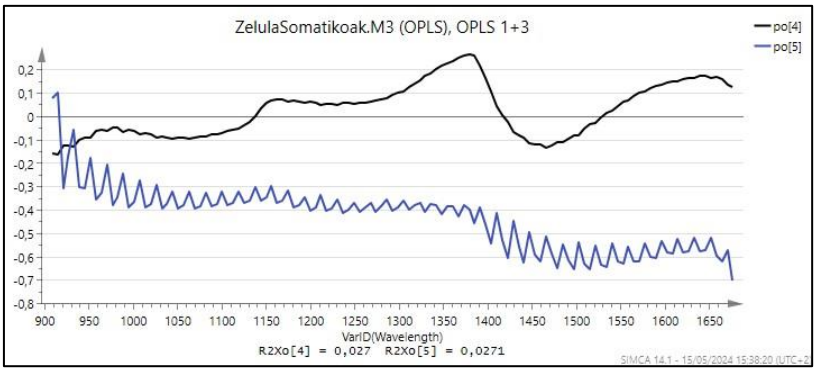


***64**. Irudia: Zelula somatikoen OPLS modeloaren iragarpen bektorea*

Osagai gehiago ez gehitzearen arrazoia, 4.osagai ortogonalak ez duela modeloa hobetzen da nahiz eta zaratatsua ez izan, horregatik

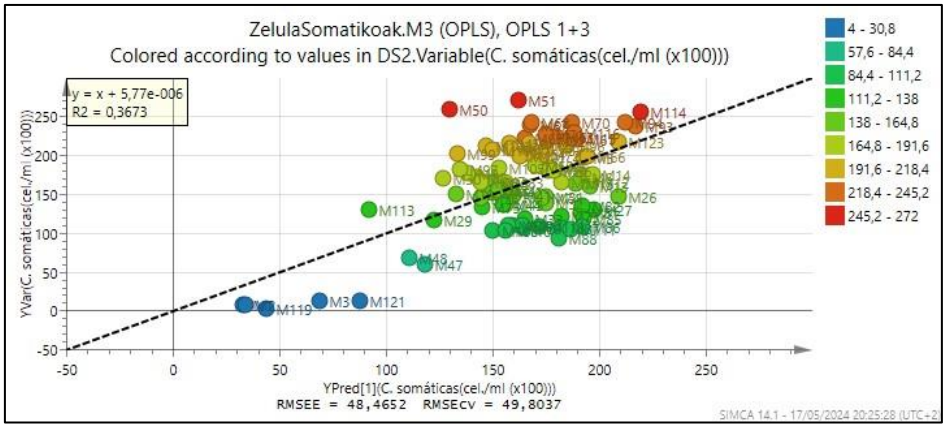
ez da modeloan sartu eta 5. osagai ortogonaletik aurrera, modeloa zarata modelizatzen hasten da.

Hurrengo irudian ikus daiteke argi nola 5. osagai ortogonala zaratatsua den.



65. Irudia: Zelula somatikoen OPLS modeloaren laugarren eta bosgarren osagaien "Loading" grafikoa

Behatutakoa Vs Iragarritakoa grafikoa, entrenamenduko laginen balidazio gurutzatuaren bidez:

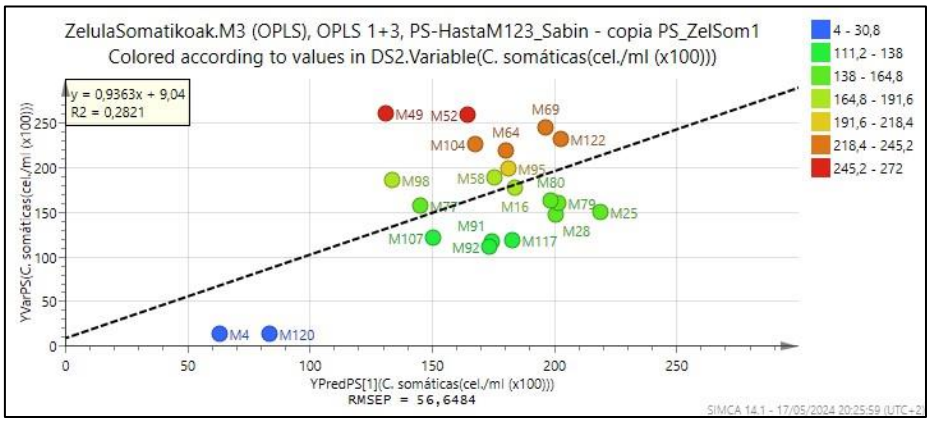


***66**. Irudia: Zelula somatikoen OPLS modeloaren entrenamendurako laginen Behatutakoa Vs Iragarritakoa grafikoa*

Ikus daiteke laginak ez daudela erregresio lerroaren oso gainean. Balidazio gurutzatuaren bidez lortutako iragarpenen errorea ia 50.000 zelula/ml da. Kontuan izanda dataset-eko laginen balio tartea 250.000 zelula/ml dela, RMSEcv balio honek %20-eko errorea dakar balio tarte honekiko.

Honen ostean, modeloak ikusi ez dituen lagin berriak sartzen dira, modeloak haien zelula somatiko kontzentrazioa aurreikusteko

duen kapazitatea ikusteko. Guztira 21 lagin berriren iragarpena egin da eta honako hau izan da emaitza:



***67**. Irudia: Zelula somatikoen OPLS modeloaren Behatutako Vs Iragarritakoa grafikoa lagin berriekin*

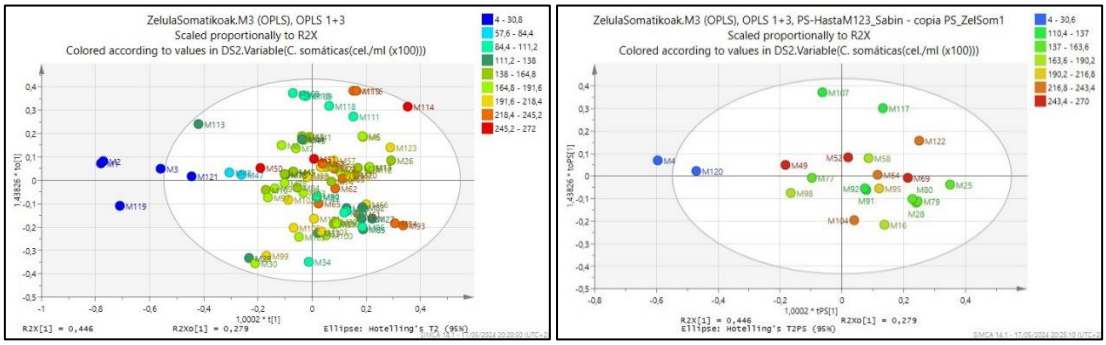
Kasu honetan ere, laginak erregresio lerroarekiko distantzia nabaria dute. Iragarpen errorea 56.000 zelula/ml da eta kontuan izanda iragarpenerako erabilitako laginen balio tartea 250.000 zelula/ml dela, RMSEP balio honek %22,4-eko errorea dakar balio tarte honekiko.

Hurrengo taulan modeloak iragarri dituen balioak eta benetakoak dituen zerrenda bat ikus daiteke lagin bakoitzerako:

	1	2	3
1	Obs ID (Primary)	M3.YVarPS(C. somáticas(cel./ml (x100)))	M3.YPredPS[1](C. somáticas(cel./ml (x100)))
2	M4	14	62,9994
3	M16	178	183,817
4	M25	151	218,671
5	M28	147	200,631
6	M49	261	130,865
7	M52	260	164,525
8	M58	189	175,271
9	M64	220	180,02
10	M69	245	196,48
11	M77	158	145,117
12	M79	160	201,404
13	M80	163	198,567
14	M91	117	174,226
15	M92	112	173,291
16	M95	199	180,932
17	M98	186	133,713
18	M104	226	167,513
19	M107	122	150,572
20	M117	119	182,828
21	M120	14	83,4938
22	M122	233	202,503

68. Irudia: Zelula somatikoen OPLS modeloak iragarri dituen balioak eta benetakoak

Jarraian, lagin berri hauek modeloan duten distribuzio bidimentsionala aztertuko da. Beheko irudian entrenamenduko laginen distribuzioa ikusten da ezkerrean eta iragarpenerako erabilitako laginena eskuman. Laginak zelula somatiko kontzentrazioaren arabera koloreztatuta daude.



69. Irudia: Zelula somatikoen OPLS modeloaren entrenamendurako (ezkerra) eta iragarpenerako (eskuma) erabilitako laginen distribuzio bidimentsionalak

Zelula somatiko kontzentrazioari erreparatuz, ikus daiteke ezkerrerago daudenak kontzentrazio txikia dutela; hala ere, gainontzeko espazio osoan ez dago inongo ordenarik. Beraz, esan daiteke osagai prediktiboa ez dela gai zelula somatiko aldagaiaren aldakortasuna ondo hautemateko.

Orokorrean esan daiteke lagin berrien distribuzioa entrenamendu laginen antzekoa dela.

Ondorioztatu daiteke MVDA-ren bidez zelula somatikoen kontzentrazioa iragartzea nahiko zaila dela, aldagai honen aldakortasunagatik, ez linealtasunagatik eta honi eragiten dioten faktore anitzengatik. Hala ere, probatutako modelo guztietatik, modelo honek iragarpen emaitza onenak lortu ditu.

4.2.5 Gihar estraktu lehorra

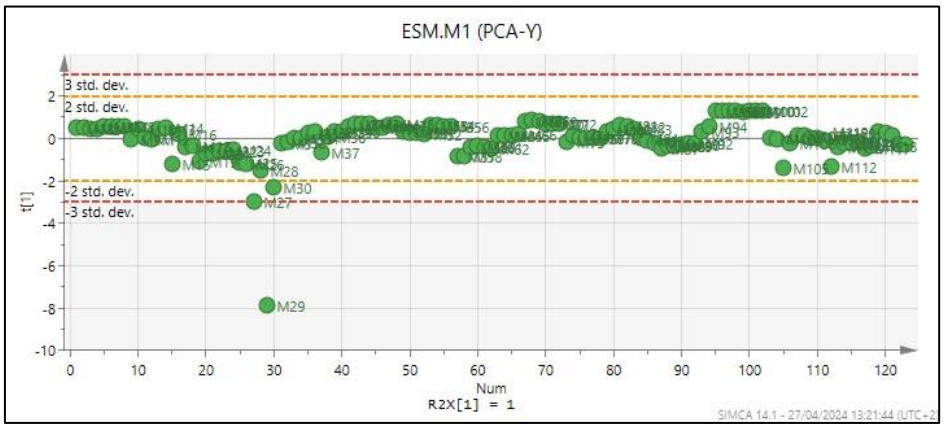
Hemendik aurrera gihar estraktu lehorraren aldagaiari ESM esango zaio, gaztelerako "Extracto Seco Magro" adierazpenaren siglak.

Batez beste, behi-esnearen ESM edukia % 8,5etik % 9ra bitartekoa izaten da. Erabilitako dataset-ean lagin gehienen ESM balioak balio tarte horren barruan daude, desbiderapen handirik gabe.

Proba hauetarako gainontzeko Y aldagai guztiak kendu dira Laktosa aldagaia izan ezik.

Lehenengo eta behin PCA-Y modelo bat sortu da lagin guztiekin, ea aldagai honi dagokion beste outlierrik dagoen ikusteko.

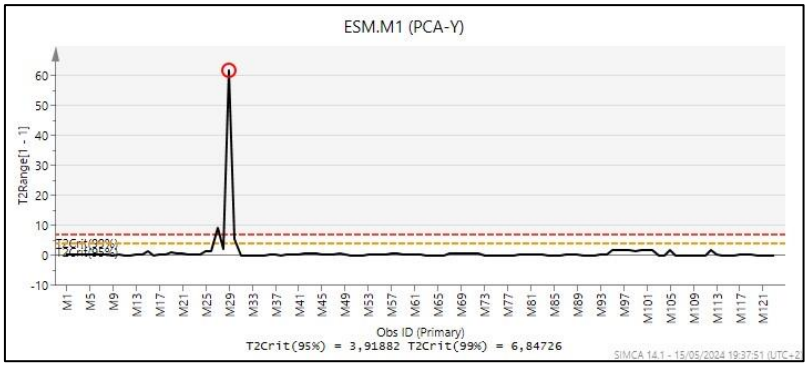
Ondorengo Score grafikoa lortu da:



***70**. Irudia: Gihar estraktu lehorraren PCA-Y modeloaren "Score" grafikoa*

Bertan argi ikus daiteke 29 lagina gainontzekoetatik oso desbideratuta dagoela eta tolerantzia lerrotik oso urrun dagoela. Beraz, argitasun osoz esan daiteke lagin hau outlier bat dela. Dataset-ari begiratuz ikusi da lagin honek oso balio atipikoa duela, %6,81.

"Hotelling's T2" neurriari begiratuz ikus dezakegu lagin honen balioa oso altua dela:



***71**. Irudia: Gihar estraktu lehorraren PCA-Y modeloaren "Hotelling's T2" neurria*

Modu honetan, X aldagaiei dagokien outlierrei (11,37,39 eta 40 laginak), azken hau gehitzen zaie eta ESM-aren analisiari dagokien modeloetatik kenduko da ere.

Hainbat modelo aztertu dira, aldagai kopuru ezberdinekin, non Q2 neurria maximizatzea eta RMSEP neurria minimizatzea bilatu den, beti ere zarata modelizatzea ekidituz.

ESM-arentzako, balio onenak lortu dituen modeloa hau izan da:

OPLS modelo bat osagai prediktibo batekin eta zortzi osagai ortogonalekin.

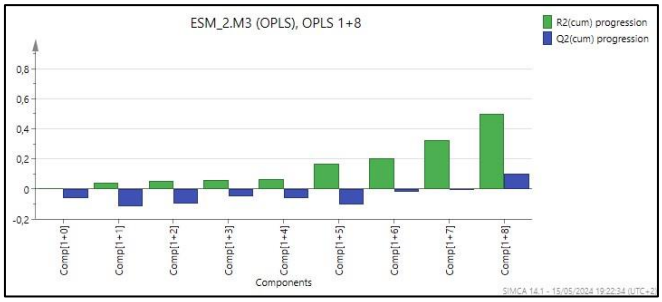
Hurrengo balioak lortu dira modelo honetarako:

R2X: 1; R2Y: 0,499; Q2: 0,104

Type: OPLS Observations (N)=94, variables (K)=126 (X=125, Y=1)

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		1			0,499			0,104		1		
Predictive		0,00822			0,499			0,104		1		
P1	0,00822	0,00822	0,773	0,499	0,499	0,104	0,01	0,104	1	1	1	NS
Orthogonal in X(OPL...		0,991			0							
O1	0,695	0,695	65,3	0	0							NS
O2	0,083	0,778	7,8	0	0							R1
O3	0,116	0,894	10,9	0	0							R1
O4	0,0368	0,931	3,46	0	0							NS
O5	0,0429	0,973	4,04	0	0							NS
O6	0,00561	0,979	0,527	0	0							R1
O7	0,00774	0,987	0,728	0	0							R1
O8	0,00456	0,991	0,429	0	0							R1

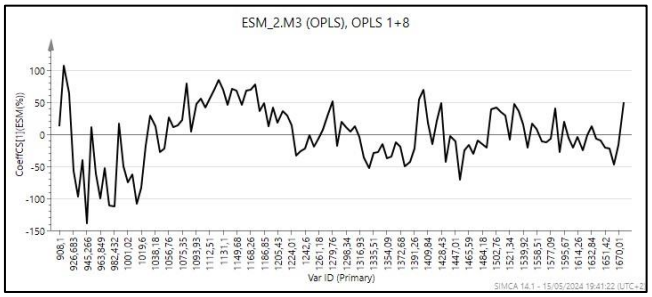
***72**. Irudia: Gihar estraktu lehorraren OPLS modeloaren R2X, R2Y eta Q2 balioak*



***73**. Irudia: Gihar estraktu lehorraren OPLS modeloaren osagaiak*

Hainbeste aldagai sartzea beharrezkoa izan da, Q2 balioa oso baxua delako aldagai gutxiagorekin, negatiboa 7. osagai ortogonalerarte gainera. Hala ere, hainbeste aldagai sartzeak bere kalteak dakartza; izan ere, honek modeloaren konplexutasuna handitzen du eta gainera zarata modelizatzen du.

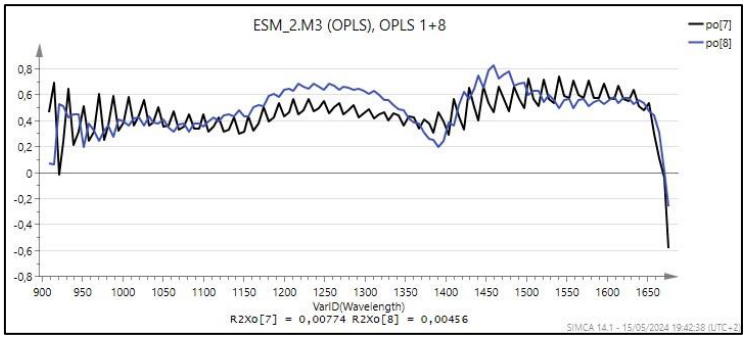
Hurrengo irudian, osagai guzti hauek batera ESM aldagairako sortzen duten iragarpen bektorea ikus dezakegu.



***74**. Irudia: Gihar estraktu lehorraren OPLS modeloaren iragarpen bektorea*

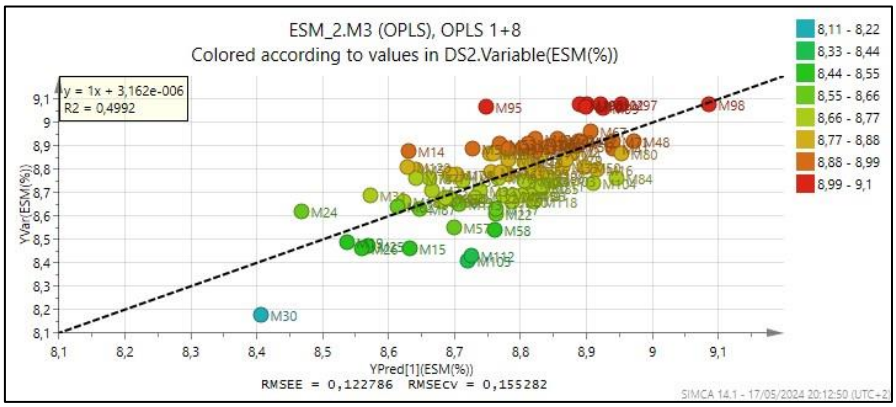
Argi eta garbi ikusten da zarata modelizatzen ari dela.

Zazpigarren eta zortzigarren osagai ortogonalen itxurari erreparatuz, ikus daiteke zaratatsuak direla:



***75**. Irudia: Gihar estraktu lehorraren OPLS modeloaren zazpigarren eta zortzigarren osagaien "Loading" grafikoa*

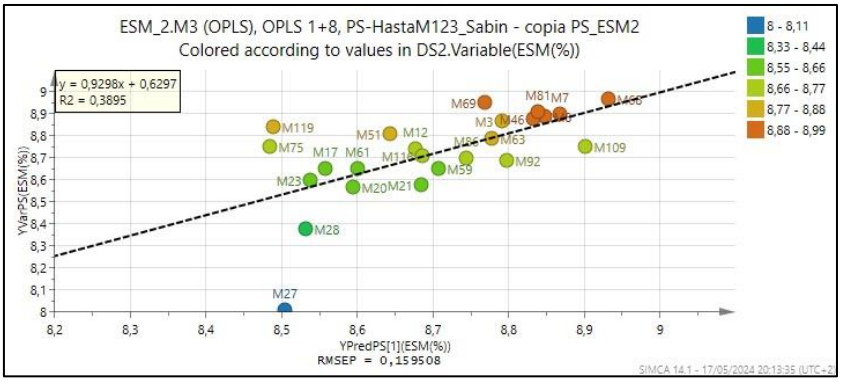
Behatutakoa Vs Iragarritakoa grafikoa, entrenamenduko laginen balidazio gurutzatuaren bidez:



***76**. Irudia: Gihar estraktu lehorraren OPLS modeloaren entrenamendurako laginen Behatutakoa Vs Iragarritakoa grafikoa*

Balidazio gurutzatuaren bidez lortutako iragarpenen errorea %0,155 da. Kontuan izanda laginen ESM balio tartea %0,9-koa dela, RMSEcv balio honek %17,22-ko errorea dakar balio tarte honekiko.

Honen ostean, modeloak ikusi ez dituen lagin berriak sartzen dira, modeloak haien ESM kontzentrazioa aurreikusteko duen kapazitatea ikusteko. Guztira 24 lagin berriren iragarpena egin da eta honako hau izan da emaitza:



***77**. Irudia: Gihar estraktu lehorraren OPLS modeloaren Behatutako Vs Iragarritakoa grafikoa lagin berriekin*

Erregresio lerroarekiko desbiderapen nabaria ikusten da lagin berrien iragarpenean. Iragarpen errorea %0,159-ekoa da eta kontuan izanda iragarpenerako erabilitako laginen ESM balio

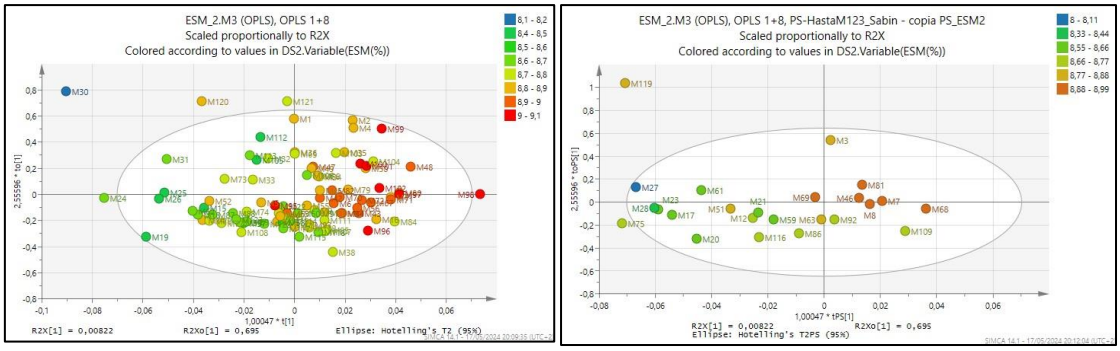
tartea %1-eko dela, RMSEP balio honek %15,9-ko errorea dakar balio tarte honekiko.

Hurrengo taulan modeloak iragarri dituen balioak eta benetakoak dituen zerrenda bat ikus daiteke lagin bakoitzerako:

	1	2	3
1	Obs ID (Primary)	M3.YVarPS(ESM(%))	M3.YPredPS[1](ESM(%))
2	M3	8,87	8,7911
3	M7	8,9	8,8671
4	M8	8,89	8,8484
5	M12	8,74	8,67643
6	M17	8,65	8,55715
7	M20	8,57	8,59348
8	M21	8,58	8,68449
9	M23	8,6	8,53708
10	M27	8,01	8,50359
11	M28	8,38	8,53088
12	M46	8,88	8,83282
13	M51	8,81	8,64334
14	M59	8,65	8,7072
15	M61	8,65	8,60048
16	M63	8,79	8,77743
17	M68	8,97	8,93212
18	M69	8,95	8,76837
19	M75	8,75	8,48366
20	M81	8,91	8,83796
21	M86	8,7	8,74385
22	M92	8,69	8,79715
23	M109	8,75	8,90156
24	M116	8,71	8,68592
25	M119	8,84	8,48826

***78**. Irudia: Gihar estraktu lehorraren OPLS modeloak iragarri dituen balioak eta benetakoak*

Jarraian, lagin berri hauek modeloan duten distribuzio bidimentsionala aztertuko da. Beheko irudian entrenamenduko laginen distribuzioa ikusten da ezkerrean eta iragarpenerako erabilitako laginena eskuman. ESM kontzentrazioaren arabera koloreztatu dira.



79. Irudia: Gihar estraktu lehorraren OPLS modeloaren entrenamendurako (ezkerra) eta iragarpenerako (eskuma) erabilitako laginen distribuzio bidimentsionalak

Nahiz eta antzeman daitekeen kontzentrazio altuagoa duten laginak eskumarago daudela eta gutxiago dutenak ezkerrerago, ez

daude era oso argian ordenaturik, ezin daiteke esan osagai prediktiboak Y aldagaiaren aldakortasuna oso ondo hautematen duenik.

Nahiz eta modelo hau aldagai honetara hobetoen egokitzen den modeloa den, ikus daiteke ez duela ESM kontzentrazioa zehaztasun handiarekin aurreikusteko indarra; batez ere, Q2 eta R2Y balioak ez direlako oso onak. Hala ere, zenbait kasutan, aurreikuspenerako erabili ditugun laginen kasuan adibidez, ikus daiteke iragarpenak nahiko onak direla. Hau, lagin hauen gehiengoak balio tarte tipikoaren (%8,5-9) barruan daudelako da. Tarte honetatik kanpo modeloak iragarpen indarra galtzen du.

4.3 Erabilitako modeloen laburpena

Argi utzi beharra dago emaitza hauek lagin multzo eta kopuru batekin lortutakoak direla. Modelo hauek lagin multzo desberdinekin probatu dira, emaitza hobeago eta txarragoekin. Balio ohikoenak dituzten modeloak aukeratu dira, hauen emaitzak ahalik eta kasu gehienetara hurbil daitezen.

Y aldagaia	Balio tartea	Modeloa	Osagai kopurua	Q2	R2Y	RMSEcv	RMSEP
Laktosa	%4,3 - 4,9	OPLS + 2.deribatua	4	0,159	0,258	%0,0946	%0,0625
Grasa	%2 - 5	OPLS + 1.deribatua	6	0,814	0,875	%0,1565	%0,0986
Proteina	%2,9 – 3,5	PLS	9	0,316	0,58	%0,102	%0,105
Zelula Somatikoak	0 – 250.000 Zelula/ml	OPLS	4	0,292	0,367	50.000 Zelula/ml	56.000 Zelula/ml
Gihar estraktu lehorra (ESM)	%8,2 – 9,1	OPLS	9	0,104	0,499	%0,155	%0,159

4. Taula: Erabilitako modeloen laburpen taula

ALDERDI EKONOMIKOA

1.GASTU-AITORPENA

Bertan proiektuaren gastu-aitorpena xehatuko da. Honekin, proiektua aurrera eramateko beharrezkoa den dirua kuantifikatu nahi da.

1.1 BARNE ORDUAK

Bertan lan-taldeko kideen soldatak sartzen dira, proiektuan lan egin duten ordu kopuruaren arabera. Proiektu honetan, lan-taldea ingeniari junior batez eta proiektuaren zuzendariaz osatua dago.

Kontzeptua	Kostua (€/o)	Orduak	Kostu totala (€)
Ingeniari junior-a	30	550	16500
Proiektuaren zuzendaria	60	60	3600
		Totala (€)	20100

1.2 AMORTIZAZIOAK

Amortizazioak proiektuan erabiltzen diren aktibo finkoak dira.

Kontzeptua	Kantitatea	Aleko kostua (€)	Ordu erabilgarriak	Erabilitako orduak	Kostu totala (€)
Simca lizentzia	1	5000	8760	400	228,31
Office lizentzia	1	450	8760	150	7,7
Ordenagailu eramangarria	1	700	40000	550	9,62
				Totala (€)	245,63

1.3 GASTU-AITORPENAREN LABURPENA

Hau da proiektuaren gastu-aitorpenaren kostu totalen laburpena.

Atala	Kostu totala (€)
Barne orduak	20100
Amortizazioak	245,63
Totala (€)	20345,63

ONDORIOAK

Lan honen helburu nagusia, NIR espektroskopiaren eta datuen aldagai anitzeko analisiaren bidez, behi esnearen laktosa, grasa, proteina, zelula somatikoak eta gihar estraktu lehorra modelatzea eta aurreikustea izan da.

Parametro bakoitzerako, hobetoen egokitzen zaion modeloa bilatu da, iragarpen ahalmena maximizatuz eta zarata modelatzea ahalik eta hein handienean ekidituz.

Emaitza onenak lortu dituen modeloa grasarena izan da. Modeloaren osagaiek, grasa aldagaiaren aldakuntza oso ondo azaltzen dute, lortutako R2Y balioan oinarrituta. Honek esan nahi du modeloak ondo iragartzen duela grasa, espektroaren aldagaietan oinarrituta. Lortutako RMSEP balioak ere, modeloak entrenamendurako erabili ez dituen eta inoiz ikusi ez dituen laginen grasa kontzentrazioa iragartzeko ahalmen handia duela adierazten du.

Ondoren, esan daiteke gainontzeko parametroentzat, modeloak iragarri nahi diren aldagaietara egokitzeko zailtasunak dituztela. Batez ere, modelo hauen osagaiek parametro hauen aldakuntza azaltzeko gaitasuna falta zaielako. Hala ere, ikusi da modelo hauek entrenamendurako erabili ez dituzten eta inoiz ikusi ez dituzten laginen parametro hauen kontzentrazioak aurreikusteko gaitasuna dutela, ez oso modu zehatz eta fidagarrian baina bai laginen kantitate esanguratsu bat balio errealetatik gertu geratzeko moduan.

Beraz, etorkizuneko lan eta ikerketetarako, datuen aldagai anitzeko analisia egiten duten teknika berriak probatzea iradokitzen da, parametro hauetara hobeto egokitu daitezkeenak. Baita lagin kopuru handiago eta anitzagoa duen dataset bat erabiltzea; modeloak eraikitzeko garaian, lagin multzo bat

erabiltzetik bestera emaitza homogeneo eta zehatzagoak lortzeko helburuarekin.

BIBLIOGRAFIA

[1] Alexis Rodriguez Reséndiz, "Quimiometria".

https://www.academia.edu/43650755/QUIMIOMETR%C3%8DA_ALEXIS_RODR%C3%8DGUEZ_RES%C3%89NDIZ

[2] "What is Multivariate Data Analysis? | Sartorius."
https://www.sartorius.com/en/knowledge/science-snippets/data-analytics-for-beginners-how-multivariate-data-analysis-can-separate-the-players-from-the-gorillas-507202

[3] "Multivariate Data Analysis for Omics | MKS Umetrics."

https://metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf

[4] "What is Principal Component Analysis (PCA) and how it is used? | Sartorius."

https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

[5] "OPLS Vs PLS | Sartorius"

https://www.sartorius.com/en/knowledge/science-snippets/opls-vs-pls-modeling-to-improve-bioprocess-yields-of-batch-processes-602762

[6] "OPLS Vs PCA: Explaining differences or groupin data? | Sartorius."

https://www.sartorius.com/en/knowledge/science-snippets/explaining-differences-or-grouping-data-opls-da-vs-pca-data-analysis-507204

[7] "Espektroscopia: la interacción de la luz y la materia | Khan Academy"

https://es.khanacademy.org/science/ap-chemistry/electronic-structure-of-atoms-ap/bohr-model-hydrogen-ap/a/spectroscopy-interaction-of-light-and-matter

[8] "Espectometria de absorción | Espectometria.com "

https://www.espectrometria.com/espectrometra_de_absorcin

[9] "Espectometria de emisión | Espectometria.com"

https://www.espectrometria.com/espectrometra_de_emisin

[10] "Espectometria de fluorescencia | Espectometria.com"

https://www.espectrometria.com/espectrometra_de_fluorescencia

[11] "Espectometria Raman | Espectometria.com"

https://www.espectrometria.com/espectrometra_raman

[12] "Espectometria de resonancia magnética nuclear | Espectometria.com"

https://www.espectrometria.com/espectrometra_de_resonancia_magntica_nuclear

[13] "¿Que es la espectroscopia NIR? | Metrohm."

https://www.metrohm.com/it_it/discover/blog/2024/cosa_%C3%A8_la_spettroscopia_NIR.html

[14] Donald A. Burns, Emil W. Ciurczak, "Handbook of Near-infrared analysis, Third Edition", Practical Spectroscopy Series Volume 35, 2007.

[15] "Components needed to build an NIR Spectrometer | Oxford Instruments"

https://andor.oxinst.com/learning/view/article/components-needed-to-build-an-nir-spectrometer

[16] "Simca | Sartorius"

https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca

[17] "PLS_Toolbox | MathWorks"

https://es.mathworks.com/products/connections/product_detail/pls-toolbox.html