USEDAT-07: IBRO USA-Europe Data Analysis Training Congress, Cambridge, UK-Bilbao, Spain-Miami, USA, 2021

Summary for PTML Chemoinformatics Linear Discriminant Analysis classification model for enantioselective reactions.

Shan He [a], Sonia Arrasate [a], and Humberto González-Díaz [a,b,]*

[a] *Department of Organic and Inorganic Chemistry, Faculty of Science and Technology University of Basque Country UPV-EHU, 48940, Leioa, Basque Country, Spain.*
[b] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.

Resumen

El Machine Learning (ML) se utiliza para el aprendizaje de un sistema. Uno de los propósitos de este aprendizaje automático es la construcción de nuevos modelos computacionales. El ML muestra éxito en diversas áreas como por ejemplo, en sistemas de referencia, interfaz cerebro-computadora, robótica y química.[1,2,3,4]

Recientemente, los operadores de la Teoría de Perturbaciones (PT) y las técnicas de ML, se han combinado para crear modelos potentes PTML (PT+ML), que se aplican a sistemas biológicos complejos en la predicción de la interacción de fármaco-proteína. Como para aquellas proteínas diana involucradas en la vía de dopamina, nanotecnología, ciencias materiales etc.[5,6]

Este método de PTML ha sido desarrollado por nuestro grupo, para buscar modelos capaces de predecir los valores v_{ij} de múltiples propiedades de un sistema i^{th} medido bajo diferentes condiciones

Abstract

Machine Learning (ML) is used to learn a system. One of the purposes of this machine learning is the construction of new computational models. ML shows success in various areas such as reference systems, brain-computer interface, robotics, and chemistry.

Recently, Perturbation Theory (PT) operators and ML techniques have been combined to create powerful PTML (PT+ML) models, which are applied to complex biological systems in predicting drug-protein interaction. As for those target proteins involved in the dopamine pathway, nanotechnology, material science, etc.

This PTML method has been developed by our group to search for models capable of predicting the v_{ij} values of multiple properties of an i^{th} system measured under different experimental conditions c_j. In general, PTML tries to predict an objective function $f(v_{ij})_{obs}$ obtained from the experimental

[1] Duda, R.; Hart, P. Stork, D. *Pattern Classification*; 2nd ed. Wiley: New York, 2001.
[2] MacKay, D. *Information theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, 2005.
[3] Bishop, C. *Pattern Recognition and Machine Learning*; Springer: Berlin, 2006.
[4] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*; 2nd ed. Springer; New York, 2009.
[5] Ferreira da Costa, J.; Silva, D.; Caamaño Olga; Brea José M; Loza, M. I.; Munteanu, C. R.; Pazos, A.; García-Mera Xerardo; González-Díaz Humbert. Perturbation Theory/machine Learning Model of Chembl Data for Dopamine Targets: Docking, Synthesis, and Assay of New. *Acs Chemical Neuroscience*, **2018**, 9 (11), 2572–2587.
[6] Blay, V.; Yokoi, T.; González-Díaz Humbert. Perturbation Theory–machine Learning Study of Zeolite Materials Desilication. *Journal of Chemical Information and Modeling*, **2018**, 58 (12), 2414–2419.

experimentales c_j. En general PTML intenta predecir una función objetivo $f(v_{ij})_{obs}$ obtenida a partir del valor experimental v_{ij} y se obtiene como una función $f(vij) = f(s_i, c_i)k$ de la estructura del sistema (s_i) y las condiciones para un tipo k. Los modelos PTML pueden predecir múltiples propiedades del sistema al mismo tiempo (multi-salida y multi-objetivo) teniendo en cuenta las variaciones (perturbaciones) con respecto a un valor de referencia o esperado en múltiples variables de entrada usadas para cuantificar las condiciones experimentales $\boldsymbol{c_j} = (c_0, c_1, c_2, \ldots c_n)$ y otras variables estructurales ó descriptores moleculares $\boldsymbol{D_{ki}} = (D_0, D_1, D_2, \ldots D_n)$ usadas para cuantificar la estructura del sistema (s_i). En la figura 1 se muestra un modelo general para un determinado sistema. [7]

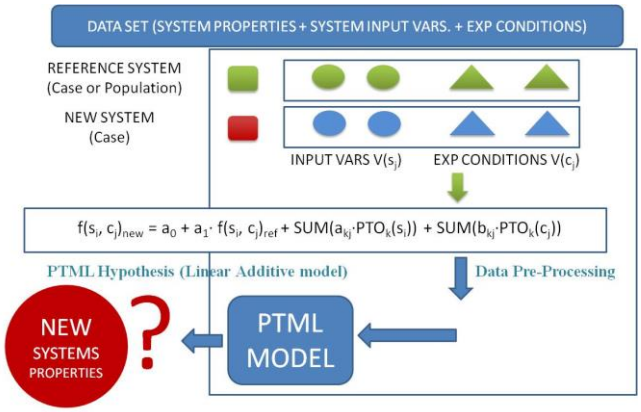

Figura 1: Modelo PTML genérico para un sistema determinado.[7]

La principal aplicación de este método es el estudio de sistemas moleculares (fármaco, proteína, vacuna, bio-marcador, nanopartículas, *etc*.) con múltiples valores v_{ij} de parámetros a optimizar, los cuáles han sido medidos en numerosos ensayos diferentes con condiciones de ensayo c_j distintas. A través de este modelo se puede obtener directamente los valores de una función calculada $f(v_{ij})_{calc} = f(s_i, c_j)_{calc}$ a partir de una función de referencia $f(v_{ij})_{ref}$ y los operadores de perturbación $PTO(s_i, c_j)_k$.

value v_{ij} and is obtained as a function $f(v_{ij}) = f(s_i, c_i)_k$ of the structure of the system (s_i) and the conditions for a type k. PTML models can predict multiple properties of the system at the same time (multi-output and multi-objective) taking into account the variations (disturbances) with respect to a reference or expected value in multiple input variables used to quantify the experimental conditions $\boldsymbol{c_j}=(c_0, c_1, c_2, \ldots c_n)$ and other structural variables or molecular descriptors $\boldsymbol{D_{ki}}=(D_0, D_1, D_2, \ldots D_n)$ used to quantify the structure of the system (s_i). Figure 1 shows a general model for a given system.

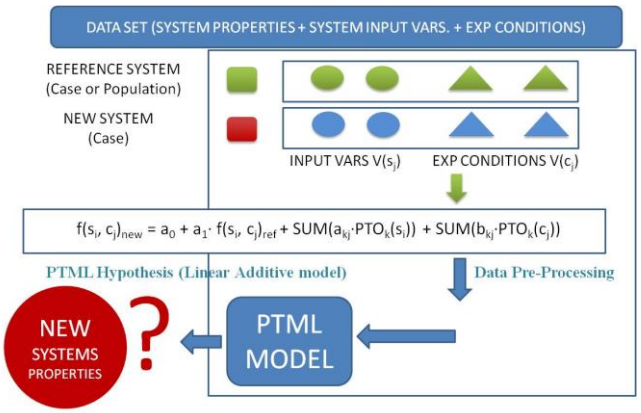

***Figure 1**: Generic PTML model for a given system.*

The main application of this method is the study of molecular systems (drug, protein, vaccine, biomarker, nanoparticles, etc.) with multiple values v_{ij} of parameters to optimize, which have been measured in numerous different tests with test conditions different c_j. Through this model it is possible to directly obtain the values of a calculated function $f(v_{ij})_{calc}=f(s_i, c_j)_{calc}$ from a reference function $f(v_{ij})_{ref}$ and the disturbance operators $PTO(s_i, c_j)_k$.

The classification models obtain the values of the probability $p(f(v_{ij}) = 1)_{calc}$ for a specific system i in a specific test, under known test conditions c_j. The probability $p(f(v_{ij}) = 1)_{calc}$ represents the probability of a system to be designed, it shows

[7] González-Díaz H. PTML: Perturbation-theory machine learningnotes, in Proceedings of the MOL2NET 2018, International Conference on Multidiciplinary Science, 4th edition, 15 January 2018-20 January 2019, MDPI: Basel, Switzerland.

los modelos de clasificación obtienen los valores de la probabilidad $p(f(v_{ij}) = 1)_{calc}$ para un determinado sistema i en un ensayo concreto, bajo unas condiciones de ensayo c_j conocidas. La probabilidad $p(f(v_{ij}) = 1)_{calc}$ representa la probabilidad de un sistema que se quiera diseñar, muestra niveles deseados $f(v_{ij})_{obs} = 1$ de los valores v_{ij} de un parámetro a optimizar.

Para una reacción química las propiedades a estudiar podrían ser el rendimiento Rdto(%) y el exceso enantiomérico *ee*(%). En nuestro caso nos enfocamos solo en el *ee*(%) por ser nuestras reacciones de α-amidoalquilación enantioselectivas y su Rdto(%) alto habitualmente. Además, si el modelo PTML buscado se plantea como función objetivo f(v_{ij}) = *ee*(%) estaríamos en presencia de un modelo de regresión y no se calcula la probabilidad p(f(*ee*(%)$_{ij}$)=1). Si el modelo intenta clasificar las reacciones como reacciones con exceso alto *ee*(%) > cut-off o exceso bajo ee(%) estaríamos en presencia de un modelo de clasificación, siendo cut-off un valor de corte definido por el investigador. Este modelo tendría como función objetivo la función f(v_{ij}) = f(*ee*(%) > cut-off)) = 1 ó 0. En ese caso si podemos obtener la probabilidad p(f(v_{ij}) =1)= p(f(*ee*(%)>cut-off)) de que el sistema tenga un nivel determinado de la propiedad *ee*(%) > cut-off. [8] En el desarrollo del programa MATEO para los modelos de regresión se usó la función objetivo f(v_{ij})= ee_R(%) =d_q·*ee*(%) que cuantifica el ee(%) de producto R siendo d_q = 1 cuando R es el enantiómero mayoritario y d_q=-1 cuando S es el enantiómero mayoritario.

Los modelos PTML añaden a los valores de la $f(v_{ij})_{ref}$ los valores de los operadores. Por ello, necesitamos calcular los valores de los PTOs (operadores de la teoría de perturbación-Perturbation Theory Operators) en el paso de procesamiento de datos. Esto nos permiten realizar

desired levels $f(v_{ij})_{obs} = 1$ of the values v_{ij} of a parameter to optimize.

For a chemical reaction, the properties to be studied could be the yield yield (%) and the enantiomeric excess ee(%). In our case we focus only on the ee(%) because our α-amidoalkylation reactions are enantioselective and their yield (%) is usually high. Furthermore, if the PTML model sought is proposed as an objective function f (v_{ij}) = ee(%), we would be in the presence of a regression model and the probability p(f (ee(%)$_{ij}$)=1) is not calculated. If the model attempts to classify the reactions as reactions with high excess ee(%)>cut-off or low excess ee(%) we would be in the presence of a classification model, with cut-off being a cut-off value defined by the researcher. This model would have as objective function the function f(v_{ij})=f(ee(%)>cut-off))= 1 or 0. In that case if we can obtain the probability p(f(v_{ij})=1)=p(f(ee(%)>cut-off)) that the system has a certain level of the ee(%)>cut-off property. In the development of the MATEO program for the regression models, the objective function f(v_{ij})=ee_R(%)=dq·ee(%) was used, which quantifies the ee(%) of product R where dq= 1 when R is the majority enantiomer and dq =-1 when S is the majority enantiomer.

The PTML models add the values of the operators to the values of $f(v_{ij})_{ref}$. Therefore, we need to calculate the values of the PTOs (Perturbation Theory Operators) in the data processing step. This allows us to carry out a process of merging information with variables and conditions from different sources. Moving Averages (MA), multi-condition MA (MMA), double MA, covariance operators, etc., are some examples of useful PTOs. Then, we can use Multiple Linear Regression (MLR), Linear Discriminant Analysis (LDA), or other linear ML techniques to find the PTML model.

[8] Bediaga, H.; Arrasate, S.; González-Díaz Humbert. Ptml Combinatorial Model of Chembl Compounds Assays for Multiple Types of Cancer. *Acs Combinatorial Science* **2018**, 20 (11), 621–632.

un proceso de fusión de información con variables y condiciones de diferentes fuentes. Los promedios móviles-Moving Averages (MA), los MA de múltiples condiciones (MMA), los MA dobles, los operadores de covarianza, etc., son algunos ejemplos de PTOs útiles. Después, podemos utilizar la regresión lineal múltiple-Multiple Linear Regression (MLR), el análisis discriminante lineal[9] u otras técnicas de ML lineal para buscar el modelo PTML.

El software MATEO que hemos verificado está basado en modelos de regresión. Sin embargo, no implementa modelos de clasificación. En el caso de las reacciones químicas los modelos de clasificación suelen ser deseables para minimizar posibles errores en las mediciones experimentales y/o para obtener una respuesta final sobre el interés de la reacción. Por este motivo, además de los modelos de regresión implementados en el MATEO nos propusimos desarrollar modelos de clasificación PTML. Para crear el modelo PTML-LDA, se ha usado el análisis discriminante lineal, el cual es una técnica estadística de clasificación que posee aplicaciones para clasificar casos.[10,11] Esta técnica es de especial interés cuando existen problemas de precisión en la medición de la variable experimental observada que dificultan la obtención de modelos de regresión, como es el caso del $ee_R(\%)$. Para usar esta técnica debemos discretizar la variable continua $ee_R(\%)$ transformándola en una variable discreta ó booleana. Inicialmente, se propusieron dos alternativas para el desarrollo del modelo. La primera se basa en la clasificación de los conjuntos de datos en tres clases, definidos por la variable o función objetivo que toma los valores $f(ee_R(\%))_{obs}$ 1, 0, ó -1. En este caso, la

The MATEO software that we have verified is based on regression models. However, it does not implement classification models. In the case of chemical reactions, classification models are usually desirable to minimize possible errors in experimental measurements and / or to obtain a final answer on the interest of the reaction. For this reason, in addition to the regression models implemented in MATEO, we set out to develop PTML classification models. To create the PTML-LDA model, linear discriminant analysis has been used, which is a statistical classification technique that has applications to classify cases. This technique is of special interest when there are precision problems in the measurement of the observed experimental variable that make it difficult to obtain regression models, such as the $ee_R(\%)$. To use this technique, we must discretize the continuous variable $ee_R(\%)$ transforming it into a discret or Boolean variable. Initially, two alternatives were proposed for the development of the model. The first is based on the classification of the data sets into three classes, defined by the objective variable or function that takes the values $f(ee_R(\%))_{obs}$ 1, 0, or -1. In this case, the observed function would be $f(ee_R(\%))_{obs}=1$ when $ee_R(\%)>$ cut-off, otherwise $f(ee_R(\%))_{obs} = -1$ if $ee_R(\%) <$-cut-off, otherwise $f(ee_R(\%))_{obs} = 0$ (cut-off> eeR (%)> -cut-off). The distribution in these groups is possible by introducing two limit values, also known as Cut-off, one of them positive and the other negative. The "-1" group is indicative of the excess S enantiomer, the "0" when it is an inefficient reaction including racemic mixtures and others (cut-off>$ee_R(\%)$>-cut-off). The value of the observed objective function f

[9] Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science*, Industry and Data Mining; StatSoft: Tulsa, 2006.

[10] Belhumeour, P.N.; Hespanha, J.P.;Kriegman, D.J.;. Eigenfaces vs. Fisherfaces: Recognitionusing class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intel-ligence*, **1997,** 19(7), 711–720.

[11] Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Curr. Top. Med. Chem*. **2013**, 13, 1713−41.

función observada sería $f(ee_R(\%))_{obs}$ = 1 cuando $ee_R(\%)$ > cut-off, sino $f(ee_R(\%))_{obs}$ = -1 Sí $ee_R(\%)$ < -cut-off, de otro modo $f(ee_R(\%))_{obs}$ = 0 (cut-off > $ee_R(\%)$ > -cut-off). La distribución en estos grupos es posible por la introducción de dos valores límites, también conocido como Cut-off, uno de ellos positivo y el otro negativo. El grupo "-1" es indicativo del enantiómero S en exceso, el "0" cuando se trata de una reacción no eficiente incluyendo las mezclas racémicas y otras (cut-off > $ee_R(\%)$ > -cut-off). El valor de la función objetivo observada $f(ee_R(\%))$ = 1 corresponde al caso de exceso de enantiómeros R. Además, en este modelo no se transformó la función de referencia (primera variable input). Por tanto se usó como variable de referencia la misma que en los modelos de regresión anteriores $f(ee_R(\%))_{ref} = ee_R(\%)_{ref}$.

El modelo obtenido a partir de esta estrategia resulta tedioso e inviable conseguir tanto para la prueba de entrenamiento como de confirmación un porcentaje mayor de 70 y lograr porcentajes de especificidad y de sensibilidad equitativos. Por esta razón, se descartó esta opción.

La segunda posibilidad al igual que el caso anterior hace uso del "Cut-off". Esta opción es más sencilla porque permite ordenar el conjunto de datos en dos grandes clases que son definidos por la función objetivo: $f(ee_R(\%))_{obs}$ = 0 ó 1. En el caso de $f(ee_R(\%))_{obs}$ = 0 es cuando la reacción tiene un bajo $ee_R(\%)$ (cut-off > $ee_R(\%)$ > -cut-off). Pero, cuando $f(ee_R(\%))_{obs}$ = 1 hay dos sub-casos. Los sub-casos son cuando hay exceso de R ó exceso de S. Para poder diferenciar estos dos sub-grupos de $f(ee_R(\%))_{obs}$ = 1 se modificó la función de referencia en el input del modelo. En este nuevo modelo PTML-LDA la nueva función de referencia es $f(ee_R(\%))_{ref} = d_q \cdot ee_R(\%)_{ref}$.

Con el software STATISTICA, se ha conseguido obtener estos modelos. Este programa tiene implementadas múltiples técnicas para la selección de variables. Los más destacables son "All effects",

$(ee_R(\%))$=1 corresponds to the case of excess of R enantiomers. Furthermore, the reference function (first input variable) was not transformed in this model. Therefore, the same reference variable was used as in the previous regression models $f(ee_R(\%))_{ref} = ee_R(\%)_{ref}$.

The model obtained from this strategy is tedious and unfeasible to achieve a percentage greater than 70 for both the training and confirmation tests and achieve equitable specificity and sensitivity percentages. For this reason, this option was discarded.

The second possibility, like the previous case, makes use of the "Cut-off". This option is simpler because it allows ordering the data set into two large classes that are defined by the objective function: $f(ee_R(\%))_{obs}$ = 0 or 1. In the case of $f(ee_R(\%))_{obs}$=0 is when the reaction has a low $ee_R(\%)$(cut-off>$ee_R(\%)$>-cut-off). But, when $f(ee_R(\%))_{obs}$ = 1 there are two sub-cases. The sub-cases are when there is an excess of R or an excess of S. In order to differentiate these two sub-groups of $f(ee_R(\%))_{obs}$ = 1, the reference function was modified in the input of the model. In this new PTML-LDA model the new reference function is $f(ee_R(\%))_{ref} = dq \cdot ee_R(\%)_{ref}$.

With the STATISTICA software, it has been possible to obtain these models. This program has implemented multiple techniques for the selection of variables. The most noteworthy are "All effects", which gives the user the option to choose between the different variables that he wishes to include based on expert criteria. On the other hand, "Forward-Stepwise" makes an automatic selection of variables based on the fact that the software itself chooses the variables by doing a Fisher (F) test. For the construction of the chemoinformatic model of the present work, the first option was used, where it is possible to choose the most important perturbations in different reaction conditions. In addition, the model uses 75% of the data sets for

que da opción al usuario elegir entre las distintas variables que desea incluir basándose en criterios de experto. Por otro lado, "Forward-Stepwise"[12], hace una selección automática de variables que se basa en que el propio software escoge las variables haciendo un test de Fisher (F)[13] . Para la construcción del modelo quimioinformático del presente trabajo, se recurrió a la primera opción, donde se posibilita la elección de las perturbaciones más importantes en diferentes condiciones de reacción. Además, el modelo utiliza el 75% de los conjuntos de datos para el entrenamiento del modelo y el 25% restante para la confirmación.

Por otra parte, la función devuelta por el modelo pertenece a la función $f(ee_R(\%))_{calc}$, la cual proporciona un resultado numérico y coeficientes de las variables ingresadas. Adicionalmente, para conseguir un modelo idóneo, se debe tener en cuenta los siguientes puntos:

- Las series previstas deben estar en el rango 70-95% tanto para la prueba de entrenamiento como la de confirmación.
- Los porcentajes de especificidad (0) y de sensibilidad (1) deben ser equilibrados. En casos de que no lo sean, puede conllevar a errores, ya que el modelo predice correctamente uno de ellos (bien sea la especificidad o sensibilidad) y malamente el otro.

Tras el entrenamiento y la validación de los modelos quimioinformáticos, se obtienen los siguientes resultados de mayor relevancia. Se muestra a continuación la matriz de clasificación utilizando el cut-off de 80, 90 y 96 respectivamente.

model training and the remaining 25% for confirmation.

On the other hand, the function returned by the model belongs to the function $f(ee_R(\%))_{calc}$ *, which provides a numerical result and coefficients of the entered variables. Additionally, to achieve an ideal model, the following points must be taken into account:*

- *Predicted sets should be in the 70-95% range for both the training and confirmation tests.*
- *The percentages of specificity (0) and sensitivity (1) must be balanced. In cases where they are not, it can lead to errors, since the model correctly predicts one of them (either the specificity or sensitivity) and the other poorly.*

After training and validation of the chemoinformatic models, the following most relevant results are obtained. The classification matrix is shown below using the cut-off of 80, 90 and 96 respectively.

***Table 1**: Training and validation series using the cut-off of 80, 90 and 96 respectively.*

Expected value	**Stadistic Para.**	**%**	$f(ee_R(\%))_{pred}$ **= 0**	$f(ee_R(\%))_{pred}$ **= 1**
Cut-off=80				
Training serie				
$f(ee_R(\%))_{obs}$	Sp(%)	100	49329	0
$f(ee_R(\%))_{obs}$	Sn(%)	82,8	5062	24341
Total	Ac(%)	93,6	54391	24341
Validation serie				

[12] Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science*, Industry and Data Mining. StatSoft: Tulsa, 2006.

[13] Fisher, R. A. On the interpretation of X^2 from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society*, **1922**, 85 (1), 87-94.

Tabla 1: Serie de entrenamiento y validación empleando el cut-off de 80, 90 y 96 respectivamente.

Valor esperado	**Para. estad**	**%**	$f(ee_R(\%))_{pred}$ **= 0**	$f(ee_R(\%))_{pred}$ **= 1**
Cut-off=80				
Serie de entrenamiento				
$f(ee_R(\%))_{obs}$	Sp(%)	100	49329	0
$f(ee_R(\%))_{obs}$	Sn(%)	82,8	5062	24341
Total	Ac(%)	93,6	54391	24341
Serie de validación				
$f(ee_R(\%))_{obs}$	Sp(%)	100	18767	0
$f(ee_R(\%))_{obs}$	Sn(%)	81,3	1893	8240
Total	Ac(%)	93,4	20660	8240
Cut-off=90				
Serie de entrenamiento				
$f(ee_R(\%))_{obs}$	Sp(%)	88,8	56518	7148
$f(ee_R(\%))_{obs}$	Sn(%)	79,1	3148	11918
Total	Ac(%)	86,9	59666	19066
Serie de validación				
$f(ee_R(\%))_{obs}$	Sp(%)	89,5	21065	2481
$f(ee_R(\%))_{obs}$	Sn(%)	75,6	1309	4045
Total	Ac(%)	86,9	22374	6526
Cut-off=96				
Serie de entrenamiento				
$f(ee_R(\%))_{obs}$	Sp(%)	90,8	69759	7029
$f(ee_R(\%))_{obs}$	Sn(%)	84,1	310	1634
Total	Ac(%)	90,7	70069	8663
Serie de validación				
$f(ee_R(\%))_{obs}$	Sp(%)	90,9	25695	2557
$f(ee_R(\%))_{obs}$	Sn(%)	87,8	79	569
$f(ee_R(\%))_{obs}$	Sp(%)	100	18767	0
$f(ee_R(\%))_{obs}$	Sn(%)	81,3	1893	8240
Total	Ac(%)	93,4	20660	8240
Cut-off=90				
Training serie				
$f(ee_R(\%))_{obs}$	Sp(%)	88,8	56518	7148
$f(ee_R(\%))_{obs}$	Sn(%)	79,1	3148	11918
Total	Ac(%)	86,9	59666	19066
Validation serie				
$f(ee_R(\%))_{obs}$	Sp(%)	89,5	21065	2481
$f(ee_R(\%))_{obs}$	Sn(%)	75,6	1309	4045
Total	Ac(%)	86,9	22374	6526
Cut-off=96				
Training serie				
$f(ee_R(\%))_{obs}$	Sp(%)	90,8	69759	7029
$f(ee_R(\%))_{obs}$	Sn(%)	84,1	310	1634
Total	Ac(%)	90,7	70069	8663
Validation serie				
$f(ee_R(\%))_{obs}$	Sp(%)	90,9	25695	2557
$f(ee_R(\%))_{obs}$	Sn(%)	87,8	79	569
Total	Ac(%)	90,9	25774	3126

Sp(%)=Specificity, Sn(%)= Sensibility y Ac(%)=Accuracy

The equations of the model from the linear discriminant analysis in reference to table 1 are revealed in the following table.

Total	Ac(%)	90,9	25774	3126

Sp(%)=Especificidad, Sn(%)= Sensibilidad y Ac(%) = Precisión

Se revela en la siguiente tabla las ecuaciones del modelo provenientes del análisis discriminante lineal en referencia a la tabla 1:

Tabla 2: Ecuaciones PTML empleando el análisis discriminante lineal.

"Cut-off"	**Ecuación**
80	$f(ee_R(\%))_{calc} = -4{,}17141 + 0{,}089 \cdot dq \cdot ee(\%)_{ij}(Rcat)_r + 0{,}03469 \cdot D(Load\%)_{qr} - 0{,}00056 \cdot D(T(°C))_{qr} + 0{,}00008 \cdot D(t(h))_{qr} + 2{,}30282 \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) + 2{,}23432 \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 1{,}32894 \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) + 2{,}29281 \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 0{,}68652 \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) - 2{,}30417 \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) + 1{,}60734 \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) - 0{,}71892 \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) + 0{,}17903 \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right) - 0{,}10199 \cdot dq \cdot D(Load\%)_{qr} - 0{,}06888 \cdot dq \cdot D(t(h))_{qr} + 0{,}04128 \cdot dq \cdot D(t(h))_{qr} - 14{,}78840 \cdot dq \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 85{,}94092 \cdot dq \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 14{,}48162 \cdot dq \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 107{,}42651 \cdot dq \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 5{,}74460 \cdot dq \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 60{,}98315 \cdot dq \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 2{,}45492 \cdot dq \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) + 0{,}58637 \cdot dq \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) - 1{,}92771 \cdot dq \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right)$
90	$f(ee_R(\%))_{calc} = -3{,}8567 + 0{,}0364 \cdot dq \cdot ee(\%)_{ij}(Rcat)_r + 0{,}0221 \cdot D(Load\%)_{qr} - 0{,}00018 \cdot D(T(°C))_{qr} + 0{,}0022 \cdot D(t(h))_{qr} + 1{,}2066 \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 27{,}2183 \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) + 1{,}3741 \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 7{,}6558 \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 0{,}5612 \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 2{,}2312 \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) + 0{,}7070 \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) - 0{,}2679 \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) + 0{,}0063 \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right) - 0{,}0615 \cdot dq \cdot D(Load\%)_{qr} - 0{,}0386 \cdot dq \cdot D(t(h))_{qr} +$

Table 2: *PTML equations using linear discriminant analysis.*

"Cut-off"	**Equations**
80	$f(ee_R(\%))_{calc} = -4{,}17141 + 0{,}089 \cdot dq \cdot ee(\%)_{ij}(Rcat)_r + 0{,}03469 \cdot D(Load\%)_{qr} - 0{,}00056 \cdot D(T(°C))_{qr} + 0{,}00008 \cdot D(t(h))_{qr} + 2{,}30282 \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) + 2{,}23432 \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 1{,}32894 \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) + 2{,}29281 \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 0{,}68652 \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) - 2{,}30417 \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) + 1{,}60734 \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) - 0{,}71892 \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) + 0{,}17903 \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right) - 0{,}10199 \cdot dq \cdot D(Load\%)_{qr} - 0{,}06888 \cdot dq \cdot D(t(h))_{qr} + 0{,}04128 \cdot dq \cdot D(t(h))_{qr} - 14{,}78840 \cdot dq \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 85{,}94092 \cdot dq \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 14{,}48162 \cdot dq \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 107{,}42651 \cdot dq \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 5{,}74460 \cdot dq \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 60{,}98315 \cdot dq \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 2{,}45492 \cdot dq \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) + 0{,}58637 \cdot dq \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) - 1{,}92771 \cdot dq \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right)$
90	$f(ee_R(\%))_{calc} = -3{,}8567 + 0{,}0364 \cdot dq \cdot ee(\%)_{ij}(Rcat)_r + 0{,}0221 \cdot D(Load\%)_{qr} - 0{,}00018 \cdot D(T(°C))_{qr} + 0{,}0022 \cdot D(t(h))_{qr} + 1{,}2066 \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 27{,}2183 \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) + 1{,}3741 \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 7{,}6558 \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 0{,}5612 \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 2{,}2312 \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) + 0{,}7070 \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) - 0{,}2679 \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) + 0{,}0063 \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right) - 0{,}0615 \cdot dq \cdot D(Load\%)_{qr} - 0{,}0386 \cdot dq \cdot D(t(h))_{qr} + 0{,}0229 \cdot dq \cdot D(t(h))_{qr} - 20{,}0419 \cdot dq \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 85{,}9859 \cdot dq \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 4{,}6302 \cdot dq \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 93{,}3769 \cdot dq \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) + 7{,}1042 \cdot dq \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 43{,}6042 \cdot dq \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 0{,}7620 \cdot dq \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) + 5{,}0707 \cdot dq \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) - 0{,}4309 \cdot dq \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right)$

	$0{,}0229 \cdot dq \cdot D(t(h))_{qr} - 20{,}0419 \cdot dq \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 85{,}9859 \cdot dq \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 4{,}6302 \cdot dq \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 93{,}3769 \cdot dq \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) + 7{,}1042 \cdot dq \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 43{,}6042 \cdot dq \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 0{,}7620 \cdot dq \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) + 5{,}0707 \cdot dq \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) - 0{,}4309 \cdot dq \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right)$
96	$f(ee_R(\%))_{calc} = -7{,}7123 + 0{,}0336 \cdot dq \cdot ee(\%)_{ij}(Rcat)_r - 0{,}0078 \cdot D(Load\%)_{qr} + 0{,}0030 \cdot D(T(°C))_{qr} - 0{,}0045 \cdot D(t(h))_{qr} + 4{,}3962 \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) + 33{,}3461 \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) + 1{,}9379 \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 10{,}8477 \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 1{,}5537 \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) - 1{,}1257 \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 0{,}0715 \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) - 1{,}1387 \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) - 0{,}0045 \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right) + 0{,}0787 \cdot dq \cdot D(Load\%)_{qr} + 0{,}0112 \cdot dq \cdot D(t(h))_{qr} + 0{,}0061 \cdot dq \cdot D(t(h))_{qr} - 21{,}2697 \cdot dq \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 87{,}6358 \cdot dq \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 21{,}8067 \cdot dq \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) + 64{,}4666 \cdot dq \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) + 8{,}0637 \cdot dq \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 1{,}2638 \cdot dq \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 3{,}3291 \cdot dq \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) + 5{,}2291 \cdot dq \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) + 0{,}2695 \cdot dq \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right)$

96	$f(ee_R(\%))_{calc} = -7{,}7123 + 0{,}0336 \cdot dq \cdot ee(\%)_{ij}(Rcat)_r - 0{,}0078 \cdot D(Load\%)_{qr} + 0{,}0030 \cdot D(T(°C))_{qr} - 0{,}0045 \cdot D(t(h))_{qr} + 4{,}3962 \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) + 33{,}3461 \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) + 1{,}9379 \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) - 10{,}8477 \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) - 1{,}5537 \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) - 1{,}1257 \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 0{,}0715 \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) - 1{,}1387 \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) - 0{,}0045 \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right) + 0{,}0787 \cdot dq \cdot D(Load\%)_{qr} + 0{,}0112 \cdot dq \cdot D(t(h))_{qr} + 0{,}0061 \cdot dq \cdot D(t(h))_{qr} - 21{,}2697 \cdot dq \cdot D\left(aPolar_{Cuns_{Cat_{Mean}}}\right) - 87{,}6358 \cdot dq \cdot D\left(aPolar_{HetNox_{Prod_{Mean}}}\right) - 21{,}8067 \cdot dq \cdot D\left(EA_{Csat_{Prod_{Mean}}}\right) + 64{,}4666 \cdot dq \cdot D\left(EA_{HetNoX_{Cat_{Mean}}}\right) + 8{,}0637 \cdot dq \cdot D\left(SAe_{Het_{Nuc_{Mean}}}\right) + 1{,}2638 \cdot dq \cdot D\left(SAe_{HetNoX_{Cat_{Mean}}}\right) - 3{,}3291 \cdot dq \cdot D\left(Vvdw_{All_{Sub_{Mean}}}\right) + 5{,}2291 \cdot dq \cdot D\left(Zv_{Cuns_{Cat_{Mean}}}\right) + 0{,}2695 \cdot dq \cdot D\left(Zv_{Cuns_{Solv_{Mean}}}\right)$