UNIVERSIDADE DA CORUÑA

Facultad de Informática
Departamento de Tecnologías de la Información y las Comunicaciones

Estudio de Aplicabilidad de Técnicas de Inteligencia Artificial en el Sector Agropecuario

Tesis Doctoral

Directores:
Daniel Rivero Cebrián
Enrique Fernández Blanco

Doctorando:
Iván Ramírez Morales

A Coruña, noviembre 2017

UNIVERSIDADE DA CORUÑA

Dr. Daniel Rivero Cebrián, profesor en el área de Ciencias de la Computación e Inteligencia Artificial, perteneciente al Departamento de Computación, Facultad de Informática, Universidade da Coruña

Y

Dr. Enrique Fernández Blanco, profesor en el área de Ciencias de la Computación e Inteligencia Artificial, perteneciente al Departamento de Computación, Facultad de Informática, Universidade da Coruña.

HACEN CONSTAR QUE:

La memoria “Estudio de aplicabilidad de técnicas de inteligencia artificial en el sector agropecuario” ha sido realizada por D. Iván Ramírez Morales, bajo nuestra dirección en el Departamento de Computación, y constituye la Tesis Doctoral que presenta para optar al Grado de Doctor en Informática de la Universidade da Coruña.

A Coruña, 3 de noviembre de 2017

Fdo: Daniel Rivero Cebrián

Fdo: Enrique Fernández Blanco

Agradecimientos

En primer lugar, agradezco a Dios, por poner en mi camino a personas y momentos que han hecho de este viaje llamado vida, una gran experiencia. De manera especial a mis padres y mi hermana que me brindaron un hogar lleno de amor y confianza, sus enseñanzas basadas en principios y valores, me han dado desde siempre las herramientas para avanzar en el marco del respeto hacia los demás. Gracias a ustedes porque desde que éramos pequeños, a costa del dolor de la distancia, nos brindaron la oportunidad de expandir las fronteras de nuestra comunidad, para ir a explorar el mundo y de esta manera, aprender de las nuevas realidades.

A mi compañera de vida, la que en un momento mágico del que fueron testigos de honor nuestros amigos y familiares, decidió emprender conmigo, en el proyecto más hermoso, uno en el que invertimos nuestro tiempo y fusionamos nuestros deseos de futuro. Ya han pasado siete años desde aquella decisión y solo tengo palabras de gratitud a ti por ser esa voz que sabe escoger las palabras precisas, para levantarme de nuevo sin importar cuantas veces caiga.

A mi hija, la creación más sublime de la que he podido ser parte y testigo. Es gracias a ella que me motiva a luchar cada día para que el mundo sea un lugar mejor. Aunque ahora eres muy pequeñita quiero agradecerte porque durante estos años siempre me has demostrado cuánto me extrañas, como muestra de tu amor. Ese sentimiento ha sido mi fuente de energía, y nada de lo logrado sería real sin ti.

A la comunidad de la Universidad Técnica de Machala, representada por su Consejo Universitario y su rector, porque ha sido gracias a su decisión que un importante grupo de profesores hemos podido llevar a cabo nuestros estudios de Doctorado. De igual manera, a los colegas profesores, a los estudiantes y a los administrativos de mi Universidad, por cada acción que realizaron para apoyarnos en este proceso de formación.

Los resultados que se exponen en el presente trabajo, no hubiesen sido alcanzados sin la guía de mis tutores, quienes compartieron sin límite, sus conocimientos y experiencia. A ustedes gracias porque jamás existió un día o una hora que fuese inadecuada para resolver mis dudas, jamás hubo tema alguno del que no se pueda hablar, por el contrario, me brindaron más que únicamente tutorías, me dieron su amistad.

A los compañeros del grupo de investigación RNASA, y de manera especial a su Director, quien tuvo la gentileza de invitarme a ser parte del mismo y estuvo siempre pendiente de que mi proceso de formación sea de calidad.

A las autoridades de los dos gobiernos que estuvieron al frente de la Universidad de la Coruña durante estos tres años, gracias a su mística de trabajo, nos hemos sentido como en casa. A los amigos de la Facultad de Informática, del Departamento de Tecnologías de la Información y las Comunicaciones, del programa de Doctorado que cursé, muchas gracias por todas las facilidades.

Siento la satisfacción de haber cumplido un compromiso que asumí al iniciar mi investigación doctoral, que estuvo llena de retos. Este logro quiero recibirlo con humildad, y dedicarlo a todas las personas que me aportaron para hacerlo realidad. ¡Muchas gracias a todos!

Lista de abreviaturas

ACC:	Exactitud, como abreviatura de *accuracy.*
ANN:	Red(es) de Neuronas Artificiales, como abreviatura de *Artificial Neural Network(s).*
ANOVA:	Análisis de Varianza, como abreviatura de *Analysis of Variance*.
BL:	*Beer – Lambert.*
CV:	Validación Cruzada, como abreviatura de *cross-validation.*
DE:	Desviación Estándar.
FN:	Falsos Negativos.
FP:	Falsos Positivos.
FI:	Intervalo de pronóstico, como abreviatura de *Forecasting Interval*.
FS:	Selección de Características, como abreviatura de *Feature Selection.*
FSD:	Primera derivada espectral, como abreviatura de *First Spectral Derivative.*
SSD:	Segunda derivada espectral, como abreviatura de *Second Spectral Derivative.*
GmbH:	Sociedad con responsabilidad limitada, como abreviatura del alemán Gesellschaft mit beschränkter Haftung.
HSD:	Diferencia significativa, como abreviatura de Honest Significant Difference.
I+D+i	Investigación, desarrollo e innovación.
IA:	Inteligencia artificial.
ML:	Aprendizaje máquina, como abreviatura de *Machine Learning.*
MLP:	Perceptrón multicapa, como abreviatura de *Multi-Layer Perceptron.*
NIR:	Reflectancia del Infrarojo Cercano, como abreviatura de *Near Infrared Reflectance.*
nm:	Nanómetros.
PCA:	Análisis de componentes principales, abreviatura de *Principal Component Analysis.*
PLS:	Cuadrados mínimos parciales, como abreviatura de *Parcial Least Squares.*
PPV:	Valor Predictivo Positivo, como abreviatura de *PPV.*
RBF:	Función de base radial, como abreviatura de *Radial Basis Function.*
RMSE:	Error cuadrático medio, como abreviatura de *Root Mean Squared Error.*
SEN:	Sensibilidad, como abreviatura de *SEN.*
SNV	Variación normalizada estándar, como abreviatura de *Standard Normal Variate.*
SPC:	Especificidad, como abreviatura de *SPC.*
SVM:	Máquina(s) de Soporte Vectorial, como abreviatura de *Support Vector Machine(s).*
SVR:	Regresión por Soporte de Vectores, como abreviatura de *Support Vector Regression.*
TN:	Verdaderos Negativos, como abreviatura de *True Negative.*
TP:	Verdaderos Positivos, como abreviatura de *True Possitive.*
WS:	Tamaño de la ventana temporal, como abreviatura de *Window Size.*

Índice

PRIMERA PARTE

SEGUNDA PARTE

Índice de Figuras

Índice de Tablas

Resumo

O aprendizaje máquina é unha rama da intelixencia artificial (IA) que utiliza algoritmos para realizar tarefas, sen que se teña programado explícitamente. Para o seu funcionamento require un proceso de formación e validación baseado en exemplos.

Nesta tese proponse estudar a aplicabilidade dalgunhas técnicas de IA na produción agrícola. A tese é apoiada por tres publicacións cun importante factor de impacto JCR. Dous deles fan referencia a unha base de datos de produción de aves de ovos e outra a unha base de datos sobre a industrialización da cana de azucre.

Na produción avícola estas técnicas foron estudadas para a alerta precoz dos problemas na curva de produción. En canto á aplicación destas técnicas no proceso industrial de cana de azucre, optimizáronse os modelos de calibración dos espectros NIR para o control de calidade nunha fábrica de azucre.

Usáronse máquinas de soporte vectorial e redes neuronais artificiais. A aplicación destas técnicas ten un alto potencial de uso na produción agrícola, xa que posibilita o desenvolvemento de sistemas intelixentes de apoio ás decisións produtivas.

Resumen

El aprendizaje máquina es una rama de la inteligencia artificial (IA) que utiliza algoritmos para realizar tareas, sin que hayan sido programados de manera explícita. Para su funcionamiento se requiere de un proceso de entrenamiento y validación en base a ejemplos.

En esta Tesis Doctoral, se propone estudiar la aplicabilidad de algunas técnicas de IA en la producción agropecuaria. El trabajo está respaldado por tres publicaciones con un importante factor de impacto JCR. Dos de ellas se refieren a una base de datos de producción avícola de huevos y la otra, a una base de datos de la industrialización de la caña de azúcar.

En la producción avícola estas técnicas fueron estudiadas para la alerta temprana de problemas en la curva de producción. En cuanto a la aplicación de estas técnicas en el proceso industrial de la caña de azúcar, se optimizó los modelos de calibración de los espectros NIR para el control de calidad en una fábrica de azúcar.

Se utilizó Máquinas de Soporte Vectorial y Redes de Neuronas Artificiales. La aplicación de estas técnicas tiene un alto potencial de uso en la producción agropecuaria, ya que posibilita el desarrollo de sistemas inteligentes de apoyo a las decisiones productivas.

Abstract

Machine learning is a branch of artificial intelligence that uses algorithms to perform tasks, without having been programmed explicitly. For its operation requires a process of training and validation based on examples.

In this thesis the application of artificial intelligence techniques in agricultural production is studied. As main result of the thesis, three articles has been published in journals with important JCR impact factors. Two of them refer to a database of poultry production of eggs and the other to a database of the industrialization of sugar cane.

In poultry production these techniques were studied for the early warning of problems in the production curve. For the application of these techniques in the industrial process of sugarcane, the calibration models of the NIR spectra for the quality control in a sugar factory were optimized.

In this work were used Support Vector Machines and Artificial Neural Networks. The application of these techniques has a high potential of use in the agricultural production, since it opens up the development of intelligent systems to support productive decisions.

Estructura de la Tesis Doctoral

Esta Tesis Doctoral se divide en dos partes. La primera de ellas está dedicada a la exposición del problema, la revisión del estado actual de la técnica, la descripción de las mejoras propuestas utilizando algoritmos de inteligencia artificial, la discusión de los resultados obtenidos, así como la presentación de las líneas de futuros trabajos. La segunda parte contiene las publicaciones científicas resultantes del trabajo desarrollado.

La primera parte, se divide en cinco secciones. En la primera sección se aborda una introducción al problema y se presenta la justificación y los objetivos de esta Tesis Doctoral. A continuación, en la sección 2 se revisan los conceptos básicos sobre el campo de aplicación de algoritmos de aprendizaje máquina en el sector agropecuario, en particular la aplicación de estos algoritmos en la producción avícola, y en la producción de caña de azúcar. Estas aplicaciones se utilizarán como base para la presente propuesta.

La sección 3 presenta una descripción de los materiales y la metodología utilizada en esta investigación doctoral. La sección 4 discute los resultados obtenidos. Finalmente, la sección 5 expone las conclusiones y las futuras líneas de investigación que surgen del desarrollo de esta Tesis Doctoral.

En la segunda parte, en la sección ANEXOS, se adjuntan las tres publicaciones científicas resultantes del trabajo experimental desarrollado en esta investigación. Las publicaciones referidas son:

- Iván Ramírez-Morales, Daniel Rivero Cebrián, Enrique Fernández Blanco, Alejandro Pazos Sierra, ***Early warning in egg production curves from commercial hens: A SVM approach***, *Computers and Electronics in Agriculture*, Volume 121, February 2016, Pages 169-179, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2015.12.009.
- Iván Ramírez-Morales, Daniel Rivero, Enrique Fernández-Blanco, Alejandro Pazos, ***Optimization of NIR calibration models for multiple processes in the sugar industry***, In *Chemometrics and Intelligent Laboratory Systems*, Volume 159, 2016, Pages 45-57, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2016.10.003.
- Iván Ramírez-Morales, Enrique Fernández-Blanco, Daniel Rivero, Alejandro Pazos, ***Automated early detection of drops in commercial egg production using neural networks***, *British Poultry Scienc*e, 2017, ISSN 0007-1668 https://doi.org/10.1080/00071668.2017.1379051

Capítulo:
I. INTRODUCCIÓN

Las proyecciones probabilísticas basadas en datos de la Organización de las Naciones Unidas indican que para el año 2050 la población mundial llegará cerca de los diez mil millones de habitantes (Azose, Ševčíková, y Raftery, 2016). En este escenario, el rol del sector agropecuario adquiere un protagonismo mayúsculo, en la provisión de alimentos y materias primas.

El sector agropecuario es un crisol de contrastes, con grandes producciones intensivas y pequeñas fincas familiares. No obstante, en ambos casos se genera una ingente cantidad de datos. Estos datos suelen registrarse y almacenarse para el análisis a corto plazo con técnicas básicas de procesamiento de la información (Stafford y Werner, 2003).

Es indudable que la gestión adecuada del conocimiento es fundamental para la correcta toma de decisiones. Contar con técnicas adecuadas de tratamiento de la información permite aprovechar la riqueza y diversidad de los datos, transformándolos en conocimiento relevante para la toma de decisiones de producción (Walker, 2002).

En la actividad agropecuaria actual, se ha generalizado el uso de estadísticas descriptivas elaboradas por expertos con la finalidad de interpretar la situación actual de las fincas y asesorar en la toma de decisiones. En el caso de la elaboración de modelos y simulaciones, generalmente son llevados a cabo por instituciones académicas y de investigación. Estas actividades representan un costo que no siempre puede ser cubierto por las empresas, las que usualmente trabajan con pequeñas muestras en base a diseños experimentales (El-Korany, El-Azhary, y Yehia, 2004; Perini y Susi, 2004)

La reducción de los costos y el incremento de los ingresos es una cuestión fundamental para las empresas en una economía de mercado globalizada. En este contexto, las empresas agropecuarias han adoptado de forma paulatina tecnologías que permiten la automatización de las tareas cotidianas (Radionov, Antonov, Makarov, y Orlov, 2015).

El análisis de los datos, es un proceso empírico que está basado en la experiencia de un grupo reducido de especialistas y en el conocimiento de los productores agropecuarios (Frost et al., 1997). Sin embargo, estos datos obtenidos en sistemas agropecuarios, se originan a partir de elementos biológicos que tienen una naturaleza compleja, dinámica, y no lineal, lo cual complica enormemente ese proceso de análisis.

Por este motivo, se requiere explorar soluciones con mayor capacidad de obtención, extracción y análisis de la información. Entre las nuevas tecnologías disponibles para dar respuesta a esta necesidad, destacan la inteligencia artificial, los sensores, la minería de datos y la visión por computadoras (Goel et al., 2003; Perini y Susi, 2004).

Hasta hace poco tiempo el término Inteligencia Artificial (IA) en el imaginario social de la comunidad no científica, guardaba una connotación relacionada con la ciencia ficción. Sin embargo, ahora es bastante común encontrar en el mercado dispositivos que aprovechan las ventajas de la IA, con el objetivo de mejorar la experiencia del usuario (Mariano, Lima, Alvarenga, Rodrigues, y Lacerda, 2014). Las formas en que se están aplicando estas tecnologías son variadas y entre ellas destacan: las recomendaciones de asistentes digitales, traducción instantánea, vehículos de conducción autónoma, reconocimiento facial, predicción de fraudes financieros y la estimación de condiciones idóneas de mercado (Müller y Bostrom, 2016).

La IA es una disciplina informática cuyas técnicas son cada vez más demandadas en diversos entornos, dada su capacidad de crear sistemas que emulan un comportamiento inteligente. De esta manera, es posible crear modelos de decisión en entornos eminentemente complejos y no lineales. Actualmente, estas técnicas representan una tendencia en países con un alto desarrollo tecnológico y con una gran inversión en I+D+i (Benítez, Escudero, y Kanaan, 2013; Bennett y Campbell, 2000).

Los algoritmos de aprendizaje máquina pueden ser entrenados de manera supervisada para el análisis de los datos agropecuarios. Los resultados del análisis elaborado por el algoritmo estrenado son similares a los de un experto. Esto es debido a que los algoritmos de ML son capaces de “extraer” las relaciones entre los datos a partir de ejemplos, en el proceso de entrenamiento (Mucherino, Papajorgji, y Pardalos, 2009).

A pesar de que el uso de técnicas de IA en sistemas productivos agropecuarios no es nuevo, existe un interés creciente por parte de los investigadores y de la industria. Estos algoritmos se han utilizado en tareas de clasificación (McQueen et al., 1995), diagnóstico (Wan y Bao, 2009; Yang y Su, 2008), de detección temprana (Rumpf et al., 2010), identificación de riesgos epidemiológicos (Hepworth, Nefedov, Muchnik, y Morgan, 2012), detección de preñez (Hempstalk, McParland, y Berry, 2015), control de calidad (Mitchell, 2014; Tomazzoli et al., 2015) entre otros trabajos.

Bajo los argumentos mencionados anteriormente se planteó la hipótesis de que el uso de técnicas de IA es aplicable como medio de soporte a la toma de decisiones basadas en datos en sistemas de producción agropecuaria.

1.1 Objetivo General

El principal objetivo de esta Tesis Doctoral es estudiar la aplicabilidad de las técnicas de IA en el análisis y procesamiento de datos de ámbito agropecuario, para el desarrollo de sistemas de soporte a las decisiones, que favorezcan la productividad y competitividad de las empresas agropecuarias.

1.2 Objetivos Específicos

Con la finalidad de que el estudio se realice en datos de producción en animales y en plantas, se propone el cumplimiento de los siguientes objetivos específicos:

- Desarrollar y evaluar modelos de alerta temprana de problemas en la curva de la producción de huevos de gallinas comerciales utilizando algoritmos de IA, para la ejecución de acciones oportunas en las granjas avícolas.
- Optimizar modelos de calibración global NIR orientados al mejoramiento del proceso de control de calidad de los parámetros °Brix y Sacarosa, en la industria azucarera.
- Establecer indicadores de rendimiento y confiabilidad de cada una de las técnicas de IA en el análisis y procesamiento de datos de ámbito agropecuario.

Capítulo:
II. ESTADO DEL ARTE

En este capítulo se describe el estado actual de la cuestión, así como base teórica de las técnicas de IA y su aplicabilidad en el sector agropecuario. Sobre estos conceptos se fundamenta la Tesis Doctoral. En las subsecciones de este capítulo, se realiza una revisión basada en las publicaciones existentes sobre el uso de algoritmos de aprendizaje supervisado en aplicaciones agrícolas y pecuarias. Además, esta sección, contextualiza la relevancia científica del presente trabajo de investigación.

2.1 Técnicas de Inteligencia Artificial

Un ámbito de mucho interés entre las técnicas de inteligencia artificial es el denominado aprendizaje máquina (machine learning - ML) o aprendizaje automático. Se trata de un conjunto de algoritmos capaces de realizar tareas complejas, sin necesidad de que sean programados de forma explícita (Bengio, 2009).

Estas técnicas adquieren un valor relevante en tareas de alta complejidad. Un ejemplo de ello es, programar todas las jugadas posibles en el ajedrez, de tal forma que una máquina sea capaz de vencer a un maestro de ajedrez humano. La complejidad de este juego fue calculada por Shannon (1950) igual a 10^{120}, como referencia de dicha cantidad, todos los átomos del universo se calcula que son igual a 10^{79}. En este tipo de tareas es preferible desarrollar un algoritmo que sea capaz de aprender a partir de ejemplos.

De forma general existen dos tipos de algoritmos de aprendizaje máquina: el supervisado y el no supervisado. El aprendizaje supervisado, se utiliza cuando existe conocimiento sobre las salidas deseadas, y se lleva a cabo un proceso de entrenamiento para obtener esas salidas deseadas. Por otra parte, cuando no se cuenta con información sobre las salidas esperadas, se suele aplicar técnicas de agrupamiento que no requieren de supervisión (Mucherino et al. 2009).

Una vez que un algoritmo ha sido entrenado, es capaz de aplicar el aprendizaje hacia nuevos datos (Hastie, Tibshirani, y Friedman, 2009). Estos algoritmos pueden ser utilizados en problemas de clasificación, regresión, predicción de series temporales, ajustes de curvas de calibración, entre otros. Su desempeño se basa en la calidad de

los patrones del conjunto de entrenamiento cuyas salidas deseadas han sido registradas (Palma y Marín, 2013; Benítez et al., 2013).

El proceso de entrenamiento se realiza con datos que han sido etiquetados previamente por un experto o en el caso de regresión, que han sido obtenidos con un instrumento de alta precisión. El conjunto de datos debe ser dividido en dos subconjuntos, uno para el entrenamiento, y otro conocido como test. Este último, se utiliza para realizar las pruebas del modelo con datos que no son parte del entrenamiento. Muchas veces se utiliza un tercer subconjunto de validación para mejorar la certeza de la confiabilidad en la selección del mejor modelo (Refaeilzadeh, Tang, y Liu, 2009).

Para la partición de los datos, se suele recurrir a técnicas como la validación cruzada también conocida como *cross-validation*. Al utilizar esta técnica se reduce la posibilidad de sobreentrenar un algoritmo, esto es, que obtenga un buen desempeño con los ejemplos conocidos, pero un mal desempeño con nuevas instancias de datos (Kuhn y Johnson, 2013).

Para la evaluación de un algoritmo de aprendizaje máquina se utiliza los datos del subconjunto test. El algoritmo a ser evaluado debe ser capaz de realizar estimaciones correctas en nuevas instancias de datos (Mucherino et al., 2009).

Debido a los efectos de la variación en la propia naturaleza de la técnica, en la partición y representatividad de datos, no se justifica el considerar simplemente el mejor resultado de las múltiples ejecuciones de un algoritmo. Por esto se recomienda presentar al menos la media y la desviación estándar de los resultados de las repeticiones. Además, es altamente recomendable realizar pruebas de hipótesis para verificar la significación de las diferencias entre las configuraciones de las ANN (Amiri, Niaki, y Moghadam, 2014; Hossain, Ayodele, Cheng, y Khan, 2016; Singh, Sarkar, y Nasipuri, 2015).

En el presente trabajo, se exploró principalmente dos técnicas de aprendizaje máquina: las Máquinas de Soporte Vectorial y las Redes de Neuronas Artificiales.

2.1.1 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (*Support Vector Machines* - SVM) se consideran unas de las principales técnicas de aprendizaje supervisado. Su aplicación está relacionada principalmente a problemas de clasificación y regresión (Benítez et al., 2013; Palma y Marín, 2013).

Los fundamentos de las SVM fueron desarrollados por Vapnik et al. (1997), en su trabajo sobre las teorías de aprendizaje estadístico, las cuales pretendían acotar el error de generalización en función de la complejidad del espacio de búsqueda. Aunque en un principio se diseñaron para resolver problemas de clasificación binaria, su aplicación se ha extendido a tareas de regresión, multi-clasificación y agrupamiento.

El actual estándar de las SVM fue propuesto por Vapnik y Cortes (1995), el objetivo es obtener modelos que estructuralmente tengan un riesgo bajo de cometer errores de estimación ante datos futuros.

Ante un problema de clasificación, esta técnica va dirigida a encontrar un hiperplano óptimo capaz de distribuir los datos en las clases a las que pertenecen. Intuitivamente parece obvio llegar a la conclusión de que ante un problema de clasificación lineal hay una alta probabilidad de obtener varias soluciones que logren clasificar los datos (Fernandez-Lozano et al., 2014)

El hiperplano óptimo usado para separar las dos clases puede ser definido a partir de una pequeña cantidad de datos del conjunto de entrenamiento llamados vectores de soporte, los que determinan el margen (Mucherino et al., 2009)

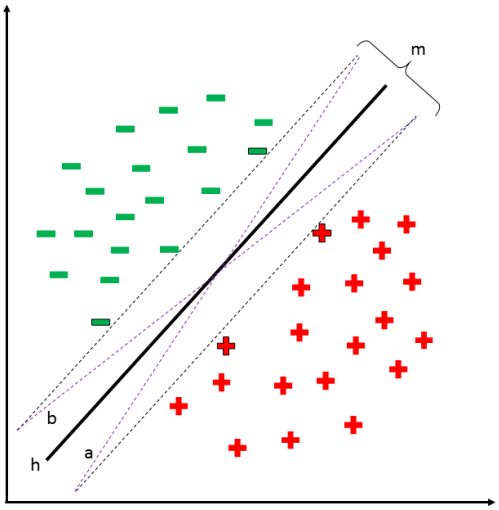


Figura 2.1 Representación de: hiperplano óptimo (h), alternativos (a, b) y los márgenes del modelo (m).

En la Figura 2.1 se ilustra el concepto de hiperplano explicado anteriormente, así como el concepto de margen óptimo del modelo. Como se puede observar existen varias posibilidades de hiperplanos alternativos (a y b). El hiperplano óptimo usado para separar las dos clases se define a partir de una pequeña cantidad de datos del conjunto de entrenamiento llamados vectores de soporte. Estos vectores de soporte son los que establecen también el margen del modelo. La elección del mejor hiperplano fue descrita por Vapnik y Kotz (1982) como la función lineal con el margen más amplio entre los vectores de soporte de ambas clases.

En la mayoría de problemas, los datos no son separables por una función lineal. Las funciones *kernel*, son utilizadas para transformar el espacio original multidimensional, en otro en el que las clases se pueden separar con una función lineal (Koch et al., 2012). El *kernel* es uno de los parámetros más importantes de las SVM.

Las SVM son entrenadas probando distintos *kernels* para seleccionar el que tenga el mejor desempeño (Mucherino et al. 2009). Entre los *kernels* más utilizados están el polinómico y el gaussiano (función de base radial), éste último cuenta con un parámetro *sigma* (σ) que ajusta el tamaño del *kernel*.

La búsqueda de parámetros óptimos de una SVM es fundamental en la construcción de un modelo de predicción para que sea preciso y estable (Prasoona et al., 2013; C.-H. Wu, Tzeng, y Lin, 2009). Los parámetros del *kernel* son ajustables en las SVM para controlar la complejidad de la hipótesis resultante y evitar el sobreajuste del modelo

(Cristianini y Shawe-Taylor, 2000; Devos, Ruckebusch, Durand, Duponchel, y Huvenne, 2009).

La modificación del parámetro C, afecta la calidad de la clasificación, ya que determina cuán severamente deben ser penalizados los errores de clasificación. De manera general valores muy altos de C pueden traer problemas de sobreajuste, reduciendo la capacidad de la SVM para generalizar (Mucherino et al. 2009).

Las SVM pueden ser utilizadas también en problemas regresión, esta versión de una SVM para regresión fue propuesta en 1997 por Vapnik et al., (1997). Este método se llama *Support Vector Regression* (SVR). El modelo depende sólo de un subconjunto de los datos (vectores de soporte), ya que la función de costos para la construcción del modelo no considera los puntos que se encuentren más allá del margen, asimismo la función de costo ignora cualquier los datos que estén cerca al modelo de predicción, dentro de un umbral ε (Basak, Pal, y Patranabis, 2007).

En el caso de las SVR se realiza la optimización de los parámetros C, *gamma* (*γ)* que es un parámetro del *kernel* RBF y *epsilon* (ε), determinan la complejidad de los límites y por lo tanto el desempeño del modelo. Se puede utilizar diferentes técnicas para optimizar estos parámetros (Devos et al., 2009; Jeng, 2006).

Las SVR se han aplicado en varios campos como series temporales (Zhang, Zhou, Chang, Yang, y Li, 2013), finanzas (J. Wu y Wei, 2007), aproximaciones de ingeniería en análisis complejos (Acar, Hudson, Miller, y Phillips, 2008), programación cuadrática convexa (Quan, Yang, Yao, y Ye, 2004), entre otras (Basak et al., 2007).

2.1.2 Redes de Neuronas Artificiales

Las Redes de Neuronas Artificiales (*Artificial Neural Network* - ANN) se definen como sistemas no lineales. Están inspiradas por el funcionamiento del sistema nervioso de los animales, el cual está compuesto por redes de neuronas biológicas que poseen bajas capacidades de procesamiento, sin embargo, toda su capacidad cognitiva se sustenta en la conectividad de éstas. De modo similar al mecanismo biológico, las ANN son capaces de realizar tareas complejas de clasificación, identificación, diagnóstico, optimización y predicción (Ahmad, 2011; Ponce, 2011).

Las ANN han atraído especialmente atención en los últimos años, sin embargo fueron McCulloch y Pitts, (1943) quienes presentaron el primer modelo de neurona artificial. Minsky y Papert (1969) escribieron el libro titulado Perceptrones: Una introducción a la geometría computacional. En este trabajo demostraron que el perceptrón sólo podía resolver funciones linealmente separables. De particular interés fue el hecho de que el perceptrón todavía no podía resolver las funciones XOR y XNOR. La crítica que realizaron en ese momento se considera la causa del poco interés en la investigación académica en redes neuronales artificiales hasta inicios de los años 80.

Hornik et al, (1989) planteó que las redes neuronales *feedforward* de múltiples capas ocultas, son capaces de aproximar cualquier función boreliana medible de un espacio de dimensión finita a otro en cualquier grado de precisión. Es por esto que las ANN se consideran aproximadores universales.

El interés en el uso de las redes neuronales va en aumento gracias a su naturaleza paralela, lo que hace que puedan aumentar su velocidad de cálculo. Las ANN además pueden adaptarse a los cambios en los datos y pueden realizar operaciones de filtrado más allá de las capacidades de técnicas de filtrado convencionales (Karayiannis y Venetsanopoulos, 2013).

Las ANN consisten en un número unidades de procesamiento (neuronas) independientes y simples, que están conectadas y generalmente organizadas en capas (Nürnberger, Pedrycz, y Kruse, 2002). Un buen ejemplo de esta organización, es el perceptrón multicapa (Multilayer Perceptron MLP). En dicha red, las conexiones suceden únicamente entre capas consecutivas, de manera general tiene una capa de entrada, una o múltiples capas ocultas y una capa de salida. La función de transferencia en las neuronas de la capa oculta y de la capa de salida usualmente es una sigmoidea, sin embargo, pueden estar presentes otras funciones como las lineales, las no lineales o las escalonadas (Kruse et al., 2013; Ruck, Rogers, Kabrisky, Oxley, y Suter, 1990).

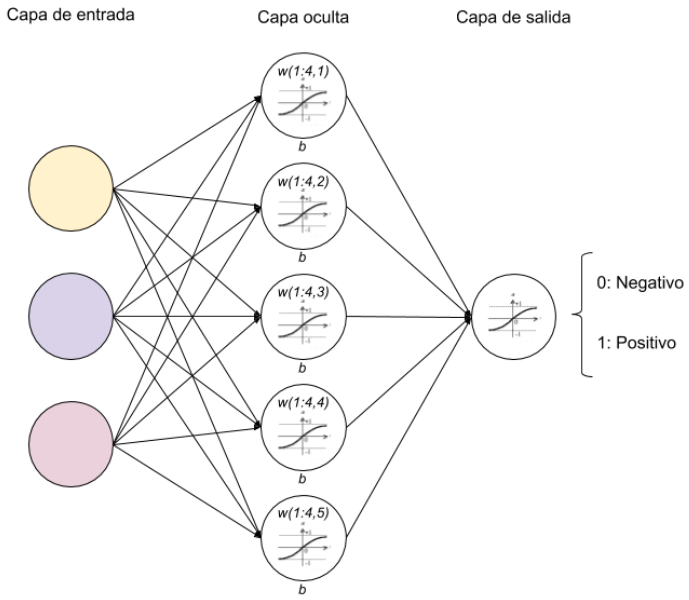


Figura 2.2. Representación de un perceptrón multicapa con una capa oculta.

En la Figura 2.2 se muestra una estructura característica del MLP, los datos (patrones de entrada) se proporcionan a la red a través de una capa que no realiza ningún proceso y simplemente envía esta información a las capas ocultas. El procesamiento es realizado en las capas ocultas y en la capa de salida (Mucherino et al., 2009). Cada neurona recibe señales de salida de las neuronas en la capa anterior y envía su salida a las neuronas de la capa siguiente. La capa de salida, recibe las entradas de las neuronas y proporciona los valores de salida que pueden ser aplicados a problemas de clasificación o de regresión (Gardner y Dorling, 1998; Kruse et al., 2013).

Uno de los métodos más usados para entrenar un MLP busca localizar el error mínimo utilizando un gradiente descendiente (Gardner y Dorling, 1998). Para el entrenamiento en primer lugar se inicializan aleatoriamente los pesos y el *bias* de las neuronas. A continuación, se determina la dirección de la pendiente más pronunciada (gradiente

descendiente) en el espacio de búsqueda, se modifican los pesos para dar un paso hacia delante, y se re-calcula el gradiente hasta llegar a un mínimo de la función (Kruse et al., 2013).

Para mejorar el desempeño de las ANN es necesario seleccionar una arquitectura adecuada, esto consiste en determinar el número de capas ocultas, el número de neuronas y la forma como estarán interconectadas. La arquitectura de red va a depender del problema a resolver, y no existe una regla o método que permita decidir cuál es la mejor. Generalmente la selección de la mejor arquitectura, resulta de un proceso empírico, en el que es necesario probar distintas alternativas hasta que se encuentra una que proporcione buenos resultados (Herrera, Hervas, Otero, y Sánchez, 2004; Rivero, Fernandez-Blanco, Dorado, y Pazos, 2011).

Actualmente las ANN son utilizadas en una amplia variedad de aplicaciones en distintos campos de las ciencias (Guo, Rivero, y Pazos, 2010; Kalhor, Rajabipour, Akram, y Sharifi, 2016; Raith et al., 2017; Samborska et al., 2014).

2.1.3 Entrenamiento con bases de datos no balanceadas

Muchos de los datos del mundo real no son equilibrados en términos de la proporción de ejemplos disponibles para cada clase. Este desequilibrio puede hacer que los algoritmos de aprendizaje máquina se sobreajusten a la clase más frecuente cometiendo errores en futuras predicciones (Blagus y Lusa, 2010; Huang, Hung, y Jiau, 2009).

Al utilizar algoritmos de ML para clasificación, se intenta realizar un entrenamiento de manera adecuada, para que sean capaces de generalizar las relaciones aprendidas de los datos de entrenamiento, hacia nuevos datos. Este análisis es cierto cuando la importancia de los errores es la misma en todas las clases. Sin embargo, en aplicaciones del mundo real los errores no suelen tener igual importancia. Un ejemplo es que, en medicina, evitar un diagnóstico falso negativo es más importante que evitar un diagnóstico falso positivo. Esto se puede argumentar ya que diagnosticar erróneamente como sano a un paciente que está enfermo, puede resultar en la pérdida de una vida, mientras que lo contrario no tiene la misma implicación y se pueden realizar pruebas confirmatorias (Zhou y Liu, 2006).

Una práctica común para prevenir el sobreajuste a la clase más frecuente en bases de datos no balanceadas, consiste en realizar una ponderación mediante un parámetro, la clasificación errónea en la clase que tiene menos ejemplos (Elkan, 2001; Pazzani et al., 1994). Esta técnica ha atraído actualmente mucha atención de la comunidad de ML. Sin embargo, existen pocos trabajos en los que se aplica este enfoque en el ámbito agropecuario (Zahirnia, Teimouri, Rahmani, y Salaq, 2015; Zhou y Liu, 2006).

2.2 Campo de aplicación: ML en el sector agropecuario

El uso de técnicas de ML en el sector agropecuario es un campo que ha tomado bastante notoriedad en la actualidad. A criterio del autor, existen muchas publicaciones que hacen alusión al uso de estas técnicas en los ámbitos más variados, sin embargo,

aún quedan muchas áreas por explorar, por lo que en los próximos años se convertirá en un *hot-topic* entre los científicos del área.

Entre los trabajos publicados sobre modelado, se encuentra el uso de redes de neuronas artificiales para modelar el secado de granos (Farkas, Reményi, y Biró, 2000) y de *Echinacea Angustifolia* (Erenturk, Erenturk, y Tabil, 2004), la simulación de daño en duraznos en una línea de transporte (Bielza, Barreiro, Rodríguez-Galiano, y Martín, 2003). Algunos sistemas refieren el uso de espectrómetros y sensores, junto con algoritmos de ML para la autenticación de uvas previo a la fermentación (Roussel, Bellon-Maurel, Roger, y Grenier, 2003)

Los sistemas de clasificación son bastante utilizados en el sector agropecuario. Es de especial interés un sistema basado en redes neuronales y lógica difusa que se aplican para la clasificación y graduación de manzanas según el tamaño (Shahin, Tollner, y McClendon, 2001). Otro sistema que resulta interesante se refiere a la clasificación de carnes en el que se utilizan tres algoritmos de ML (Díez et al., 2003).

En cuanto a los sistemas expertos para el soporte a las decisiones, se reportan trabajos como medio para mejorar y optimizar la productividad agropecuaria (Jones, 1993; Lokhorst y Lamaker, 1996). Un área de interés reciente es el diagnóstico de enfermedades. Por otro lado, A nivel comercial se encuentra referencias a un sistema de pronóstico de precios futuros para alimentos, utilizando técnicas de ML (Doganis, Alexandridis, Patrinos, y Sarimveis, 2006).

2.2.1 Campo de aplicación en una producción avícola

Para abordar el campo de aplicación de técnicas de ML en el sector avícola, es importante destacar que los avicultores han registrado datos para monitorizar la salud y la producción de los lotes de producción durante más de 40 años. Datos como el consumo de alimentos y agua, el crecimiento y la mortalidad han sido recogidos con la finalidad de controlar los rendimientos y, a partir de esto, mejorarlos (Hepworth et al., 2012).

Para el análisis de los datos de producción en sistemas agropecuarios, han sido ampliamente utilizados métodos matemáticos (Frost et al., 1997), técnicas estadísticas (Mench et al., 1986; Narinc, Uckardes, y Aslan, 2014) y técnicas de visualización de datos (Mertens et al., 2009). Estos métodos permiten identificar diferencias significativas en los indicadores productivos, como indicio de presencia de anomalías (De Vries y Reneau, 2010).

El concepto de los gráficos estadísticos de control, se utiliza comúnmente para denotar los sistemas de control de procesos industriales. Varios autores demostraron su uso en el contexto de la cría de animales, aunque las propiedades estadísticas de los datos relativos a los animales a menudo no cumplen con los principios básicos de estas técnicas (Mertens et al., 2009)

Sin embargo, debido a la tendencia a manejar cada vez poblaciones de animales más grandes, los sistemas automáticos de monitorización en la gestión de la producción

ganadera se convierten en necesarios como un complemento a la observación humana (Aydin, 2016).

La curva de la producción de huevos en una explotación avícola, puede ser afectada por varios factores tales como la ingesta de alimentos (calidad y cantidad), el consumo de agua, la intensidad y la duración de la luz recibida, la infestación por parásitos, enfermedades, causas ambientales, entre otras (Jacob, Wilson, Miles, Butcher, y Mather, 2014). En estudios como los realizados por Grossman et al., (2001) y Narinc et al., (2014) han desarrollado modelos matemáticos para describir la curva de producción y la persistencia de la curva en gallinas ponedoras. Otros trabajos, como los publicados por Long y Wilcox, (2011), estudiaron la curva de producción de gallinas ponedoras para determinar el aprovechamiento óptimo de los lotes de gallinas ponedoras.

Basándose en la recopilación masiva de datos que son convertidos en información relevante, Lokhorst y Lamaker (1996) diseñaron un sistema experto aplicado en la producción de gallinas ponedoras, posteriormente Lokhorst et al., (1998) introdujeron los primeros prototipos de sistema de monitoreo de la producción en ponedoras. Más recientemente, Xiao et al. (2011), desarrollaron un software que elabora análisis estadísticos de los datos de producción en gallinas ponedoras para detectar problemas en la curva producción.

La monitorización en tiempo real es un gran desafío porque la recolección de datos incluye una variabilidad natural dependiente de la hora de recogida, del estado nutricional de las aves, del fotoperiodo, entre otros factores. Woudenberg et al., (2014) desarrolló un método para la detección temprana de problemas avícolas basados en el cálculo de los residuos, lo que permite la identificación de problemas potenciales en la producción de huevos de 10 lotes de producción.

Actualmente, está recibiendo mayor interés la implementación de alertas tempranas en sistemas de gestión de la producción de animales, no obstante para lograr su aplicación se requiere recopilar información suficiente para establecer reglas de inferencia, que constituyan en su conjunto la base de conocimiento del modelo de alerta temprana (Leiyua, Ruizhia, y Zhenlib, 2012). En este sentido, es necesario considerar que las anomalías en la curva de producción son valores inesperados, por lo que identificarlos de manera automática es todo un reto, dada la variabilidad natural y las fluctuaciones aleatorias a las que están expuestos este tipo de datos (Woudenberg et al., 2014).

Schaefer et al., (2004) y Cameron, (2012) coinciden en que la detección temprana de un problema en los animales, permite actuar de manera oportuna. Esto minimiza los daños y reduce el potencial de contagio en caso de enfermedades, lo que repercute finalmente en la reducción de los costos e incremento de la eficacia de un tratamiento. En este sentido, Gates et al., (2015) proponen el desarrollo de sistemas innovadores para recopilar, analizar y comunicar la información en tiempo real, de tal forma que se pueda actuar a tiempo en la solución de los problemas en la producción ganadera comercial.

En avicultura, los algoritmos de ML han demostrado que son capaces de reemplazar con éxito a los modelos matemáticos y estadísticos tradicionales para modelar la curva de producción de huevos. Esto se debe a que son fáciles de utilizar, requieren menos

variables y resultan más eficientes (Ahmad, 2011; Ahmadi y Golian, 2008; Felipe, Silva, Valente, y Rosa, 2015). Sin embargo, existen pocos estudios previos sobre el uso de algoritmos de aprendizaje máquina para la detección temprana de problemas en la curva de la producción de huevos de gallinas comerciales.

2.2.2 Campo de aplicación en la industria de la caña de azúcar

El flujo de producción en la industria azucarera tiene varios procesos y subprocesos que requieren ser analizados para mantener un estándar de calidad (Polanco et al., 2014). Las plantas agroindustriales, requieren de sistemas costo-eficientes y no destructivos para controlar la calidad de sus procesos de producción, la seguridad alimentaria y el cumplimiento de las especificaciones técnicas (Kumaravelu y Gopal, 2015).

Uno de esos sistemas no destructivos para asegurar la calidad es la quimiometría, la cual se desarrolló desde la década de 1970 como un campo de estudio interdisciplinar. Este campo abarca una amplia y variada gama de técnicas matemáticas y estadísticas para el análisis de la composición química de los materiales (Kowalski, 1980).

Para analizar la calidad de materias primas orgánicas, una técnica comúnmente utilizada es la espectroscopía de Reflectancia del Infrarrojo Cercano (Near Infrared Reflectance - NIR spectroscopy), asociada a la quimiometría, no obstante la relación entre la absorción en la región espectral del infrarrojo cercano y el analito es frecuentemente de tipo no lineal (Bertran et al., 1999).

El origen de estas relaciones no lineales, es diverso y difícil de identificar, en algunos casos se debe a las diferencias en viscosidad, temperatura, pH, tamaño de partícula y a la propia naturaleza química del analito, por este motivo es común que la calibración se realice utilizando métodos de análisis multivariado (H. Martens y Naes, 1992). Una selección correcta de las variables a fin de reunir un pequeño subgrupo con una menor sensibilidad a las no linealidades o para descartar aquellas longitudes de onda más pronunciadas suele ser efectiva para mejorar el desempeño de los modelos (Leardi, Boggia, y Terrile, 1992; Saeys, Inza, y Larrañaga, 2007).

Recientemente, con el desarrollo de la informática y la quimiometría, las aplicaciones de la espectroscopía NIR se han popularizado y atraen cada vez más la atención de los investigadores, se plantea que la técnica es capaz de detectar concentraciones de analito de 0,1% (Cen y He, 2007). Actualmente se encuentra en constante expansión literatura que aplica técnicas de ML en quimiometría (Brereton, 2015; Mitchell, 2014; Tomazzoli et al., 2015; Torrione, Collins, y Morton, 2014), entre las que se destaca el uso de *Artificial Neural Networks* y *Support Vector Machines*, estas técnicas se basan en el reconocimiento de patrones (Brereton, 2015).

En la industria de los alimentos, se ha empleado ampliamente la espectroscopía NIR para analizar la calidad nutricional de lácteos (Tajammal Munir, Yu, Young, y Wilson, 2015; L. Wang, Sun, Pu, y Cheng, 2016), aceites (L. Wang et al., 2016), cárnicos (Zamora-Rojas, Pérez-Marín, De Pedro-Sanz, Guerrero-Ginel, y Garrido-Varo, 2012), peces (He, Wu, y Sun, 2014), cereales (Henry y Kettlewell, 2012) y frutas (Magwaza et al., 2011).

En la industria de la caña de azúcar se encuentran varios trabajos en los que se evidencia una buena correlación entre los espectros NIR y los indicadores de calidad de la caña de azúcar (Valderrama, Braga, y Poppi, 2007). Se ha investigado sobre las técnicas de pre-procesamiento de los espectros NIR en caña de azúcar (Zayas-Ruiz, Lorenzo-Izquierdo, y Fragoso-Concepción, 2015) la selección de características (Sorol, Arancibia, Bortolato, y Olivieri, 2010) así como algoritmos de quimiometría para mejorar la predicción de los analitos objetivos (Chen, Wen, Chen, Li, y Huo, 2014; X. Wang et al., 2010).

En el artículo de Tange et al. (2015) se demostró que el uso de modelos de calibración con Máquinas de Soporte Vectorial (SVM) para regresión resulta eficiente para estimar los valores de ºBrix y Sacarosa, parámetros de calidad del proceso industrial del azúcar. El uso de SVM mejora en cuanto al error de predicción a la técnica de *Partial Least Squares* (PLS), sin embargo, el modelo propuesto utiliza todo el espectro NIR, lo que induce a pensar que aún es posible optimizar el modelo, implementando una técnica de pre-procesamiento adecuada, una correcta selección de características y la optimización de los parámetros de la máquina de soporte de vectores.

2.2.2.1 Espectroscopía NIR

La espectroscopía es el estudio de la interacción de la radiación electromagnética con una sustancia química, la naturaleza de esta interacción depende de las propiedades de la sustancia. Cuando la radiación pasa a través de una muestra (sólido, líquido o gas), ciertas frecuencias de las radiaciones son absorbidas por las moléculas de la sustancia que conduce a las vibraciones moleculares. Las frecuencias de radiación absorbidas son únicas para cada molécula, lo que proporciona un patrón o huella que caracteriza a la sustancia (Ozaki, McClure, y Christy, 2006).

El sector agroalimentario fue el primero en adoptar la espectroscopía NIR como técnica analítica de la calidad. En 1974, la *Canadian Grain Commission* reemplazó el tradicional método de *Kjeldahl* para la determinación de proteína en trigo con un método, basado en NIR. En 1995, se estimó que el nuevo método ahorraba $ 2,5 millones y evitaba la generación de 47 toneladas de residuos cáusticos cada año (Stark, 1996)

Desde entonces, la espectroscopía NIR ha sido ampliamente y efectivamente utilizada para evaluar la calidad de alimentos (Rady y Guyer, 2015). Esto se debe a que es una técnica de análisis rápido potente, mínimamente invasiva y no destructiva. Sin embargo, la única manera de extraer la información física y química relevante de los datos espectrales de la muestra es mediante algoritmos de aproximación estadística llamados quimiométricos (Kumaravelu y Gopal, 2015).

Los espectros NIR se obtienen como resultado de las transiciones vibratorias fundamentalmente asociadas con los enlaces químicos que contienen hidrógeno, C-H, N-H, S-H y O-H, y que están presentes en la mayoría de compuestos orgánicos e inorgánicos. La información de dicho espectro NIR se extiende a lo largo del rango de longitud de onda entre 780 nm - 2500 nm. (Dos Santos, Lopo, Páscoa, y Lopes, 2013; Kumaravelu y Gopal, 2015; Teye, Huang, y Afoakwa, 2013).

Una clara desventaja de la espectroscopía NIR es que no existen modelos preestablecidos para medir la interacción entre la reflectancia de la luz en el espectro del infrarrojo cercano y la materia, por lo que se requiere de calibración. Este proceso es puramente empírico en la mayoría de los casos (Kumaravelu y Gopal, 2015), además para obtener modelos de calibración robustos y precisos se requiere de un número suficientemente grande de muestras que abarque las variaciones en las propiedades físicas o químicas a modelar por ejemplo muestras tomadas en distintos lugares, climas y sistemas de producción (Florkowski, Prussia, Shewfelt, y Brueckner, 2009).

El espectro NIR, en todas sus longitudes de onda, puede contener información respecto a varios analitos (Blanco y Villarroya, 2002). Esto implica que el espectro resultante es producto de las modificaciones que se realizan simultáneamente en todos los analitos de la muestra, haciendo más complicado el proceso de calibración (Florkowski et al., 2009).

Además de las características químicas, características físicas como temperatura, viscosidad, cristales y pH pueden alterar el espectro NIR. Por añadidura, otros factores que aumentan la complejidad de los espectros son las bandas de absorción de agua, los efectos de dispersión, el ruido propio del instrumento de medición y otros efectos ambientales. Esto hace que se requiera el uso de técnicas estadísticas multivariantes junto a técnicas de pre-procesamiento para obtener modelos de calibración robustos (Dos Santos et al., 2013; Florkowski et al., 2009; H. Martens y Naes, 1992).

La quimiometría es una parte esencial de la espectroscopía NIR en el sector de la alimentación (Dos Santos et al., 2013), las técnicas quimiométricas utilizan métodos de análisis multivariado. Uno de estos métodos es el análisis de componentes principales (Principal Component Analysis – PCA), la cual es una técnica de análisis cualitativo de los datos espectrales. También se utiliza el análisis de regresión por cuadrados mínimos parciales (Partial Least Squares – PLS) como una técnica para obtener predicción cuantitativa de los parámetros de interés de la muestra (Kumaravelu y Gopal, 2015).

En la literatura científica, se reporta también el uso de regresión lineal, regresión multivariada, redes de neuronas artificiales, SVR, entre otros. (Blanco y Villarroya, 2002; H. Martens y Naes, 1992; Tange et al., 2015)

Debido a que la espectroscopía NIR es una técnica de análisis relativo, es decir que, para realizar la calibración, se requiere relacionar el espectro con el valor del analito, el mismo que debe ser determinado previamente mediante un método de referencia. El propósito de los métodos quimiométricos de análisis multivariado es construir modelos capaces de estimar con precisión las características y propiedades de nuevas muestras desconocidas (Blanco y Villarroya, 2002).

Capítulo:
III. MATERIALES Y MÉTODOS

Este capítulo describe los materiales, métodos y metodologías que se utilizaron durante el desarrollo de la investigación a fin de cumplir con los objetivos propuestos en el Capítulo I. Primero se describe las bases de datos y en la medida que se va presentando la metodología propuesta para el abordaje de cada problema, se describen los métodos y técnicas que se van a emplear.

El presente trabajo de Tesis Doctoral se realizó utilizando dos bases de datos de ámbito agropecuario, una aplicada a la industria avícola y la otra aplicada a la industria de la caña de azúcar. La metodología para el análisis de cada base de datos debe ser la adecuada, por consiguiente, en esta sección se realiza una descripción detallada de los materiales y los métodos empleados para cada aplicación por separado.

Para el procesamiento de los datos, se realizarán pruebas utilizando varias técnicas de aprendizaje máquina y minería de datos como: Redes de Neuronas Artificiales para clasificación y Máquinas de Soporte Vectorial utilizadas tanto para clasificación como para regresión.

3.1 Base de datos aplicada a la industria avícola

La empresa Agrolomas Cía. Ltda. con sede en la provincia de El Oro, Ecuador facilitó sus datos productivos del periodo comprendido entre enero del 2008 a diciembre del 2015. La base de datos contiene registros diarios de la cantidad de huevos sanos, huevos trizados y número de aves que mueren.

Los datos corresponden a 24 lotes con aproximadamente 20.000 aves cada uno. Las líneas genéticas utilizadas fueron ISA Brown, Lohmann Brown y HyN. El sistema de reemplazo utilizado fue “todo dentro - todo fuera” (Flanders y Gillespie, 2015), es decir que cada lote contiene únicamente aves de la misma edad durante todo el período de producción. El período de producción utilizado para los experimentos fue de 60 semanas (desde la edad de 19 a 79 semanas).

Se conformó un panel de expertos para el etiquetado de los datos. Cuando sucedía una caída de producción, el panel de expertos etiquetó ese día como positivo. Los días normales fueron etiquetados como negativos.

El número medio de días etiquetados como positivo para cada lote fue de 8 días, sin embargo, se observa que hay lotes, que no presentan ningún problema, mientras que otros, que presentan hasta 33 días.

En total, los 24 lotes presentaron 188 etiquetas positivas. Esto representa el 1,85% de los 10.142 registros. Por este motivo, el algoritmo tiene una tarea de aprender las relaciones entre los datos a partir de una gran cantidad de patrones negativos (días en que no hay problemas) y unos pocos patrones positivos (días que presentan algún problema). Esta situación desequilibra los resultados esperados y añade dificultad a la tarea de clasificación y predicción.

La Tabla 3.1 describe cada lote con sus indicadores generales correspondientes: tiempo de producción; aves alojadas en el inicio del ciclo de producción; aves muertas durante el tiempo de producción; total de huevos producidos por el lote durante el tiempo de producción; promedio de huevos producidos por día; huevos por gallina alojada al día; la máxima producción% (pico) que alcanzó y el número de etiquetas positivas de cada lote.

Tabla 3.1. Indicadores de producción de los lotes estudiados.

Lote	tiempo (d) producción	aves alojadas	aves muertas	huevos totales	Huevos/ día	huevos por ave alojada	pico de producción	etiquetas positivas
1	473	20,300	2,929	7,211,252	15,246	0.7510	98.14%	30
2	429	20,361	3,022	6,951,132	16,203	0.7958	97.04%	0
3	516	20,137	3,630	8,160,013	15,814	0.7853	96.87%	1
4	148	18,874	1,430	1,767,577	11,943	0.6328	N/A	0
5	480	19,770	2,421	7,185,831	14,970	0.7572	97.25%	0
6	461	20,408	1,573	7,145,492	15,500	0.7595	97.11%	33
7	518	20,187	2,718	7,974,633	15,395	0.7626	95.97%	14
8	501	20,130	1,984	8,093,083	16,154	0.8025	97.03%	0
9	104	19,740	436	1,594,527	15,332	0.7767	97.03%	0
10	389	19,668	2,153	6,078,320	15,626	0.7945	95.86%	0
11	543	19,920	2,409	7,900,793	14,550	0.7304	97.32%	0
12	491	19,934	1,969	7,230,558	14,726	0.7387	98.70%	17
13	431	19,492	1,382	6,787,937	15,749	0.8080	96.30%	13
14	419	19,920	1,600	7,147,832	17,059	0.8564	97.17%	0
15	468	20,120	1,549	7,172,119	15,325	0.7617	98.91%	0
16	517	20,234	2,865	7,692,698	14,879	0.7354	97.42%	12
17	498	19,971	2,051	7,744,766	15,552	0.7787	96.83%	24
18	391	20,104	1,238	6,463,994	16,532	0.8223	97.72%	13
19	307	20,094	693	5,301,566	17,269	0.8594	98.52%	0
20	450	19,895	1,984	6,905,452	15,345	0.7713	98.16%	13
21	480	19,910	2,702	7,590,784	15,814	0.7943	96.94%	0
22	529	19,950	2,973	8,429,271	15,934	0.7987	98.20%	10
23	374	19,907	2,050	6,023,519	16,106	0.8090	98.30%	8
24	202	19,893	814	3,407,626	16,869	0.8480	97.63%	0

La Figura 3.1. muestra un diagrama de cajas y bigotes, en el que son presentadas las medidas de tendencia y de dispersión de la variable huevos por gallina al día, para los 24 lotes empleados en el desarrollo del modelo. Se observa que cada lote presenta características distintas en cuanto a su distribución de datos, esta heterogeneidad en los datos es favorable para el proceso de entrenamiento ya que el conjunto de datos tendrá una casuística variada y por lo tanto se espera que el modelo resultante tenga mayor representatividad.

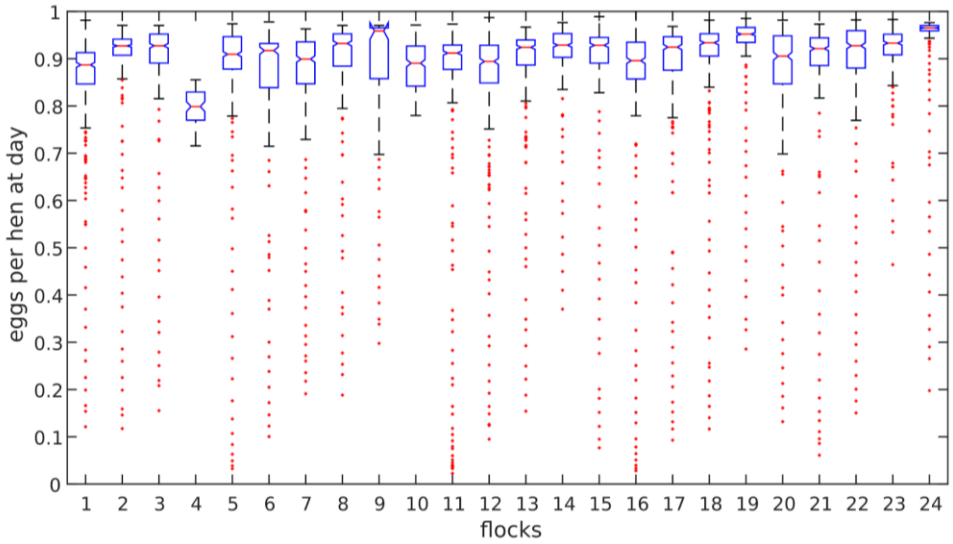


Figura 3.1. Diagrama de cajas y bigotes de huevos por gallina al día en los 24 lotes empleados para el desarrollo de los modelos.

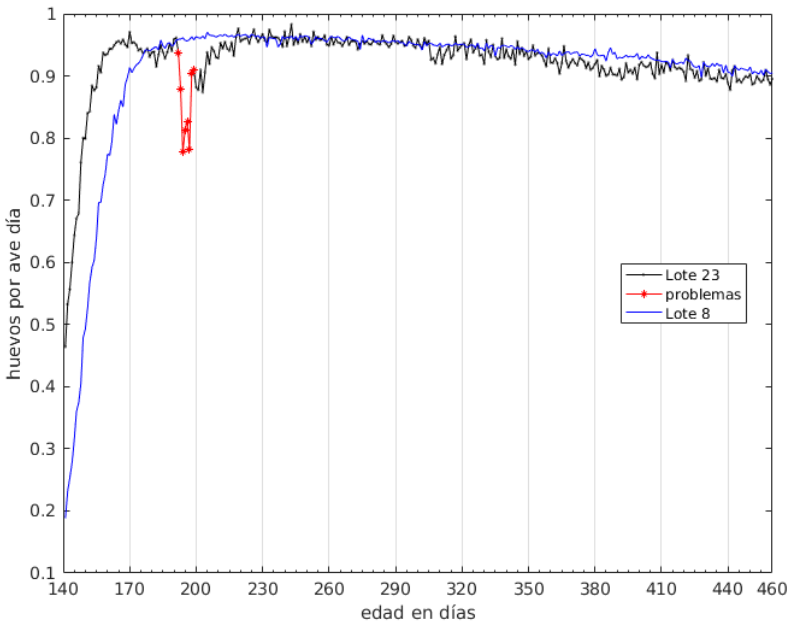


Figura 3.2. Producción diaria en los lotes representativos 8 y 23.

La Figura 3.2 muestra dos lotes representativos de la base de datos. El lote 8 tiene una curva de producción característica, sin anomalías en todo el tiempo de producción. Por

su parte, el lote 23, que a pesar de que inicia su producción con menos edad, se aprecia una fuerte caída entre los 191 días y los 199 días, este intervalo de tiempo fue etiquetado como problemas en la curva, ya que a partir del día 199 las aves ya han empezado a recuperarse.

Se ha encontrado que en las granjas avícolas en las que la recolección de huevos se realiza de manera constante a la misma hora. Al compararlos con los datos de las granjas en los que no es posible establecer una hora fija para la recolección y conteo de los huevos, los datos registrados tienen mayor consistencia.

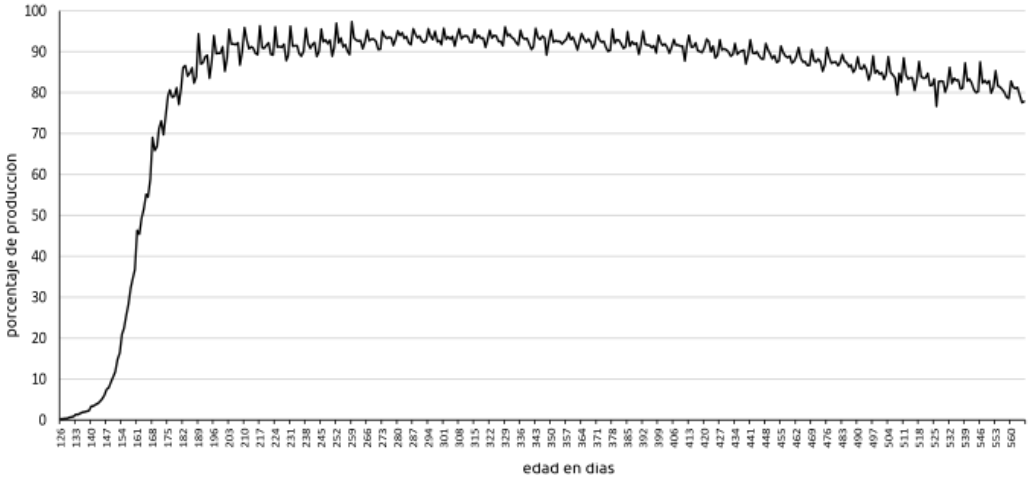


Figura 3.3 registros diarios de producción de huevos por ave en el Lote 11.

En la base de datos experimental, los datos fueron registrados una vez al final del día, aunque no fue siempre a la misma hora. La propia dinámica de la producción hace que algunos días se registre antes de las 14h00 y otros luego de las 16h00. Esta variabilidad puede observarse en la Figura 3.3, que representa la producción diaria de huevos por ave. Este hecho representa un desafío adicional para el modelo de alerta temprana, ya que debe ser capaz de distinguir entre un problema real y estas caídas que se producen debido a las variaciones cíclicas semanales relacionadas con la rutina y el momento de la recolección.

La Figura 3.4 muestra un ejemplo de zonas etiquetadas como problema de acuerdo con el criterio del panel de expertos. Los días en que hubo una caída de producción fueron etiquetados como 1, y los días con producción normal fueron etiquetados como 0.

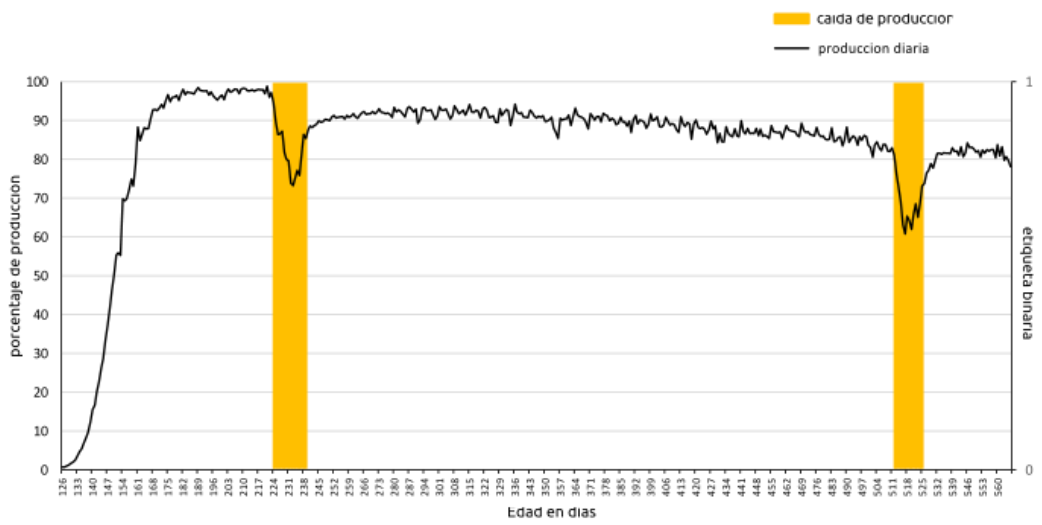


Figura 3.4. Ejemplo de etiquetado de los problemas en Lote 1.

3.1.1 Partición de los datos

Para este trabajo fue utilizada una técnica de validación cruzada de 5 capas. De esta manera, los patrones se dividieron en 5 subconjuntos. La división se realizó de manera aleatoria y estratificada, de tal manera que cada subconjunto contenía proporciones similares de patrones positivos y negativos.

Para cada entrenamiento, un subconjunto fue usado como test y los otros 4 para el entrenamiento. La validación cruzada fue repetida 100 veces para obtener la información estadística del desempeño de los modelos que permitan su evaluación adecuada.

Durante el proceso de validación cruzada *k* veces, los datos se dividen en *k* subconjuntos; uno se utiliza como un subconjunto de prueba y los otros (*k*-1) como subconjuntos de formación (Mucherino et al., 2009). El proceso de validación cruzada se repite para *k* pliegues, con cada uno de los posibles subconjuntos, y finalmente se lleva a cabo una media aritmética de los resultados para cada pliegue para obtener un único resultado, que se pasa a la SVM.

3.1.2 Conformación de los patrones de entrada

En datos de series temporales, es posible reestructurar los patrones de entrada para organizarlos previo al entrenamiento de algoritmos de aprendizaje supervisado. Esto se hace, mediante la utilización de los datos previos como variables de entrada y utilizar un dato del siguiente día como la variable de salida (Kapoor y Bedi, 2013).

Este método se conoce como ventana temporal deslizante o *sliding time window*, consiste en la creación de diferentes secuencias de puntos de datos consecutivos de la serie temporal. Existen dos parámetros en este método: el tamaño de la ventana y el tamaño de paso en la ventana. El parámetro más importante es el tamaño de la ventana, normalmente se experimenta con distintos valores hasta encontrar el óptimo, mientras

que el tamaño de paso de la ventana se mantiene típicamente igual a uno (Saeed y Snášel, 2014).

A partir de los datos de producción, se crearon los patrones de entradas a los algoritmos y sus salidas deseadas. Los patrones de entrada se componen tomando datos de una ventana deslizante (Lindsay y Cox, 2005), con una muestra del día actual y algunas muestras anteriores y consecutivos, de acuerdo con el tamaño de las ventanas.

3.1.3 Optimización de los algoritmos de ML

El enfoque propuesto para el modelado del problema está basado en la combinación de un clasificador que utiliza los datos de una ventana temporal como patrones de entrada. A partir de estos datos, el algoritmo debe ser capaz de detectar las anomalías. Se consideró el enfoque del estudio de Bennett y Campbell (2000) quienes plantearon que los algoritmos de clasificación pueden ser usados para detectar anomalías. En ambos casos, el intervalo de pronóstico se fijó entre cero y cinco días.

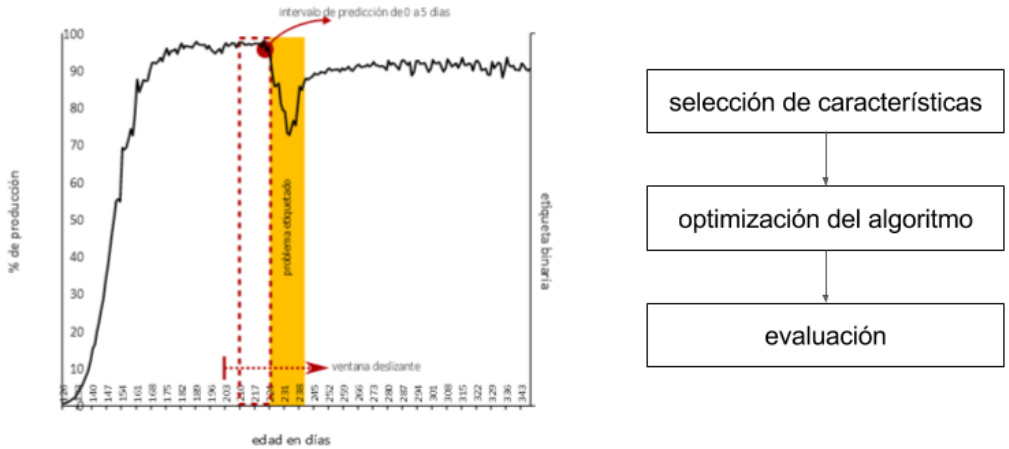


Figura 3.5 Enfoque propuesto para el modelado del problema.

En la Figura 3.5 se observa un esquema del enfoque de Bennett y Campbell (2000), en una ventana temporal deslizante. De esta manera, se buscar generar modelos de alerta temprana basados en IA, para detectar problemas en las curvas de producción. Esto permitiría anticiparse y actuar oportunamente en la solución de los mencionados problemas.

En el caso del modelo con SVM, el proceso de entrenamiento consistió en: la selección de características por parte del experto, para, a continuación, optimizar el algoritmo, es decir la selección del *kernel* y optimizar el parámetro C. Finalmente los modelos fueron evaluados a diferentes tamaños de ventana, e intervalos de predicción.

En el caso del modelo con ANN, la selección de características se realizó de manera automática a partir de los datos en crudo, la selección de las características se realizó automáticamente. Dado que el tamaño de ventana y el umbral del filtro univariado afectan directamente a la selección de características, la optimización se realizó de manera simultánea. El siguiente paso fue seleccionar la mejor arquitectura de la ANN y finalmente se ajustó el parámetro de ponderación (S), que mide la importancia que tienen los patrones positivos en el modelo.

3.1.4 Análisis de desempeño de los algoritmos

El desempeño de un algoritmo de clasificación, normalmente se analiza mediante el cálculo de la fracción de instancias del conjunto de prueba que han sido correctamente clasificadas. Esta medida recibe el nombre de exactitud o *Accuracy* (ACC). Generalmente la ACC es el único requisito de rendimiento utilizado para evaluar el desempeño de técnicas de aprendizaje máquina (D. Martens y Baesens, 2010; Venkatesan, Thangavelu, y Prabhavathy, 2013). Sin embargo, de acuerdo con Sun et al. (2009) cuando se trata de analizar el desempeño en bases de datos no balanceadas, puede suceder que una ACC alta se produzca sólo por el hecho de clasificar todas las instancias como miembros de una clase. Por este motivo, se requiere recurrir a otras medidas como la especificidad o *Specificity* (SPC), sensibilidad o *Sensitivity* (SEN) y el valor predictivo positivo o *Positive Predictive Value* (PPV). Las fórmulas de cálculo de estas medidas de desempeño han sido descritas por Fawcett (2006) de la siguiente manera:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.1) \qquad SPC = \frac{TN}{TN+FP} \quad (3.3)$$

$$SEN = \frac{TP}{TP+FN} \quad (3.2) \qquad PPV = \frac{TP}{TP+FP} \quad (3.4)$$

Donde *TP* son los valores positivos que son detectados correctamente, *FP* son los valores negativos que son detectados de manera incorrecta como positivos, *TN* son los valores negativos que son detectados de manera correcta y, finalmente, *FN* son los valores positivos que son detectados incorrectamente como negativos.

La ACC es la capacidad general del sistema para detectar correctamente, los valores normales y las anomalías. La SPC se define como la capacidad de detectar la ausencia de anomalías como falsa. La SEN es la capacidad de detectar la presencia de una anomalía como verdadera. El PPC es la probabilidad de que una anomalía en realidad se produce cuando la prueba es positiva (Altman y Bland, 1994; Hastie et al., 2009; Pepe, 2003).

Para tomar la decisión sobre cuáles parámetros resultaron óptimos en los modelos finales, se realizó Análisis de Varianza (ANOVA) y Análisis de Rango Múltiple (MRT) con el método de *Tukey's Honestly Significant Difference* (HSD) para un valor de $p < 0{,}01$.

Estas medidas de desempeño fueron evaluadas en las tres etapas de optimización de los modelos para un intervalo de predicción igual a uno. Al final, cuando se obtuvo el modelo optimizado, se realizó pruebas a distintos intervalos de predicción (de cero a cinco), para determinar hasta cuantos días de anticipación, se puede contar con modelos confiables para detectar una anomalía en la curva de producción.

3.2 Base de datos aplicada a la industria del procesamiento de la caña de azúcar

La base de datos fue publicada por Tange et al. (2015) en la dirección electrónica http://www.models.life.ku.dk/nirsugarcanedata, los datos fueron obtenidos en una fábrica de azúcar japonesa (Daito Togyo Co) en la que se procesa caña de azúcar. Las

muestras se obtuvieron durante tres meses en la temporada de cosecha, durante cada uno de los pasos del proceso: después de la molienda (jugo), seguido del proceso de evaporación (jarabe), para a continuación realizar la cristalización (masa cocida) y finalmente la centrifugación (melaza). Se realizaron tres ciclos de cristalización y centrifugación por lo que, las muestras de masa cocida y melaza son más numerosa que las de las otras dos etapas. Inmediatamente después del muestreo, se extrajo las señales NIR y la técnica de referencia a temperatura de proceso.

Un total de 1797 espectros NIR entre 400 nm a 1888 nm con incrementos de 2 nm, fueron obtenidos utilizando un espectrómetro NIR DS2500, FOSS AB. En el trabajo presentado por Tange et al (2015), se excluyó del análisis las señales espectrales con una absorbancia mayor a dos, en contraposición, el presente trabajo utilizó la base de datos completa.

El parámetro de calidad ºBrix, expresa la totalidad de los sólidos disueltos (azúcar y no azúcar) como porcentaje del peso total, donde su escala coincide con el porcentaje de sacarosa en soluciones puras. En todo material azucarero (jugo, mieles, etc) los ° Brix, son siempre mayores a los de Sacarosa, mientras que en materiales de elevada pureza como los licores de una refinería la diferencia entre estos indicadores es mínima. En el presente trabajo, el parámetro de calidad ºBrix, se midió utilizando un refractómetro Abbemat-WR, Anton Paar GmbH.

El parámetro de calidad Sacarosa (Pol), expresa la cantidad de sacarosa contenida en una disolución, expresada como porcentaje del peso. En soluciones puras el porcentaje de Pol equivale exactamente al porcentaje de sacarosa, mientras que, en otras impuras como el jugo de caña y las mieles, existe una diferencia entre estos dos valores, diferencia que será mayor, cuanto más impura sea la disolución, por esta razón el valor de Pol es aceptado internacionalmente como sacarosa aparente. En el presente trabajo se midió la Sacarosa utilizando un polarímetro MCP500, Anton Paar GmbH.

En el caso de los datos de °Brix, se observa una marcada separación en las distribuciones de frecuencias, con una ligera superposición entre masa cocida y melaza en la franja alrededor de 90 °Brix. Por otra parte, las mediciones de sacarosa muestran zonas de entrecruzamiento entre jarabe, masa cocida y melaza, en la franja alrededor de 50%, de sacarosa, como se puede observar en la Figura 3.6.

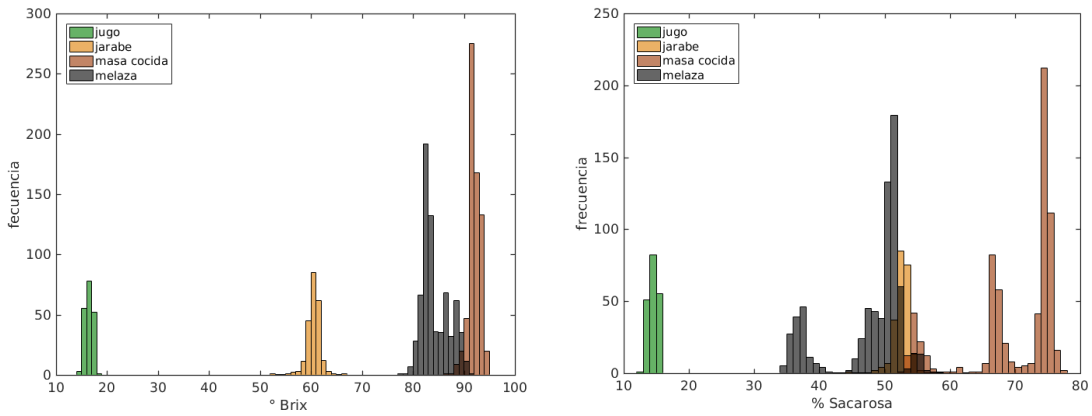


Figura 3.6. Histogramas del contenido de °Brix y % Sacarosa, en cada subproceso.

3.2.1 Partición de los datos

Los datos espectrales fueron divididos en subconjuntos de entrenamiento (calibración) y test utilizando una técnica de validación cruzada (Garcia y Filzmoser, 2015; Kuhn y Johnson, 2013) se escogió una técnica de 10 capas. En esta, los datos se dividen en 10 grupos, 9 se utilizan como conjunto de calibración y el uno restante como conjunto de test. Se cambia el test set hasta que todos los grupos hayan sido puestos a prueba. La validación cruzada se repitió 100 veces.

3.2.2 Técnicas de selección de características

Debido a la gran cantidad de información espectral proporcionada por los espectrofotómetros NIR, se requiere reducir sustancialmente el número de muestras necesarias para la construcción de modelos de clasificación y de calibración. (Blanco y Villarroya, 2002). En la última década, la selección de características (Feature Selection - FS) en la construcción de modelos ha pasado de ejemplos ilustrativos en cuanto su funcionamiento, a convertirse en un requisito, particularmente por la naturaleza de los problemas con alta dimensionalidad como los *microarray analysis* y los análisis espectrales (Saeys et al., 2007).

La selección de características contrasta con otras técnicas de reducción de dimensionalidad, como el PCA. Este contraste se debe a que la selección de variables no altera la representación original de las variables, sino que consiste simplemente seleccionar un subconjunto de las mejores características, conservando su naturaleza original. Esto favorece que sean fácilmente interpretables por un experto en campo (Saeys et al., 2007), lo que los convierte en datos relevantes y deseables.

Considerando que muchas de las técnicas de reconocimiento de patrones no fueron diseñadas en sus orígenes para lidiar con grandes cantidades de información poco relevante, la aplicación de técnicas de FS se ha convertido en una necesidad actualmente en muchas aplicaciones (Isabelle Guyon y Elisseeff, 2003)

Aplicar previamente una técnica de FS, previene el sobreajuste del modelo, mejora el desempeño y reduce el tiempo de cómputo. Además, permite obtener una comprensión más profunda de los datos. Esto agrega complejidad en la tarea de modelado, ya que en lugar de sólo optimizar los parámetros del modelo, ahora es necesario encontrar las características óptimas que definen al modelo, y en el caso de regresión se utiliza para buscar las variables que maximizan el ajuste del modelo (I. Guyon, Gunn, Nikravesh, y Zadeh, 2008; Keleş, van der Laan, y Eisen, 2002; Saeys et al., 2007).

Un grupo de técnicas comúnmente utilizadas para FS, son los filtros que evalúan la relevancia de una característica. Estos analizan las propiedades intrínsecas de los datos y, en general calculan un valor de relevancia en el que aquellas características con menor valor que el umbral predefinido son eliminadas (Saeys et al., 2007; Szymańska et al., 2015). En general estos filtros son de dos tipos, los univariados y los multivariados (Isabelle Guyon y Elisseeff, 2006).

El filtro univariado es un paradigma simple pero eficiente, ya que el *ranking* de salida es fácil de entender. Con ellos se suele definir un método de umbral para seleccionar aquellos que cumplan con una condición por encima o por debajo del mismo. Los filtros

trabajan de forma independiente del modelo y utilizar las propiedades intrínsecas de los datos (Szymańska et al., 2015).

Los filtros univariados tienen la ventaja de que son rápidos, escalables e independientes de la técnica de clasificación/regresión que se utilice. Sin embargo, estos filtros tienen como principal desventaja que ignoran la dependencia y correlación entre características (H. Liu y Motoda, 2012).

El método de filtro univariado se puede aplicar utilizando un *t*-test (Jafari y Azuaje, 2006), F-test (Bhanot, Alexe, Venkataraghavan, y Levine, 2006) o la prueba de suma de rangos de Wilcoxon (Szymańska et al., 2015) con lo que se calcula un *p-value* que representa la significación estadística de cada variable en el modelo, así las variables son ordenadas en dependencia de su *p-value* (Saeys et al., 2007).

3.2.3 Optimización del algoritmo de regresión

De manera general, el procesamiento de las muestras consta de los siguientes pasos: adquisición de datos espectrales, pre-procesamiento de los datos para eliminar ruido, selección de características, desarrollo del modelo de calibración utilizando un conjunto de espectros de los que se conoce los valores de los analitos objetivo, obtenidos por técnicas de referencia y, finalmente la validación del modelo utilizando datos distintos a los de calibración (Cen y He, 2007).

En el presente trabajo se presenta una metodología para optimizar modelos de calibración NIR para múltiples procesos en la industria azucarera. En este caso particular se utilizó un algoritmo SVR con un *kernel* RBF (Gunn, 1998), la optimización de los parámetros se realizó en una secuencia de pasos que se describen a continuación:

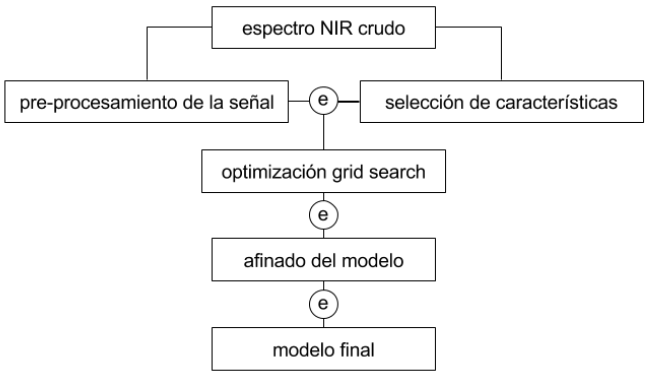


Figura 3.7 Metodología para la optimización de modelos de calibración NIR utilizada en el presente trabajo.

Como se aprecia en la Figura 3.7, paralelamente se realizan la elección de la técnica de pre-procesamiento más adecuada y la selección de características que modelan de una mejor manera las no linealidades del espectro. Para esto el umbral del percentil de *p-value* usado en el proceso de FS fue evaluado para las nueve técnicas de pre-procesamiento, con valores fijos de los parámetros C igual a 100, γ igual a 0.1 y ε igual a 0.1. En este punto se evalúa el modelo y sus valores óptimos son fijados, para continuar con la optimización.

La combinación óptima de los parámetros C y γ, fue optimizada de manera simultánea usando el método de *grid search* (Koch et al., 2012; Ma, Zhang, y Wang, 2015) en escala logarítmica, el parámetro C fue evaluado en el rango de 10^1 a 10^6 y γ en el rango de 10^{-1} a 10^{-3}, el intervalo de evaluación fue de 0.25 en el exponente, con un valor fijo de ε igual a 0.1.

La técnica *grid search* es un enfoque de fuerza bruta para optimización de parámetros que se utiliza ampliamente en técnicas de ML. Una vez que se definen los rangos y las medidas de parámetros, cada combinación de los parámetros se pone a prueba para encontrar la mejor en base a una medida de desempeño (Ma et al., 2015).

Finalmente, con el objetivo de afinar el desempeño de la SVR, fue optimizado el parámetro ε de la SVR, en el rango de 0 a 1 en escala lineal, el intervalo de evaluación fue de 0.01.

En cada fase se realizó 100 repeticiones para cada CV. En total 478000 SVR fueron evaluadas para determinar la configuración óptima del modelo de calibración.

3.2.4 Técnicas de pre-procesamiento

El pre-procesamiento de los datos espectrales NIR es una parte integral del modelado en quimiometría. El objetivo del pre-procesamiento es eliminar los fenómenos físicos en los espectros con el fin de mejorar la regresión multivariante posterior, modelo de clasificación o análisis exploratorio (Blanco y Villarroya, 2002; Rinnan, Berg, y Engelsen, 2009).

Es común utilizar técnicas para reducir el ruido en los espectros, instrumental o información de fondo por ejemplo la técnica de *detrending* suele utilizarse para eliminar los efectos de la acumulación los conjuntos de datos a partir de una tendencia (Dos Santos et al., 2013).

La espectroscopía del infrarrojo cercano es la técnica espectroscópica que ha conllevado la mayor cantidad y diversidad de técnicas de pre-procesamiento, principalmente debido a que los espectros pueden ser influenciados de manera significativa por las no linealidades introducidas por la dispersión de la luz (Rinnan et al., 2009).

Los espectros NIR de muestras sólidas se ven influidos por las propiedades físicas de las muestras. Esto plantea algunos problemas en la evaluación, por lo que resulta importante el pre-procesamiento para minimizar las contribuciones que incorporan información irrelevante con lo que se puede desarrollar modelos más simples y robustos (Blanco y Villarroya, 2002; Dos Santos et al., 2013).

Las técnicas de pre-procesamiento tienen el objetivo de reducir la variabilidad en los datos con el fin de mejorar y simplificar la función buscada en los espectros. Estas técnicas, son indispensables en análisis espectral multivariado, aunque siempre existe el peligro de aplicar el tipo incorrecto o la aplicación de un procesamiento previo demasiado severo que elimine información valiosa. Por esta razón, el pre-procesamiento de espectros NIR se realiza aún por prueba y error (Rinnan et al., 2009; Xu et al., 2008).

La elección adecuada de la técnica de pre-procesamiento es difícil de evaluar antes de la validación del modelo. En general, la realización de varios pasos pre-procesamiento no es aconsejable, y, como requisito mínimo el pre-procesamiento debe mantener o disminuir la complejidad del modelo eficaz (Rinnan et al., 2009).

En el presente trabajo se aplicaron cuatro técnicas básicas para las combinaciones de pre-procesamiento de los espectros NIR: la ley de Beer-Lambert, la primera derivada espectral, Standard Normal Variate, y *detrending*.

3.2.4.1 Ley de Beer-Lambert:

Es empírica para los espectros NIR y sugiere una relación lineal entre la absorbancia de los espectros y la concentración (s) de la constituyente. Esta ley es válida únicamente para sistemas de transmitancia pura sin dispersión. En mediciones de reflectancia la ley se puede expresar de la siguiente manera (Rinnan et al., 2009).

$$A\lambda = -log_{10}(R) \cong \epsilon\lambda \times l \times c \quad (3.5)$$

Donde $A\lambda$ es la absorbancia dependiente de la de longitud de onda, $\epsilon\lambda$ es la absortividad molar dependiente de la longitud de onda, l es la longitud efectiva de la trayectoria de la luz a través de la matriz de la muestra, y $c$es la concentración del componente de interés.

3.2.4.2 Derivadas espectrales:

Tienen la capacidad de eliminar tanto los efectos aditivos como los multiplicativos en los espectros y se han utilizado en la espectroscopía analítica durante décadas. La primera derivada suprime únicamente la línea de base, la segunda derivada elimina tanto la línea de base como la tendencia lineal (Dos Santos et al., 2013). El método más básico para la derivación es el de diferencias finitas (Rinnan et al., 2009), para el caso de la primera derivada, ésta se calcula como la diferencia entre dos puntos de medición espectrales posteriores:

$$X_{i,fsd} = X_i - X_{i-1} \quad (3.6)$$

3.2.4.3 Standard Normal Variate (SNV):

Es posiblemente el segundo método más aplicado para la corrección de dispersión de NIR / NIT de datos (Barnes, Dhanoa, y Lister, 1989). El formato básico para SNV y corrección normalización es el siguiente:

$$X_{i,snv} = \frac{X_i - \bar{x}}{S} \quad (3.7)$$

En donde $X_{i,\,snv}$ es el espectro SNV a la longitud de onda *i*, $\bar{x}$ es la media del espectro de la muestra a ser corregido, y *S* es la desviación estándar de la muestra de espectro.

3.2.4.4 Técnica *detrending*:

Se aplica sobre los espectros para eliminar los efectos del cambio de línea de base y curvilinealidad, es característico para los espectros NIR a los que se les ha aplicado la técnica de pre-procesamiento basada en la ley de Beer-Lambert. Este efecto es generalmente lineal (Luypaert, Heuerding, de Jong, y Massart, 2002). El método

consiste en modelar la línea de base como una función lineal de la longitud de onda y esta función a continuación, se resta de cada valor del espectro de forma independiente.

$$X'_i = X_i - b_i \quad (3.8)$$

Donde b_i es la línea de base del espectro en la longitud de onda *i*, calculada con el modelo lineal.

3.2.5 Análisis del desempeño

Todos los modelos resultantes de las diferentes combinaciones de parámetros fueron evaluados utilizando como medidas del desempeño el coeficiente de determinación (R^2) y la raíz del error cuadrático medio (RMSE). Para su cálculo se utilizaron los datos de test de la validación cruzada como indicadores de la exactitud del modelo basado en NIR (Dos Santos et al., 2013; Viscarra Rossel, 2008).

La ecuación para el coeficiente de determinación *R-squared* es:

$$R - squared = \frac{\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{i=n}(y_i - \bar{y})^2} \quad (3.9)$$

La ecuación para el error cuadrático medio (RMSE) es:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2}{n}} \quad (3.10)$$

Dondey_i es el valor real de la i-ésima observación, $\hat{y}_i$ es el valor predicho de la i-ésima observación, $\bar{y}$ es el valor medio de las mediciones reales y n es el número de observaciones.

R^2 es un valor estadístico que determina la calidad de un modelo. Esto lo hace en función de la proporción de variación entre los valores reales y valores predichos, que puede explicarse por el modelo. El RMSE, consiste en las diferencias entre los valores predichos por un modelo y los valores realmente observados. Esta técnica suele preferirse respecto a otras ya que su interpretación está en la misma escala que los datos, esta medida ha ganado popularidad debido a su relevancia teórica en el modelado estadístico (Armstrong y Collopy, 1992; Hyndman y Koehler, 2006).

Capítulo:
IV. RESULTADOS Y DISCUSIÓN

En este capítulo se presentan los principales resultados de esta investigación. Para ello, se ha realizado una división del capítulo en dos partes, una respecto a la aplicación de las técnicas de ML en la industria avícola de producción de huevos comerciales, y otra que se refiere a la aplicación de las técnicas de ML en la calibración del proceso de control de calidad en la industria de la caña de azúcar.

4.1 Aplicación de técnicas de ML en la industria avícola

En este trabajo se utilizó dos técnicas de ML orientadas a clasificación, con el objetivo de realizar una detección temprana de problemas en la curva de producción de huevos de gallinas ponedoras comerciales, De acuerdo con Bennett y Campbell, (2000), los clasificadores tradicionales pueden ser utilizados para la detección de casos raros, también llamados anomalías. Lindsay y Cox, (2005) afirman que las técnicas tradicionales de aprendizaje máquina, pueden ser una alternativa viable a la técnica de análisis de series de tiempo clásico. Bajo este enfoque, los clasificadores pueden ser utilizados en una serie de datos temporales para predecir si *k* días luego sucederá o no un caso anómalo.

De acuerdo con la metodología descrita a detalle en el capítulo anterior, se evaluó modelos con SVM y con ANN. En el caso del modelo con SVM se utilizó un enfoque dirigido a que sea un experto quien genera las características de entrada a partir de los datos crudos. En el caso del modelo con ANN se aplicó una técnica automática de selección de características.

4.1.1 Modelo SVM

Las entradas del modelo partieron de una colección de más de 30 características, con las que se llevó a cabo algunas pruebas preliminares, las cuales fueron evaluadas por un experto para dejar finalmente seis características relevantes como entrada al modelo. Se determinó que características como la línea genética de las aves, las variaciones estocásticas en la producción de huevos, la mortalidad acumulada o la pendiente de la curva de la semana, no proporcionan una mejora significativa en el modelo, y por lo tanto fueron descartadas (Mollazade, Omid, y Arefi, 2012).

Las características relevantes seleccionadas como patrones de entrada de la SVM se definen de la siguiente manera:

A. El porcentaje de producción de más de un día (número de huevos producidos durante un día / número de aves existentes) menos el porcentaje de la producción histórica para un día similar.
B. El porcentaje de la producción durante el día al final de la ventana deslizante, menos el porcentaje de producción durante el día al inicio de la ventana deslizante.
C. La producción durante el día, menos la producción de siete días antes.
D. El coeficiente de variación (desviación estándar / media * 100) de la segunda mitad de la ventana deslizante.
E. La desviación estándar de la primera mitad de la ventana deslizante menos la desviación estándar de la segunda mitad de la ventana deslizante.
F. La edad de las aves en semanas al final de la ventana deslizante.

4.1.1.1 Conformación de patrones

Las características relevantes, fueron seleccionadas dentro de la ventana deslizante, donde cada patrón de entrada tiene su valor correspondiente en el conjunto de salida. Estas salidas fueron cero o uno, dependiendo de si la etiqueta del día al final del intervalo de pronóstico fue positiva o negativa con respecto a la presencia de un problema en la curva. Este procedimiento se realiza para cada día durante el período de estudio, siempre extraer las mismas características fijas.

Para evaluar el intervalo de pronóstico, las salidas esperadas para cada ventana deslizante se ha tomado de patrón correspondiente en el conjunto de salida (intervalo de pronóstico de día cero), y un desplazamiento de tiempo (Lindsay y Cox, 2005) de uno a cinco días más tarde. De esa manera el algoritmo se entrena en base a las salidas próximos días esperados, y por lo tanto la SVM entrenada podría ser capaz de detectar los problemas antes que los expertos se percaten de que han sucedido.

4.1.1.2 Optimización de los parámetros de la SVM

La etapa experimental consistió en estudiar diferentes valores de configuración del sistema como: *kernel*s, parámetros, tamaño de ventana e intervalo de ejecución. Los diferentes ajustes de los parámetros de SVM se evaluaron en el conjunto de datos de test, utilizando pruebas estadísticas ANOVA y pruebas de comparación múltiple de *Tukey's* HSD para un valor de $p < 0,01$.

La configuración inicial del sistema fue la siguiente: valor del parámetro C igual a 0.1, tamaño de la ventana igual a siete e intervalo de pronóstico igual a uno. Con estos valores, se evaluó en primer lugar los *kernels*, seleccionando el mejor, para a continuación evaluar los distintos valores del parámetro C, seleccionando y fijando el mejor valor. Seguidamente, se probó algunos valores de tamaño de ventana y se fijó el mejor valor. Finalmente, se evaluó varios intervalos de predicción, con lo que se obtiene los resultados finales. A continuación, se detalla los resultados alcanzados.

Durante la optimización, se llevaron a cabo algunas pruebas preliminares por ensayo y error usando los *kernels* más comunes para SVM, según lo planteado por (Mollazade et al., 2012). Finalmente se decidió realizar un estudio exhaustivo con cuatro *kernel*s: 1) polinómico, 2) función de base radial o *radial basis function* (RBF), 3) cuadrática y 4) lineal. La evaluación de estos *kernel* se realizó teniendo valores fijos del parámetro C igual a 0.1, un tamaño de ventana igual a 7 y un intervalo de pronóstico igual a uno.

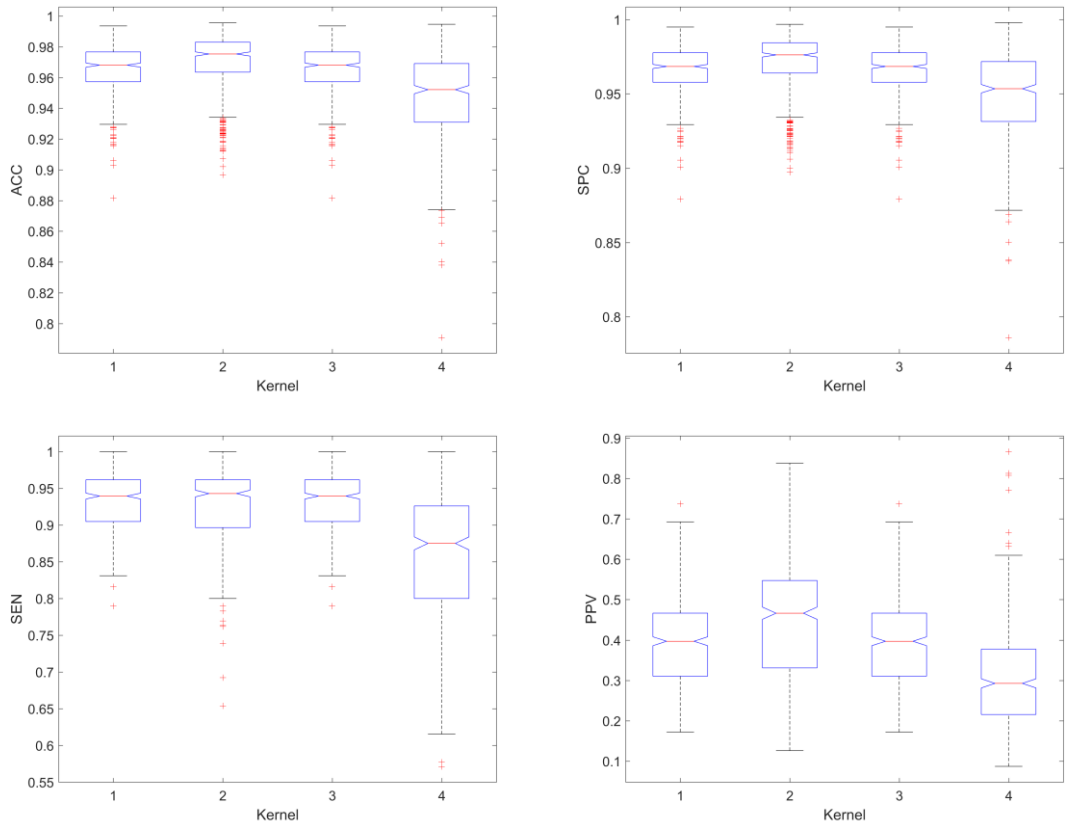


Figura 4.1. Diagramas de caja y bigotes de las métricas de desempeño para cada uno de los *kernels* evaluados.

En la Figura 4.1 se observa un diagrama de cajas con los resultados de la evaluación de los *kernels*. Como se observa, el *kernel* 4 (lineal) es el que da los peores resultados, mientras que los *kernels* 1, 2 y 3 (polinómica, RBF y cuadrático) dan resultados similares. Se realizó una prueba estadística ANOVA con el método de comparación múltiple *Tukey's* HSD, que se puede revisar en la Tabla 4.1.

Tabla 4.1. Comparación múltiple de los *kernels* evaluados

	Kernel			
	1 polinómica	**2 RBF**	**3 cuadrática**	**4 lineal**
ACC	0,9654 [a]	0,9687 [a]	0,9654 [a]	0,9475 [b]
SPC	0,9661 [a]	0,9696 [a]	0,9661 [a]	0,9492 [b]
SEN	0,9289 [a]	0,9203 [a]	0,9289 [a]	0,8546 b
PPV	0,3932 [b]	0,4445 [a]	0,3932 [b]	0,3045 c

Las filas con letras distintas, difieren significativamente según el método Tukey's HSD para un valor de p<0.01.

Del análisis anterior se decide elegir al *kernel* 2 (radial basis function), ya que es estadísticamente mejor en los cuatro parámetros evaluados y tiene una ventaja adicional sobre los demás *kernel*. El *kernel* es posiblemente el componente más importante de un algoritmo SVM (Mollazade et al., 2012; Suttorp y Igel, 2007; Zhao et al., 2010). De acuerdo con Fernández Pierna et al., (2006), y Zhiliang et al. (2015) el *kernel* RBF, ha demostrado ser una excelente función de núcleo para varias aplicaciones.

Dicho *kernel* RBF cuenta con un parámetro σ que ajusta el tamaño del *kernel*. Según Bennett y Campbell, (2000) y Zhao et al., (2010) este parámetro debe ser ajustado para mejorar las medidas de eficiencia de la predicción. Una técnica común para optimizarlo es mediante un gradiente ascendente (Suttorp y Igel, 2007). A continuación, se realiza un ajuste de éste parámetro, con valores entre uno y seis manteniendo fijos todos los demás parámetros.

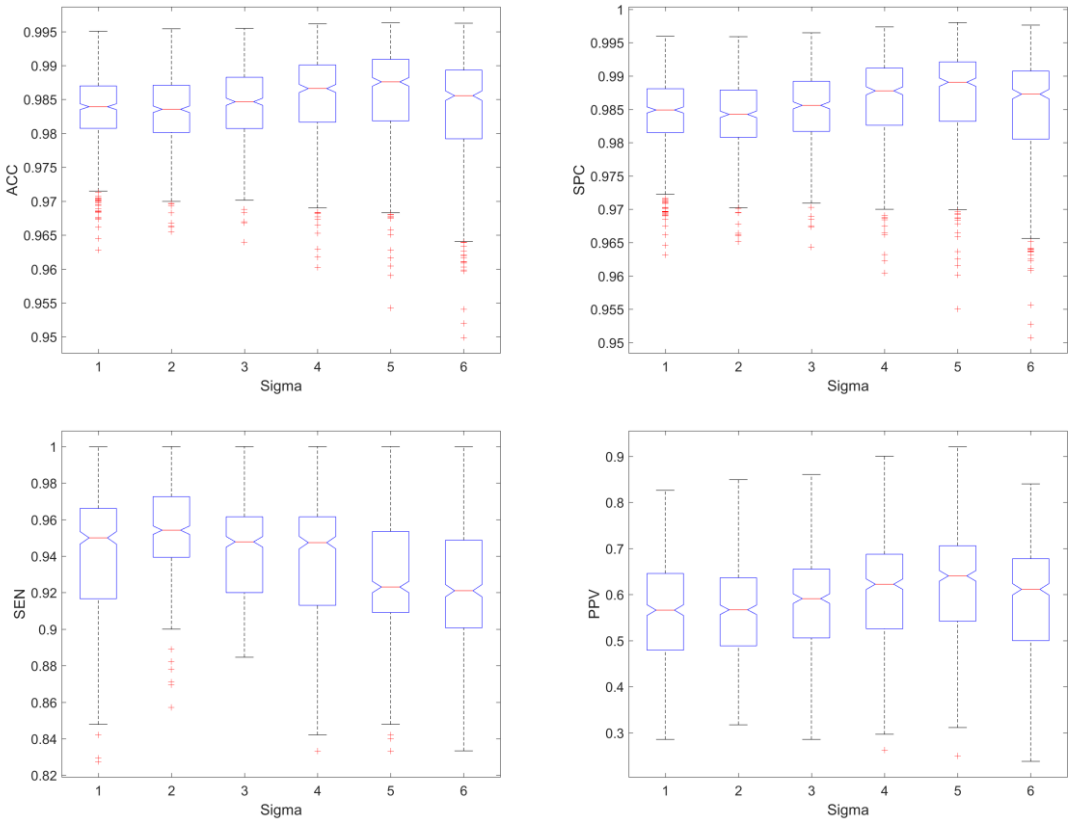


Figura 4.2. Diagramas de cajas y bigotes de las medidas de desempeño, de acuerdo al valor de σ.

En la Figura 4.2, se puede observar los resultados de la evaluación de distintos valores del parámetro σ. Como se observa, tanto para precisión, especificidad y valor predictivo positivo, los valores tienden a mejorar a medida que σ es mayor (hasta llegar a 5). Sin embargo, para sensibilidad, los valores tienden a empeorar a medida que σ es mayor. Para tener una herramienta que permita tomar la mejor decisión, se realizó una prueba estadística ANOVA con el método de comparación múltiple *Tukey's* HSD, que se puede revisar en la Tabla 4.2.

Tabla 4.2. Comparación múltiple de distintos valores de σ para cada medida de eficiencia de la predicción.

	σ = 1	**σ = 2**	**σ = 3**	**σ = 4**	**σ = 5**	**σ = 6**
ACC	0,9833 [b]	0,9833 [b]	0,9842 [ab]	0,9853 [a]	0,9856 [a]	0,9833 [b]
SPC	0,9843 [b]	0,9840 [b]	0,9851 [b]	0,9864 [ab]	0,9869 [a]	0,9847 [b]
SEN	0,9406 [b]	0,9520 [a]	0,9419 [b]	0,9365 [b]	0,9259 [c]	0,9199 [c]
PPV	0,5657 [b]	0,5647 [b]	0,5836 [b]	0,6100 [ab]	0,6223 [a]	0,5858 [b]

Las filas con letras distintas, difieren significativamente según el método Tukey's HSD para un valor de p<0.01.

Debido a que la base de datos de la producción de huevos, tiene muchos más patrones negativos que patrones positivos, la métrica especificidad tiene un interés especial respecto a la cantidad de falsos positivos. A partir de este enfoque, un valor de *σ* igual a cinco es mejor. Otro enfoque para apoyar esta decisión se afirma por Fernández Pierna et al., (2006), quienes sostienen que la capacidad de generalización aumenta mientras *σ*, se obtienen mejores desempeños de los modelos.

Se decide un valor de *σ* igual a 5, ya que presenta la mejor precisión, especificidad y valor predictivo positivo, con una significación estadística. En el caso de la sensibilidad, un valor de *σ* igual a 5 da un resultado de 0,9259, que de acuerdo al test de *Tukey's* HSD está en el grupo c, es decir el de menor valor entre los evaluados.

Una vez decidido fijar el *kernel* RBF, con un valor de *σ* igual a 5, se realizó una evaluación variando el parámetro C. Para ello se mantuvieron fijos los valores iniciales de tamaño de ventana e intervalo de pronóstico y se probó únicamente con valores bajos de C, considerando que valores altos de éste parámetro pueden acarrear problemas de sobreajuste (Mucherino et al., 2009).

En la Figura 4.3 se puede revisar los resultados de la evaluación de diferentes valores del parámetro C. Como se observa, pequeñas modificaciones del parámetro C no generan diferencias apreciables, tanto para ACC, SPC o PPV. No obstante, en la SEN, la modificación de éste parámetro si genera ligeros incrementos. Para tomar la mejor decisión, se realizó una prueba estadística ANOVA con el método de comparación múltiple *Tukey's* HSD, que se puede revisar en la Tabla 4.3.

Tabla 4.3. Comparación múltiple de distintos valores del parámetro C, para cada medida de eficiencia de la predicción.

	C = 0,01	**C = 0,1**	**C = 0,15**	**C = 0,2**	**C = 0,25**
ACC	0.9845 a	0.9856 a	0.9852 a	0.9849 a	0.9847 a
SPC	0.9859 a	0.9869 a	0.9864 a	0.9860 a	0.9858 a
SEN	0.9185 b	0.9256 b	0.9328 a	0.9343 a	0.9358 a
PPV	0.6063 a	0.6222 a	0.6131 a	0.6057 a	0.6025 a

Las filas con letras distintas, difieren significativamente según el método Tukey's HSD para un valor de p<0.01.

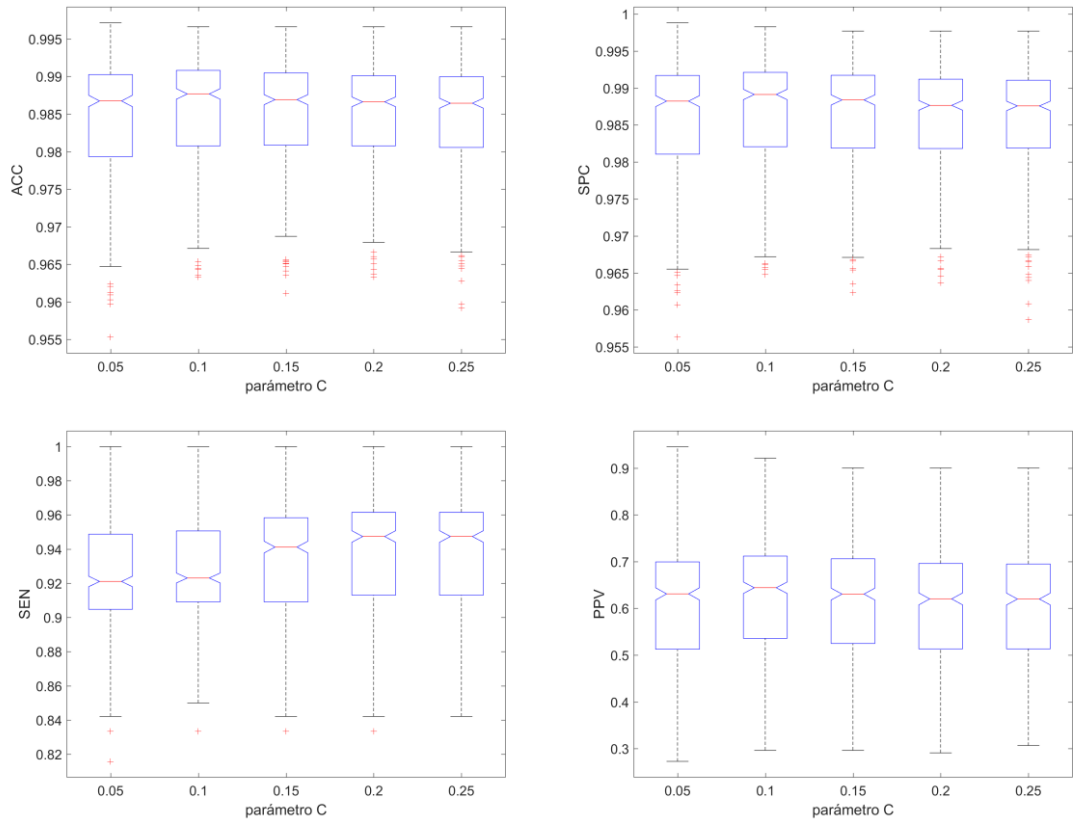


Figura 4.3. Diagramas de cajas y bigotes de las medidas de desempeño, según el valor del parámetro C.

De acuerdo a los resultados se podría decidir cualquier valor del parámetro C, mayor o igual a 0,15. Considerando que mientras menor sea el valor, menos posibilidad de sobreajuste existe, decidimos fijar 0,15 como valor del parámetro C.

4.1.1.3 Selección del tamaño de ventana óptimo

El tamaño de la ventana, expresa la cantidad de datos de los días anteriores al evento que se le suministra al sistema para configurar los patrones de entrada. Considerando que las características calculadas por el experto guardan un componente cíclico (semanas), se decidió utilizar múltiplos de 7 para evaluar el tamaño óptimo de la ventana temporal. Esta decisión se respalda en los resultados empíricos de pruebas preliminares en las que valores distintos a un múltiplo de 7, resultaba en un peor desempeño.

Tabla 4.4. Comparación múltiple de distintos valores de tamaño de ventana, para cada medida de eficiencia de la predicción.

	WS = 7	**WS = 14**	**WS = 21**	**WS = 28**
ACC	0,9659 c	0,9852 a	0,9838 ab	0,9821 b
SPC	0,9670 c	0,9864 a	0,9850 ab	0,9836 b
SEN	0,9020 c	0,9318 a	0,9320 a	0,9168 b
PPV	0,4300 c	0,6122 a	0,5972 ab	0,5808 b

Las filas con letras distintas, difieren significativamente según el método Tukey's HSD para un valor de $p<0.01$.

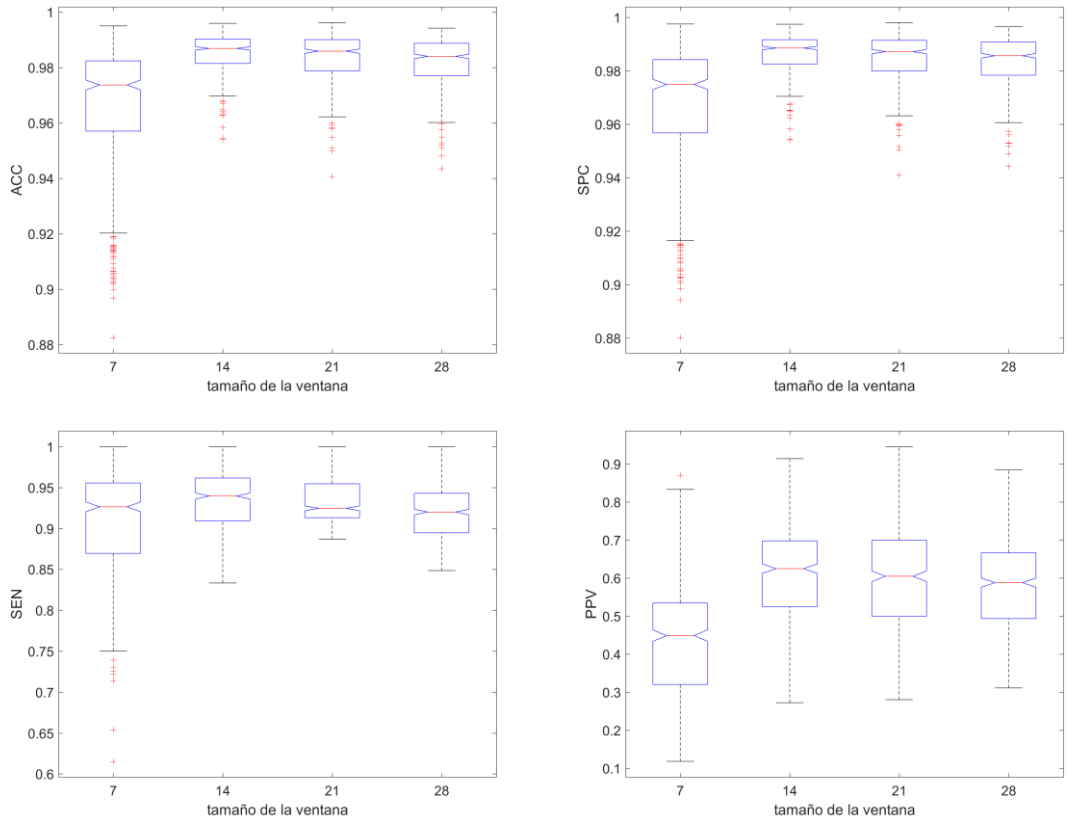


Figura 4.4. Diagramas de cajas y bigotes: precisión, especificidad, sensibilidad y valor predictivo positivo según el tamaño de la ventana.

Por tanto, el tamaño de la ventana temporal se evaluó con valores desde 7 días hasta 28 días. En la Figura 4.4 se puede observar los resultados de la evaluación en diagramas de cajas. Claramente se observa que un tamaño de ventana igual a 7 genera los peores resultados en las 4 medidas de evaluación y que valores de 14, 21 y 28 generan resultados similares. Por lo que se procede a realizar una prueba estadística ANOVA con el método de comparación múltiple *Tukey's* HSD, que se puede revisar en la Tabla 4.4.

Las características relevantes B, D y E, dependen de la cantidad de datos proporcionados en la ventana temporal con el fin de calcular un valor único para cada característica, que constituye una parte de un patrón.

Los resultados de este trabajo, mostraron que un tamaño de ventana igual a 14 días genera los mejores resultados en todas las métricas de rendimiento. Esto se entiende ya que un tamaño de ventana de 7 días, no proporciona datos suficientes, en consecuencia. Por otra parte, un tamaño de ventana de más de 28 días, contiene una excesiva cantidad de información que enmascara a los patrones positivos. Los resultados del test de *Tukey's* HSD, apuntan inequívocamente a que el tamaño de ventana óptimo para este tipo de problema, es de 14 días.

4.1.2 Modelo ANN

Al igual que el modelo SVM, el enfoque propuesto se basa en un clasificador que, combinado con ventana temporal, debe ser capaz de detectar anomalías con respecto

a comportamientos típicos. Siguiendo lo propuesto por Bennett y Campbell (2000), quienes se plantearon la posibilidad de que las técnicas de clasificación fueran utilizadas para detectar casos anormales.

Todos los resultados aquí mostrados corresponden a datos del conjunto test, obtenidos mediante una técnica de *cross-validation* con 5 capas, de acuerdo a lo enunciado en el capítulo anterior.

4.1.2.1 Conformación de los patrones de entrada

La conformación de los patrones de entrada se realizó tomando los datos de producción de los días anteriores basándose en el método de ventana deslizante. El número de huevos producidos en el día, el número de aves que mueren en el día, el número de aves vivas en el día, y el número de huevos trizados, fueron considerados como datos en crudo para este estudio. El tamaño de la ventana, determina la cantidad de días hacia atrás que fueron considerados para la captura de estos datos.

A diferencia del modelo SVM, en que se requirió que un experto en avicultura proponga más de 30 características como entradas del modelo, en este estudio fue realizado un proceso automático de selección de características relevantes. Partiendo del análisis de los datos en crudo, y sin que se requiera de la revisión de un experto se realizó dicha selección, para ello fue necesario aplicar una técnica basada en un filtro univariado dentro de la ventana temporal deslizante. Actualmente esta práctica se ha convertido en una necesidad en muchas aplicaciones (Isabelle Guyon y Elisseeff, 2003) debido a que previene el sobreajuste del modelo, mejora el desempeño y reduce el tiempo de cómputo (I. Guyon et al., 2008; Saeys et al., 2007).

En pruebas preliminares se fijó una arquitectura inicial de una capa oculta con 10 neuronas, un parámetro S de 0.5 y un adelanto igual a 1. Con estos valores de parámetros fijos, el tamaño de la ventana temporal y el valor del umbral del filtro univariado, fueron evaluados simultáneamente utilizando la técnica *grid search* explicada en el Capítulo III.

La Figura 4.5 muestra los resultados de los promedios de 100 repeticiones de la técnica de *cross-validation* con 5 capas. Estos se encuentran dispuestos en una *grid search* cuyos ejes son el tamaño de ventana y el umbral del filtro univariado. En la Figura 4.5a, se observa que un umbral mayor a 40 junto con un tamaño de ventana mayor a 15, muestran los mejores resultados de ACC. Algo similar sucede para la especificidad que puede ser observada en la Figura 4.5b. Esta particularidad, era esperada, dado que, en una base de datos no balanceada cuyas salidas deseadas o *targets* sean mayoritariamente negativos, la influencia de estos valores en la ACC general del modelo es alta. La SEN y PPV observadas en las Figura 4.5c y 4.5d, respectivamente, denotan un área más marcada que se encuentra entre 40 y 70 del umbral de filtro univariado, y entre 15 y 20 del tamaño de ventana.

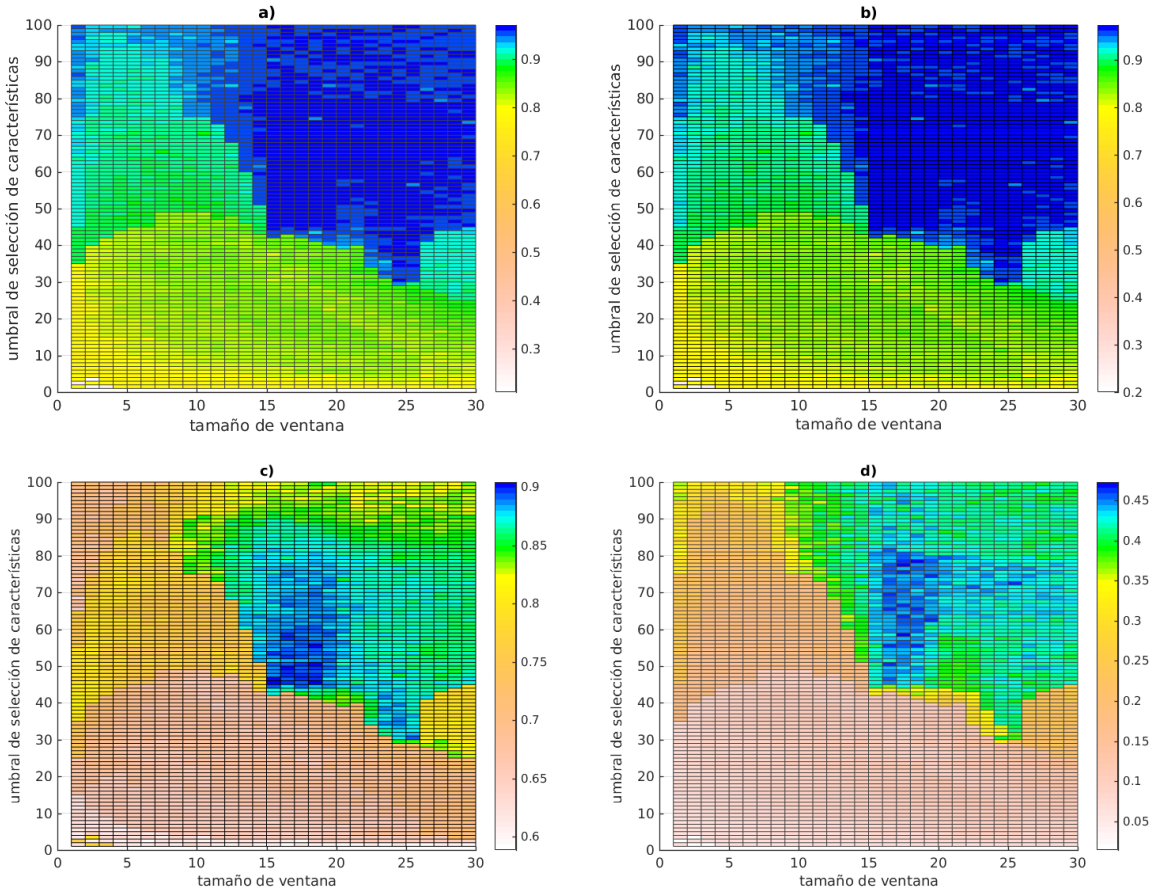


Figura 4.5 Gráfico de la técnica *grid search* para la selección del tamaño de ventana y el umbral de selección de características. a) ACC, b) SPC, c) SEN, d) PPV.

En esta etapa el tamaño de paso de la ventana se mantuvo igual a 1 (Saeed y Snášel, 2014), fueron evaluados tamaños de ventana entre 0 y 30. Según el criterio del autor, estudiar tamaños de ventana mayores es poco práctico dado que sería necesario que un lote esté en producción por un periodo de tiempo demasiado largo antes de que el modelo pueda ser utilizado. De manera simultánea fue optimizado un umbral entre 0 y 10 para la selección de características, basándose en el percentil *p-value* de una técnica de filtro univariado que usa un *t*-test.

Considerando que la selección de características relevantes según las métricas de desempeño del modelo está condicionado por el tamaño de la ventana (Frank, Davey, y Hunt, 2001), y el umbral del filtro univariado, se ha considerado un tamaño de ventana igual a 18 junto con un umbral del filtro univariado igual a 65. En este punto se observa los mejores resultados para las cuatro métricas de desempeño. Los valores de tamaño de la ventana y del umbral del filtro univariado que se alejan de los óptimos, tienen un desempeño menor, ya sea porque la información contenida en las características seleccionadas no es suficiente, o debido a que resulta excesiva y ruidosa para el modelo.

En el modelo SVM, es requerido que los tamaños de ventana que sean múltiplos de 7, esta restricción está dada por la manera en que las características fueron calculadas por parte del experto; mientras que, en el modelo aquí propuesto, no existe tal restricción. De acuerdo con los resultados experimentales, la información contenida en

las características relevantes seleccionados, permite modelar de mejor manera el problema. Estas características relevantes son:

- El número de huevos producidos 5 días al inicio y 4 días al final de la ventana temporal.
- El número de aves muertas en los últimos 10 días de la ventana temporal.
- El número de aves existentes 14 días al inicio de la ventana temporal.
- Todos los datos de huevos trizados en la ventana temporal.

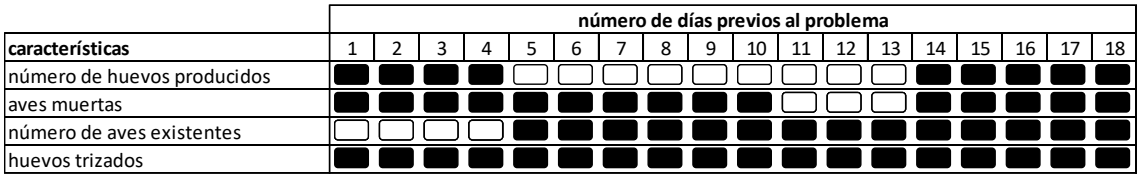


Figura 4.6. Características seleccionadas con un umbral de FS de 65 en una ventana de tamaño igual a 18 días.

Una ilustración de estas características relevantes puede ser observada en la Figura 4.6, en la que las características seleccionadas se muestran marcados en negro. Las características seleccionadas contienen la información más relevante para el sistema. Con respecto al número de huevos producidos, resultan de mayor importancia la diferencia entre los huevos puestos en los primeros cuatro días y los cinco últimos. Con el número de gallinas muertas, ocurre algo similar. Sin embargo, con respecto al número de huevos rotos, el sistema requiere tener la información completa sobre, cómo ha sido su evolución en los 18 días de la ventana temporal deslizante. En cuanto al número de gallinas existente, no resulta importante la cantidad de los primeros cuatro días, sino del día 5 al 18.

4.1.2.2 Arquitectura de la ANN

La búsqueda de la mejor arquitectura de la ANN comprende un proceso en el que se define previamente un rango de posibles arquitecturas y se selecciona la mejor basándose en su desempeño en datos del conjunto test (Moody y Utans, 1994).

Fueron evaluadas arquitecturas de una y de dos capas ocultas, con varias combinaciones de 0, 25 y 50 neuronas en cada capa oculta, el algoritmo de optimización utilizado fue una variación de la técnica de gradiente descendiente conocida como *scaled conjugate gradient backpropagation* (Møller, 1993).

En esta etapa del estudio las siete arquitecturas fueron evaluadas utilizando una técnica de análisis de varianza (ANOVA) y la técnica de comparación *Tukey's* HSD (Abdi y Williams, 2010) para un valor de $p < 0.01$, en concordancia con lo planteado por Herrera et al., (2004) y Rivero et al., (2011) quienes afirman que es necesario probar varias arquitecturas diferentes hasta que encuentra una que proporciona buenos resultados.

Para la selección de la arquitectura fueron fijados en la etapa anterior, el tamaño de ventana y el umbral del filtro para selección de características relevantes. El valor del parámetro S fue igual a 0.5 y el intervalo de pronóstico igual a 1.

Tabla 4.5. Comparación múltiple de las arquitecturas evaluadas.

Arquitectura	ACC		SPC		SEN		PPV	
de la ANN	media	D.E.	media	D.E.	media	D.E.	media	D.E.
[no HL]	0.955[a]	0.062	0.956[a]	0.063	0.900[a]	0.083	0.387[b]	0.159
[25]	0.968[a]	0.021	0.970[a]	0.022	0.880[a]	0.057	0.451[a]	0.174
[25,25]	0.961[a]	0.026	0.962[a]	0.027	0.884[a]	0.062	0.411[ab]	0.178
[25,50]	0.959[a]	0.028	0.961[a]	0.029	0.884[a]	0.058	0.405[ab]	0.175
[50]	0.967[a]	0.020	0.970[a]	0.021	0.879[a]	0.059	0.450[a]	0.175
[50,25]	0.962[a]	0.026	0.964[a]	0.027	0.876[a]	0.070	0.418[ab]	0.180
[50,50]	0.962[a]	0.025	0.964[a]	0.026	0.876[a].	0.061	0.416[ab]	0.175

Las filas con letras distintas, difieren significativamente según el método Tukey's HSD para un valor de $p<0.01$.

La Tabla 4.5 muestra las arquitecturas de la red de neuronas artificiales que fueron evaluadas, es necesario señalar que ninguna de las arquitecturas de dos capas superó a una arquitectura de red con una sola capa oculta conformada por 25 neuronas. Esta es la que muestra las mejores métricas de desempeño. Una arquitectura de una capa oculta con 50 neuronas no muestra diferencias estadísticamente significativas respecto a la de 25.

Al tener que decidir entre dos iguales, es preferible optar por la de menor complejidad, ya que los modelos más simples suelen ser menos tendientes al sobreajuste. Por lo tanto, para continuar con la optimización se decide una arquitectura de una sola capa oculta con 25 neuronas.

4.1.2.3 Optimización del parámetro de ponderación del aprendizaje

Tal como sucede en otras aplicaciones de detección de eventos poco frecuentes, los patrones raros son los que transmiten la información más relevante. Esto se evidencia en casos como la detección temprana de fallas en equipamientos industriales, en la que el sistema debe detectar los fallos a partir de datos en los que lo normal es que existan mayor cantidad de patrones de funcionamiento normal que anormal (Dai y Gao, 2013). Otro caso es en el campo de la medicina para el diagnóstico de enfermedades, en el que hay más individuos sanos que enfermos (Tavares, Oliveira, Cabral, Mattos, y Grigorio, 2013).

La detección de anomalías en la curva de producción presenta un reto similar, ya que lo común es que no existan caídas de producción, y se encuentran pocos casos positivos. Esto hace que además de la ACC, sea necesario evaluar la SEN, SPC y PPV acorde con lo planteado por Sun et al. (2009). De esta manera seleccionó los valores del parámetro que confieren al modelo un mejor desempeño en las cuatro métricas.

Vannucci y Colla, (2016), plantean que los métodos de clasificación tradicionales suelen fallar cuando se enfrentan a conjuntos de datos no balanceados. Esto se debe a que principalmente la clasificación errónea de las muestras raras es penalizada de igual manera que la clasificación errónea de los frecuentes. Como consecuencia, los clasificadores tienden a centrarse en eventos más probables, por este motivo en el presente trabajo se definió un parámetro ajustable S a partir del cual se asigna una matriz de pesos de compensación a los patrones raros, de acuerdo a lo planteado por

Elkan (2001), este ajuste permitió afinar el modelo para obtener una mejor respuesta en cuanto al valor predictivo positivo y la sensibilidad.

En esta etapa fue modificado el parámetro S, con valores entre 0.05 y 0.95 a intervalos de 0.05. El modelo utilizó el tamaño de ventana y el umbral del filtro univariado fijados en la primera etapa, la arquitectura fijada en la segunda etapa y un adelanto igual a 1.

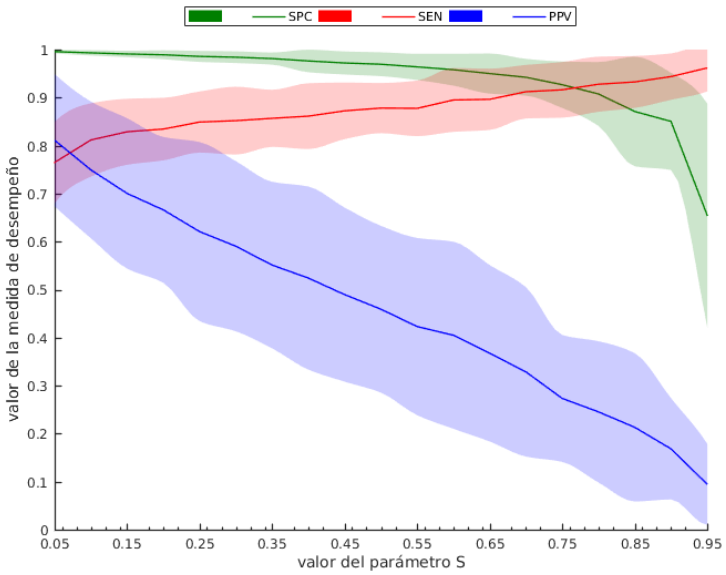


Figura 4.7 Gráfico de desempeño del modelo a distintos valores del parámetro S.

En la Figura 4.7 se observa el comportamiento de las medidas de desempeño según los valores del parámetro S, la línea representa el promedio de 100 evaluaciones y el área sombreada se corresponde con la desviación típica.

Como se puede apreciar, los valores bajos del parámetro S, mantienen una especificidad alta, debido a que hay una mayor ponderación de la importancia de los patrones negativos, sin embargo, la sensibilidad es baja. A medida que se modifica el valor del parámetro S incrementándolo, ocurre lo contrario, debido a que se les asigna mayor ponderación a los patrones positivos en detrimento de los negativos, con lo que aumenta la sensibilidad y baja la especificidad. El PPV, se reduce también en la medida que se incrementa el valor del parámetro S, ya que aumenta el número de falsas alarmas.

Un buen sistema, debe mantener un balance adecuado entre una especificidad, y una sensibilidad alta. Además, simultáneamente, el sistema debe tener un valor razonable de falsas alarmas. Por esa razón se elige un valor del parámetro S igual a 0.1, ya que, en este punto, se maximiza la sensibilidad sin que eso afecte excesivamente al valor predictivo positivo.

4.1.3 Evaluación comparativa de ambos modelos a distintos intervalos de predicción

Con la finalidad de evaluar el número de días previos a la presencia de la anomalía, en los que el modelo mantiene un adecuado desempeño, los datos fueron reorganizados

aplicando un cambio en el día. Esto se traduce en un adelanto a las salidas deseadas o *targets*. Esta característica genera una modificación entre los patrones originales y la salida deseada, lo que según Lindsay y Cox, (2005), hace que el modelo sea entrenado en base a las salidas esperadas en los próximos días, confiriéndole la capacidad de detectar de forma anticipada las anomalías.

El intervalo de pronóstico se debe ajustar a la medida de la exigencia del momento, un valor igual cero implica que el sistema funciona como alerta temprana, es decir que es capaz de identificar el problema en los primeros momentos del mismo, valores mayores o iguales a uno implican que el sistema funciona como sistema de predicción, es decir que es capaz de identificar el problema con n días antes de que se evidencie el problema. Se evaluó valores de intervalo de pronóstico entre cero y cinco días.

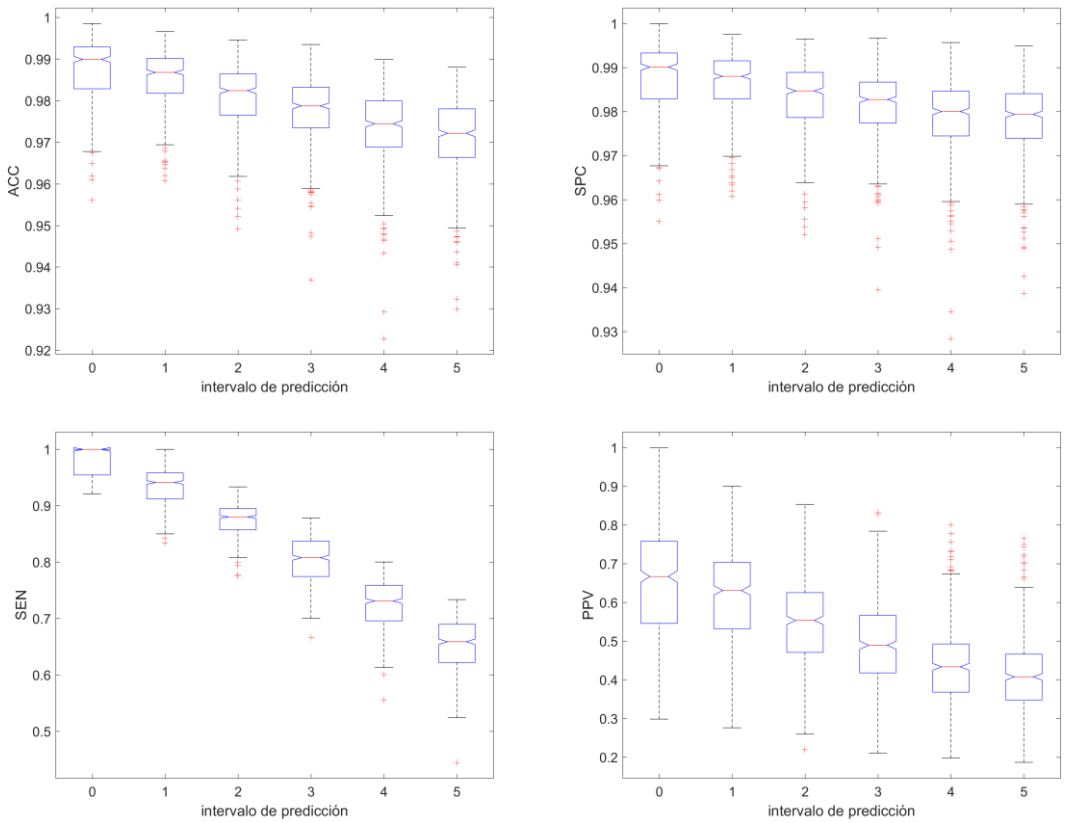


Figura 4.8. Diagramas de cajas y bigotes de las medidas de desempeño según el intervalo de pronóstico.

Los resultados se observan en la Figura 4.8. Como era de esperar, mientras menor sea el intervalo de pronóstico se obtiene un mejor desempeño en todas las medidas de eficiencia de la predicción. A medida que aumenta el intervalo de pronóstico, el desempeño disminuye. Para el caso de la sensibilidad, el intervalo de pronóstico le afecta con más intensidad que a las otras medidas de eficiencia de la predicción.

Considerando que un buen sistema de predicción es aquel que logra una sensibilidad mayor o igual a 0.8, se asume, que el intervalo de pronóstico óptimo está entre 0 días (únicamente alarma) hasta 3 días previos a la caída de producción; la selección de uno u otro valor va a depender de cuán preciso, sensible y específico se desea que el

sistema se desempeñe. Se realizó una prueba estadística ANOVA con el método de comparación múltiple *Tukey's* HSD, que se puede revisar en la Tabla 4.6.

Tabla 4.6. Comparación múltiple de distintos valores de intervalo de pronóstico para cada métrica de desempeño del modelo.

	FI = 0	**FI = 1**	**FI = 2**	**FI = 3**	**FI = 4**	**FI = 5**
ACC	0,9874 a	0,9854 b	0,9811 c	0,9776 d	0,9735 e	0,9713 f
SPC	0,9876 a	0,9865 a	0,9835 b	0,9814 c	0,9789 d	0,9783 d
SEN	0,9783 a	0,9333 b	0,8738 c	0,8030 d	0,7229 e	0,6483 f
PPV	0,6518 a	0,6135 b	0,5480 c	0,4940 d	0,4419 e	0,4090 f

Las filas con letras distintas, difieren significativamente según el método Tukey's HSD para un valor de $p<0.01$.

El modelo SVM realiza una ACC de 0,9874, SPC de 0,9876, la SEN de 0,9783 y un VPP de 0,6518, en un intervalo de pronóstico de cero, en este caso, el modelo funciona como una alerta temprana. A medida que aumenta intervalo de pronóstico, las métricas de desempeño disminuyen, en el caso de SEN, el intervalo de pronóstico es afectado con mayor intensidad que las otras métricas.

A valores de intervalos de predicción óptimos, el modelo es capaz de identificar el problema antes de que haya sido evidente para los expertos. En algunos casos, se encontró que el modelo era capaz de detectar como falsos positivos, algunos días antes de que ocurra un evento.

Tabla 4.7. Comparación múltiple de distintos valores de intervalo de pronóstico para cada medida de eficiencia de la predicción.

	FI = 0	**FI = 1**	**FI = 2**	**FI = 3**	**FI = 4**	**FI = 5**
ACC	0.993±0.003	0.990±0.005	0.986±0.006	0.983±0.009	0.980±0.01	0.976±0.012
SPC	0.995±0.003	0.994±0.005	0.992±0.006	0.991±0.009	0.989±0.011	0.987±0.014
SEN	0.864±0.061	0.817±0.078	0.736±0.091	0.664±0.104	0.580±0.132	0.482±0.152
PPV	0.813±0.119	0.756±0.14	0.713±0.166	0.672±0.19	0.624±0.22	0.567±0.233

La Figura 4.9 muestra la distribución de los resultados de las medidas de desempeño en 100 repeticiones de la técnica de *cross-validation* con 5 capas, para los intervalos de pronóstico, los valores de media y desviación estándar se encuentran en la Tabla 4.7.

En la Tabla 4.7 y en la Figura 4.9a y Figura 4.9b se observa que la ACC y SPC muestran valores por encima de 0.9758 y 0.9874 respectivamente, para todos los intervalos de pronóstico evaluados, sin embargo, se evidencia una media mayor con menor dispersión para los intervalos de predicción igual a cero y un día.

En la Tabla 4.7 y en la Figura 4.9c, se observa la SEN, cuyos resultados en promedio caen por debajo de 0.8 a partir de un FI de 2 días, en la Figura 4.9d se muestra el PPV, que se conserva por encima de 0.7 hasta el intervalo de pronóstico igual a tres días.

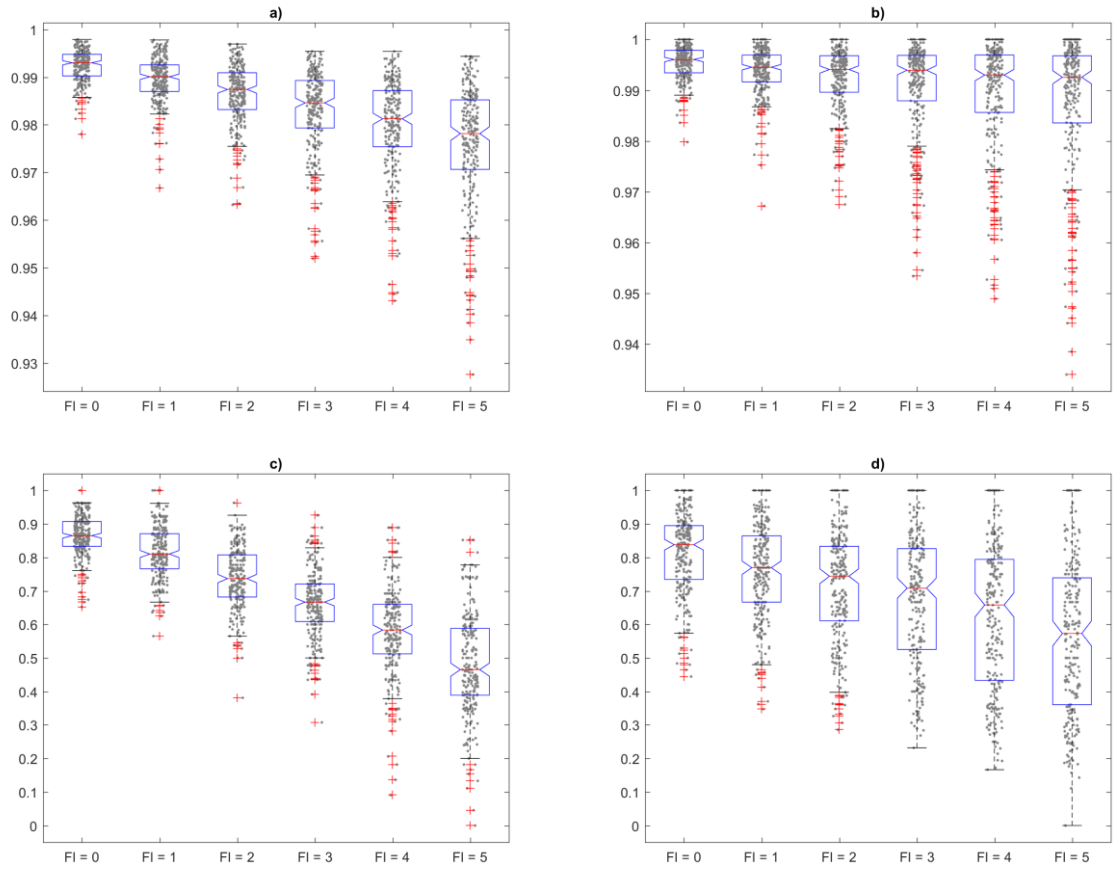


Figura 4.9. Diagramas de cajas y bigotes: para intervalos de predicción de cero a cinco: a) ACC, b) SPC, c) SEN, d) PPV.

Con respecto al desempeño del modelo, otros sistemas como el desarrollado por Lokhorst y Lamaker (1996), lograron una sensibilidad del 0.64 y una especificidad de 0.72 para el mismo día en que sucede la anomalía. El sistema propuesto aquí, basado en ANN tiene una SEN de 0.8639 y una SPC de 0.9954, además la ACC es de 0.9925 y el valor predictivo positivo de 0.8125, para detectar anomalías el mismo día en que sucede la anomalía.

El modelo SVM con las características que fueron elaboradas por un experto, obtuvo una ACC de 0.9854, SPC de 0.9865 SEN de 0.9333, y PPV, para un intervalo de pronóstico de un día previo a la presentación de la anomalía. El modelo ANN alcanza un valor menor en cuanto a SEN 0.8169, sin embargo, mejora al modelo SVM en las otras 3 métricas de desempeño del modelo ACC 0.9896, SPC 0.9936 y PPV 0.7564.

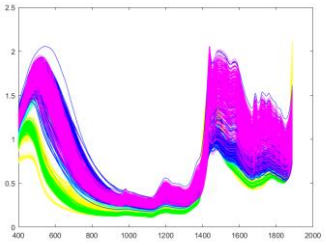
La mejora en cuanto a PPV, es de más de 23% lo que implica una menor cantidad de falsas alarmas, y por lo tanto un avance significativo en cuanto al uso práctico en el monitoreo continuo de los lotes de producción avícola.

A mayores intervalos de predicción, los parámetros de rendimiento disminuyen, en el caso de la sensibilidad, el intervalo de pronóstico que afecta con más intensidad que a otras métricas. En opinión del autor, los valores de sensibilidad por encima de 0,8 son aceptables. Por lo tanto, el intervalo de pronóstico óptimo se considera que es de un día.

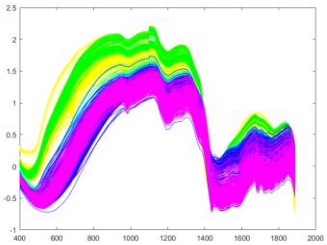
4.2 Aplicación de técnicas de ML en la industria de la caña de azúcar

Para la conformación de los patrones de entrada, en primer lugar, se realiza el pre-procesamiento de las señales obtenidas por el espectrómetro NIR. En contraposición a lo anteriormente planteado, Xu et al (2008) indica que la combinación de métodos de pre-procesamiento mejora la estabilidad de los modelos y los resultados en términos de RMSE ya que toma ventaja de la información complementaria de las distintas técnicas de pre-procesamiento. En consecuencia, se realizó pruebas preliminares con los métodos de pre-procesamiento más comunes, que fueron explicados en la sección de metodología. En base a los resultados de estas pruebas, se definió nueve combinaciones que se explican a continuación:

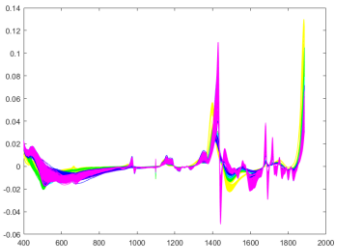
1. Raw: señal de reflectancia sin procesar (tal como se extrae del instrumento).
2. *Beer Lambert* (BL): a la señal sin procesar se le aplica la teoría de *Beer Lambert*.
3. First spectral derivative (FSD): cálculo de la primera derivada espectral a partir de la señal sin procesar.
4. SNV to Raw: a la señal sin procesar, se le calcula la Standard Normal Variate.
5. SNV to BL, cálculo de la SNV a la señal procesada en la combinación 2.
6. SNV to FSD, cálculo de la SNV a la señal procesada en la combinación 3.
7. SNV to Raw + *detrend*: a la señal procesada en la combinación 4, se le aplica una técnica de *detrending*.
8. SNV to BL + *detrend*: a la señal procesada en la combinación 5, se le aplica una técnica de *detrending*.
9. SNV to FSD + *detrend*: a la señal procesada en la combinación 6, se le aplica una técnica de *detrending*.



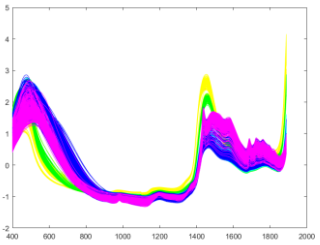
1. Raw



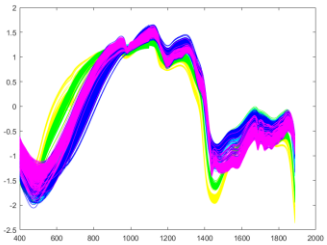
4. *Beer Lambert*



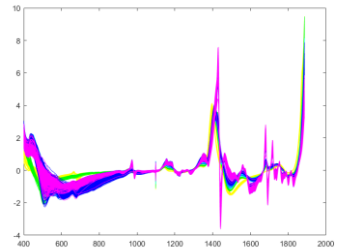
7. First spectral derivative



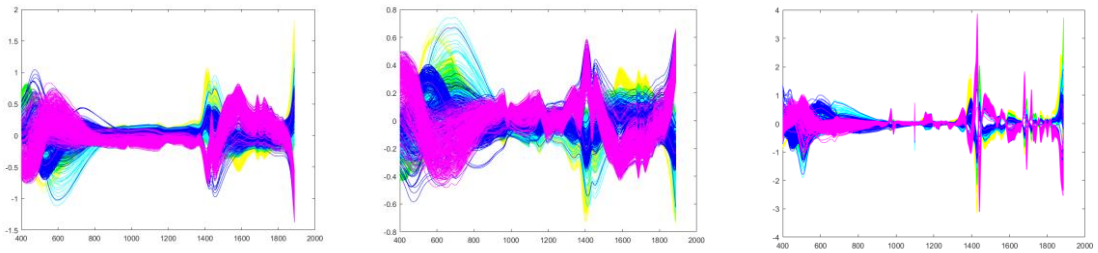
2. SNV to Raw



5. SNV to BL



8. SNV to FSD



3. SNV to Raw + *detrend* 6. SNV to BL + *detrend* 9 SNV to FSD + *detrend*

Figura 4.10. Cambios en los espectros NIR al aplicar las técnicas de pre-procesamiento

En la Figura 4.10 se puede observar nueve alternativas transformaciones de los espectros NIR que resultan de combinar las técnicas de pre-procesamiento básicas planteadas anteriormente, estas serán evaluadas en la primera fase. Para una mejor ilustración del lector en esta sección los espectros fueron coloreados de acuerdo a sus respectivos valores de ° Brix, nótese cómo se suavizan y acentúan las diferencias entre las distintas tonalidades en el espectro NIR, al aplicar las técnicas de pre-procesamiento.

4.2.1 Selección de la técnica de pre-procesamiento y de las características relevantes

Las nueve alternativas de técnicas de pre-procesamiento propuestas fueron evaluadas a distintas intensidades del umbral de selección de características, lo que se realizó modificando los valores del percentil umbral para el *p-value* de la técnica *T*-test para selección de características.

° Brix % Sacarosa

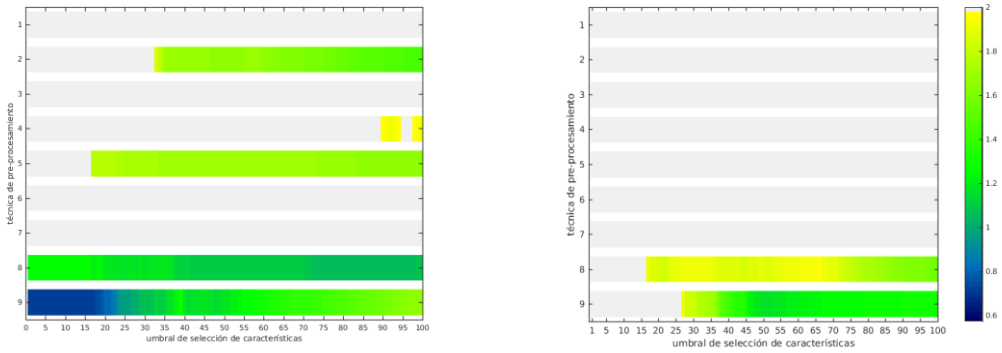


Figura 4.11. RMSE en CV alcanzado con las 9 técnicas de pre-procesamiento a distintos valores de umbral de selección de características.

La Figura 4.11 presenta los resultados del RMSE en CV alcanzados con las 9 técnicas de pre-procesamiento a distintos valores de umbral de selección de características en la que se observa que, tanto para el caso de ° Brix como para Sacarosa, la técnica de pre-procesamiento número 9 (listada en la Figura 4.10) es la que ofrece mejores resultados, dado que los menores valores de RMSE se registran en el rango de percentil comprendido entre 1 y 17, para ° Brix, mientras que para Sacarosa, los menores valores de RMSE fueron registrados en el rango de percentil comprendido entre 45 y 55. La

técnica de pre-procesamiento 9 consiste en la combinación del cálculo de la primera derivada, que luego es normalizada con la técnica SNV y finalmente se le extrae la tendencia.

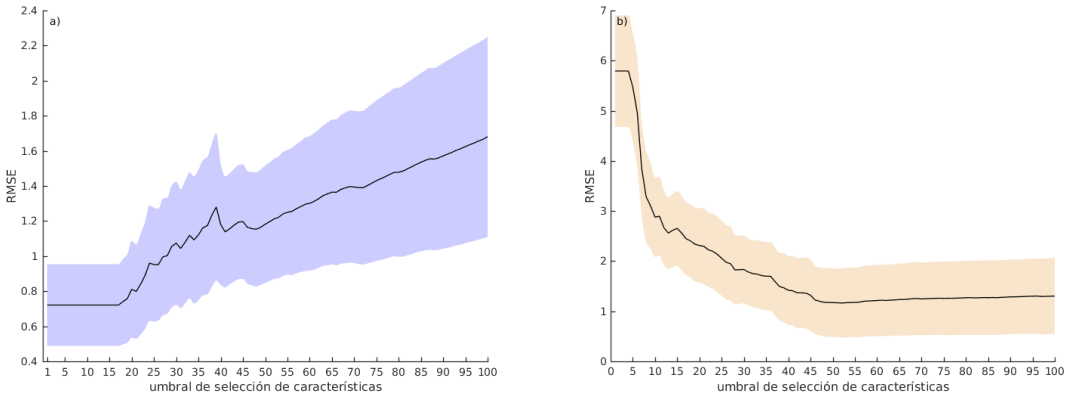


Figura 4.12. RMSE en CV alcanzado en el modelo a) ºBrix y b) Sacarosa, con la técnica de pre-procesamiento número 9, a distintos valores de umbral de selección de características.

La Figura 4.12 muestra el RMSE en CV alcanzado con la técnica de pre-procesamiento número 9. A distintos valores de umbral de selección de características, se puede observar que para el caso del modelo de ° Brix los valores óptimos del umbral de selección de características están entre 1 y 17. Se selecciona un valor igual a 10, y, por su parte el modelo de Sacarosa tiene un valor óptimo del umbral de selección de características igual a 52.

Las características seleccionadas con un umbral igual a 10, se presentan en la Figura 6, en la que se resaltan en bandas verticales sombreadas, las características más relevantes para el modelo de ° Brix.

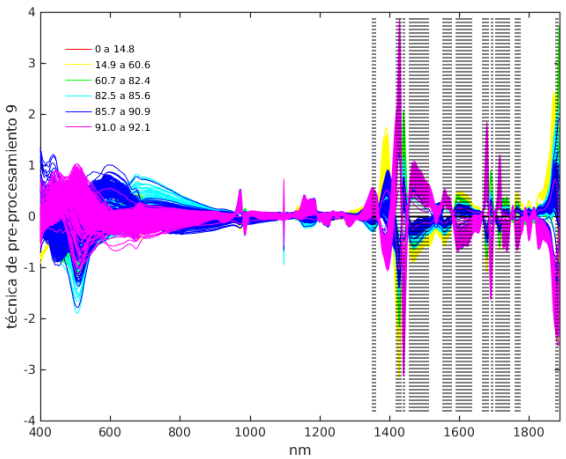


Figura 4.13. Bandas espectrales (características) seleccionados para construir el modelo de predicción de ° Brix.

Las características seleccionadas con un umbral igual a 52, se presentan en la Figura 4.13, en la que se resaltan en bandas verticales sombreadas, las características más relevantes para el modelo de Sacarosa.

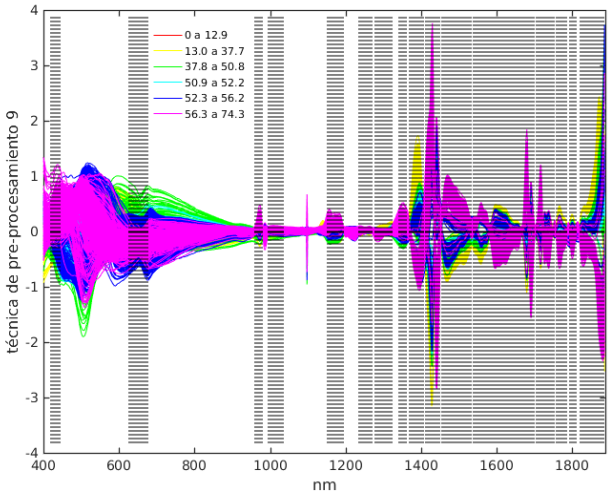


Figura 4.14. Bandas espectrales (características) seleccionados para construir el modelo de predicción de Sacarosa.

4.2.2 Optimización de los parámetros C y *γ* de la SVR

Aplicando la técnica de pre-procesamiento y las longitudes de onda (características) seleccionadas, fueron evaluados los modelos de calibración a fin de encontrar la combinación óptima de los parámetros C y *γ*, para un valor fijo de *ε* igual a 0.1. El método de *grid search* (Koch et al., 2012; Ma et al., 2015) en escala logarítmica fue aplicado con el parámetro C fue evaluado en el rango de 10^1 a 10^6 y *γ* en el rango de 10^{-1} a 10^{-3}, siendo el intervalo de evaluación del exponente de 0.25.

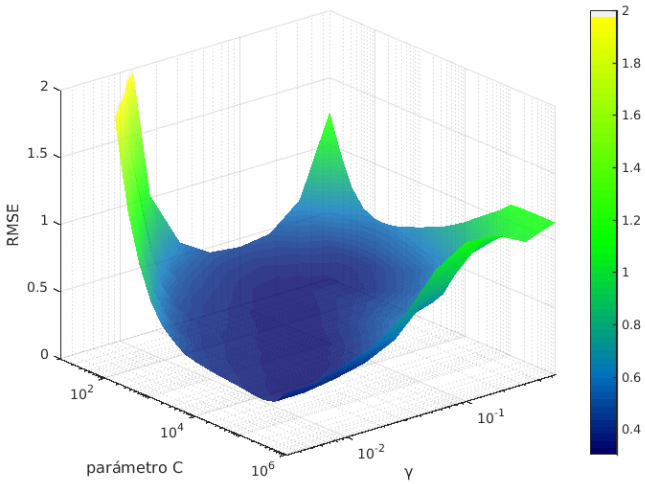


Figura 4.15. Gráfico de superficie del RMSE en validación cruzada del modelo de predicción de ° Brix, según los valores de los parámetros C y *γ* en todos los pasos de proceso.

La Figura 4.15 muestra un gráfico de superficie en el que se puede observar los valores de RMSE que arroja el modelo global de ° Brix para las diferentes combinaciones del parámetro C y *γ*. Los valores óptimos para ambos parámetros son $10^{2.75}$ y $10^{-1.25}$ respectivamente. Esto se debe a que se trata de los valores que en conjunto generan el RMSE menor evaluado en el promedio de los resultados de test de la técnica de CV con repetición.

La Figura 4.16, detalla el desempeño del modelo global en cada paso del proceso, y se confirma que los valores de los parámetros óptimos son los planteados anteriormente.

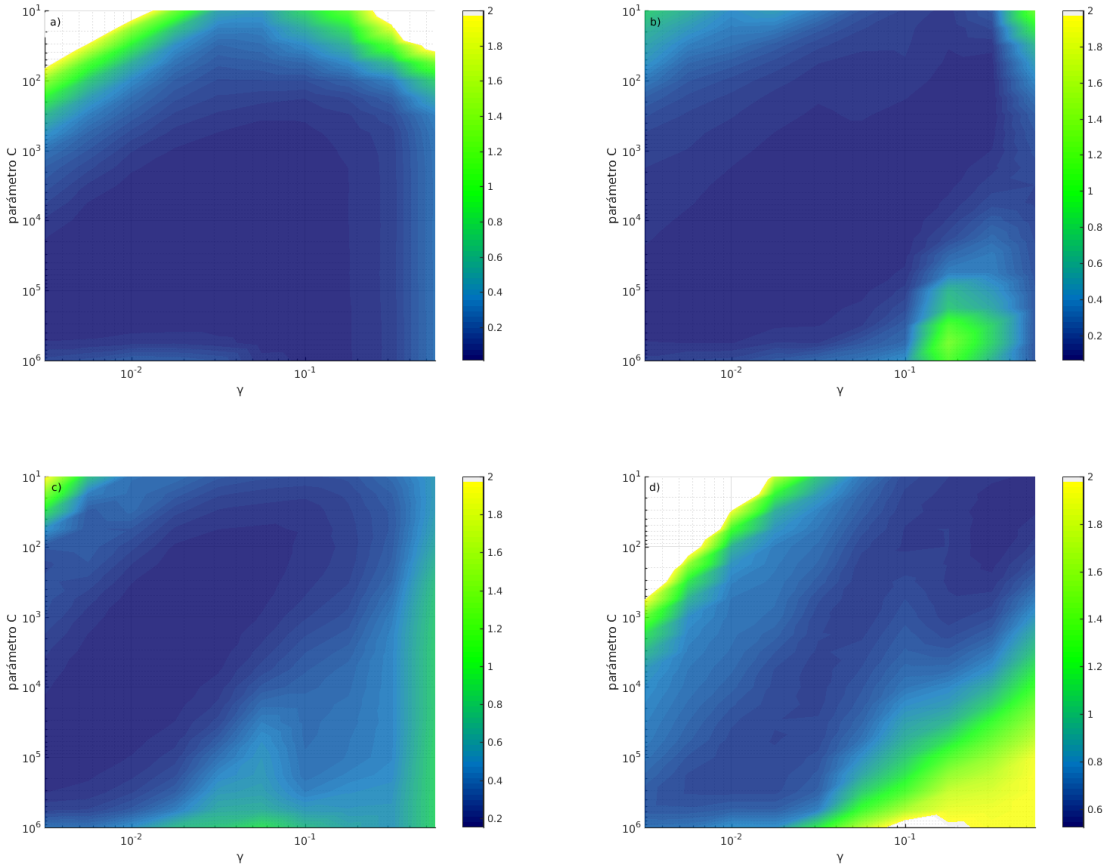


Figura 4.16. Mapas de calor del RMSE en validación cruzada del modelo de predicción de ° Brix, según los valores de los parámetros C y *γ* en los 4 pasos de proceso: a) jugo, b) jarabe, c) masa cocida, d) melaza.

La Figura 4.17 muestra un gráfico de superficie en el que se puede observar los valores de RMSE que arroja el modelo global de Sacarosa para las diferentes combinaciones del parámetro C y *γ*. Los parámetros óptimos son 10^{3} y $10^{-1.75}$ respectivamente, ya que son los que en conjunto generan el RMSE menor evaluado en el promedio de los resultados de test de la técnica de CV con repetición.

La Figura 4.18, detalla el desempeño del modelo global en cada paso del proceso, y se confirma que los valores de los parámetros óptimos son los planteados anteriormente.

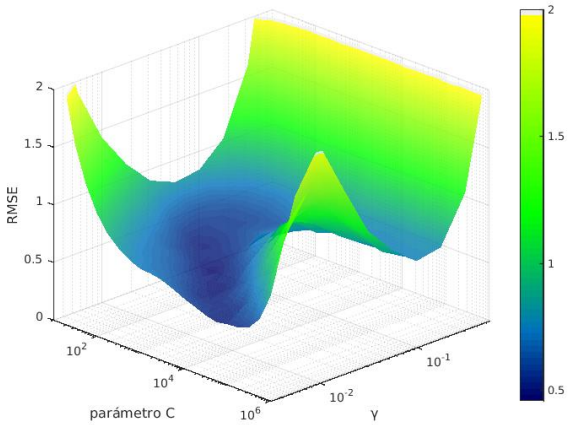


Figura 4.17: Gráfico de superficie del RMSE en validación cruzada del modelo de predicción de Sacarosa según los valores de los parámetros C y *γ* en todos los pasos de proceso.

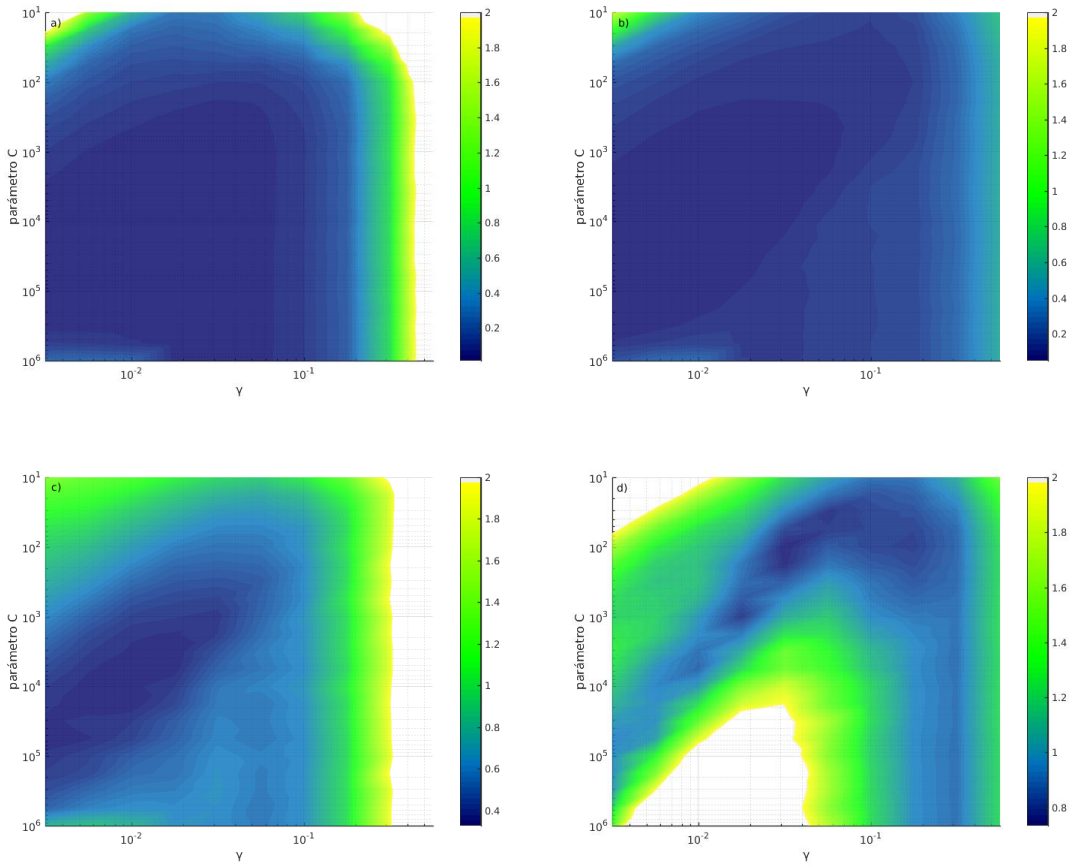


Figura 4.18. Mapas de calor del RMSE en validación cruzada del modelo de predicción de Sacarosa según los valores de los parámetros C y *γ* en los 4 pasos de proceso: a) jugo, b) jarabe, c) masa cocida, d) melaza.

4.2.3 Optimización del parámetro *ε*

Una vez que se ha seleccionado la técnica de pre-procesamiento, las longitudes de onda (características) y la combinación óptima de los parámetros C y *γ*, fue evaluado el parámetro *ε* de la SVR, en el rango de 0 a 1 en escala lineal, el intervalo de evaluación fue de 0.01.

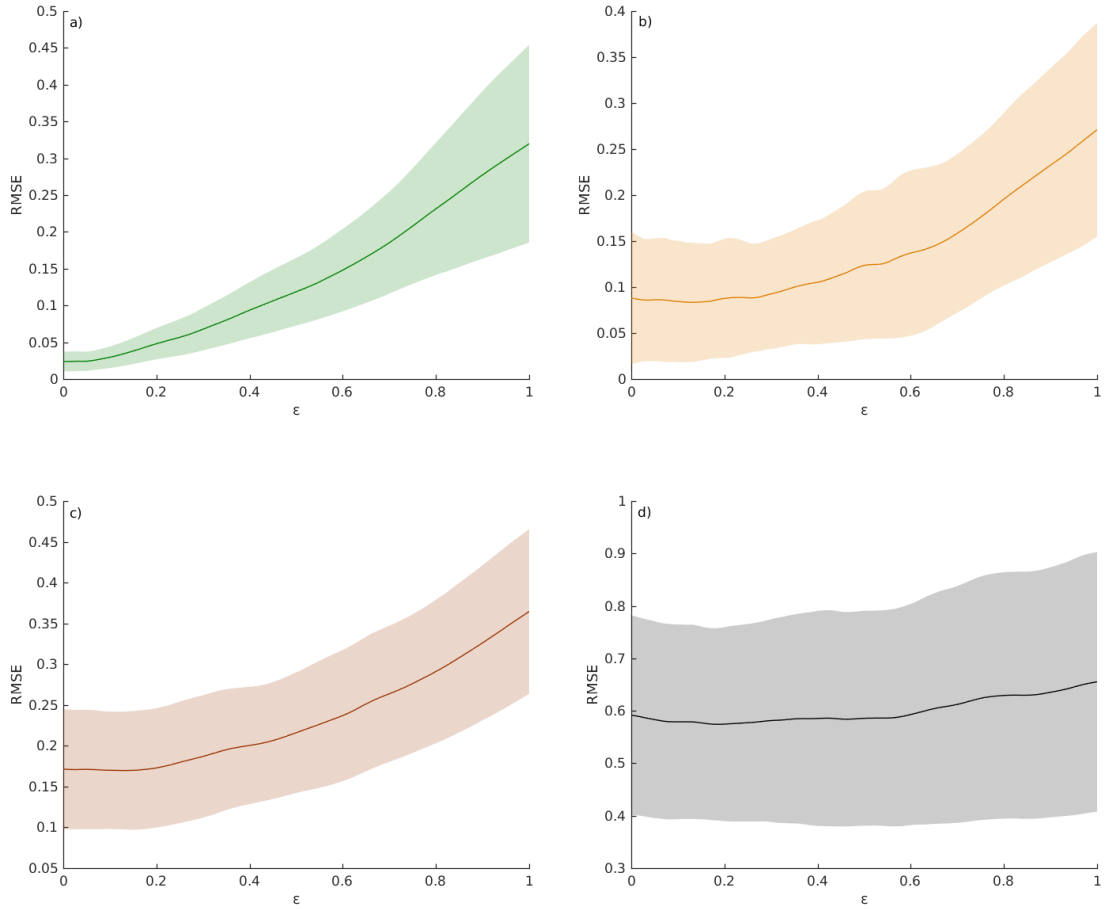


Figura 4.19. Curvas y bandas de una desviación estándar del RMSE en validación cruzada del modelo de predicción de ° Brix según el valor de ε en los 4 pasos de proceso.

La Figura 4.19 muestra las curvas para la optimización del parámetro *ε* en el modelo de predicción de ° Brix, se puede evidenciar que un valor de ε igual a 0.16 optimiza el RMSE en los cuatro pasos de proceso: a) jugo, b) jarabe, c) masa cocida, d) melaza.

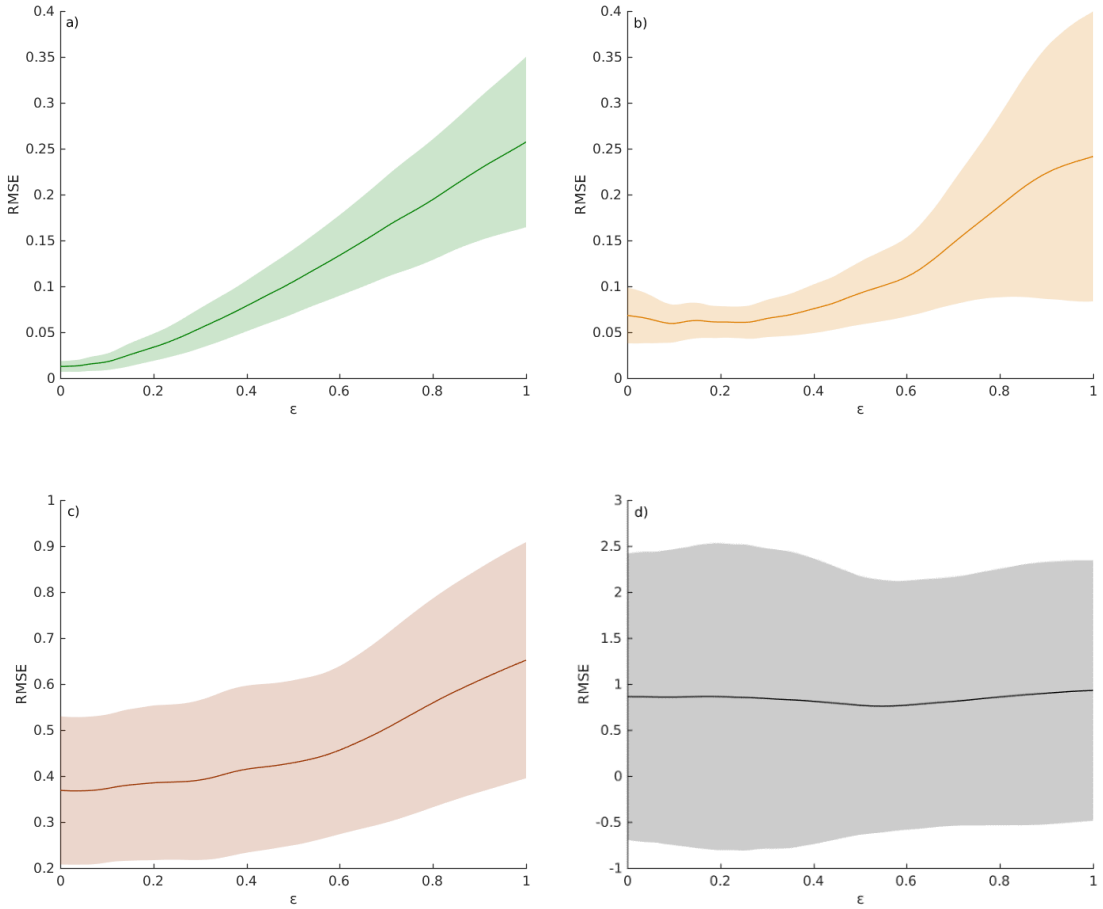


Figura 4.20. Curvas y bandas de una desviación estándar del RMSE en validación cruzada del modelo de predicción de Sacarosa según el valor de ε en los 4 pasos de proceso: a) jugo, b) jarabe, c) masa cocida, d) melaza.

La Figura 4.20 muestra las curvas para la optimización del parámetro ***ε*** en el modelo de predicción de Sacarosa, se evidencian que el valor de ***ε*** óptimo es 0.07. Sin embargo, en el caso de la melaza el valor óptimo de ***ε*** es 0.51, para la discusión se evaluará los modelos optimizados con ambos valores a fin de considerar una mejora en la estimación de Sacarosa en el paso del proceso melaza.

Finalmente, para evaluar estadísticamente el desempeño de los modelos cuyos parámetros fueron optimizados, se calculó *R-squared*, *R-squared* ajustado y *p-value*. El modelo para ° Brix se analizó con sus parámetros óptimos, en el caso de Sacarosa, el modelo se analizó con la alternativa de ***ε*** igual a 0.07.

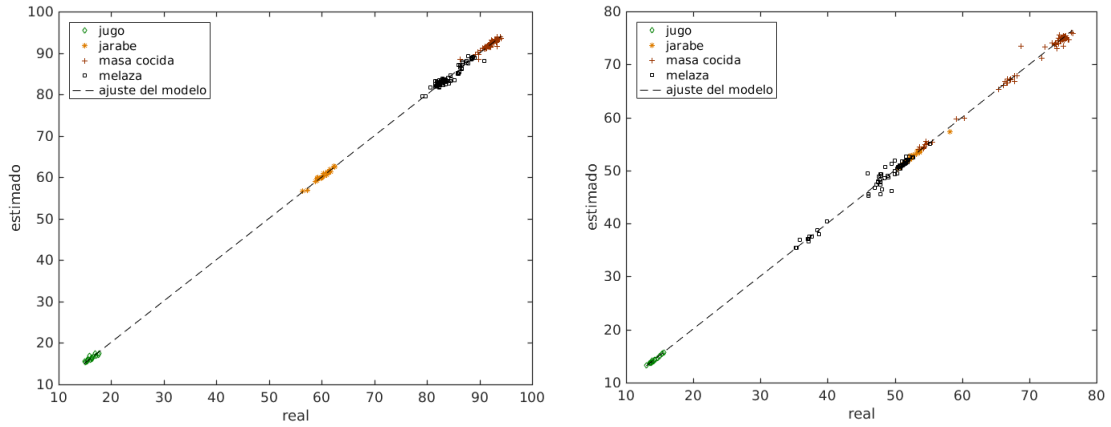


Figura 4.21. Gráficos de regresión (real *versus* estimado) en el conjunto *test* de la validación cruzada, de los modelos globales de Brix y Sacarosa.

La Figura 4.21 muestra dos gráficas de regresión (real *versus* estimado), el modelo de calibración para ° Brix, obtuvo un *R-squared* de 0.999, *R-squared* ajustado de 0.999 y un *p-value* de 1.77^{-285}. Mientras tanto que el modelo de calibración para Sacarosa obtuvo un *R-squared* de 0.998, *R-squared* ajustado de 0.998 y un *p-value* de 4.21^{-245}. Para ambos modelos se determina que existe una alta correlación estadísticamente significativa entre los valores reales y los valores predichos, lo que consolida la importancia de este hallazgo.

Los resultados fueron comparados con los publicados por Tange et al (2015), quienes utilizaron SVR para realizar el modelo de calibración de ° Brix y Sacarosa en el proceso industrial de la caña de azúcar. Los resultados aquí presentados, arrojan valores de RMSE inferiores a los de Tange et al (2015), que representan estimaciones más precisas de los parámetros de calidad. Esto se explica por cuanto en el presente trabajo fueron empleadas técnicas de pre-procesamiento y de selección de características relevantes. Por lo tanto, se eliminó el ruido en los espectros, así como las longitudes de onda que no aportan significativamente al modelo.

Rinnan et al., (2009) plantean que la realización de varias etapas de pre-procesamiento, no es aconsejable en espectros NIR, sin embargo los resultados del presente trabajo evidencian una mejora con la técnica de pre-procesamiento 9. Esta técnica consiste en calcular la primera derivada espectral, luego aplicar SNV y finalmente extraer la tendencia. Otros trabajos recientes en espectroscopía NIR han combinado varias técnicas de pre-procesamiento con muy buenos resultados (Hadad, Ra, y Elkhoudarya, 2015; Martelo-Vidal y Vázquez, 2014; Pan et al., 2015; Xie et al., 2015; Xu et al., 2008). Xu et al (2008) plantean que al combinar técnicas de pre-procesamiento se toma ventaja de la información complementaria lo que mejora la estabilidad de los modelos y los resultados en términos de RMSE.

En el modelo propuesto, la técnica de selección de características, permitió reconocer las longitudes de onda más relevantes, eliminar el ruido que generan las otras longitudes de onda y obtener estimaciones más precisas. Este modelo es consistente con otras investigaciones en las que se ha utilizado una técnica de umbral de selección de características para eliminar las variables irrelevantes con lo que han demostrado su

efectividad en la obtención de resultados más precisos (Christin et al., 2013; Erho et al., 2013; I. Guyon et al., 2008; Jie et al., 2015; Zhu et al., 2014).

La optimización de los parámetros de la SVR, demostró ser importante a efectos de lograr los RMSE mínimos del modelo. En el presente caso se realizó utilizando una técnica de search grid lo que permitió conocer la combinación del parámetro C, γ y ε óptimos, lo cual es coherente con lo planteado por Jeng (Jeng, 2006) y por Devos et al (2009). Estos afirman que los valores combinados de los parámetros de las SVM determinan la complejidad de los límites y por lo tanto el desempeño del modelo.

En este sentido se puede inferir que el modelo propuesto es preciso y estable, esto debido a la optimización de parámetros lo que es consistente con lo planteado por Cristianini y Shawe (2000) y por Devos et al (2009) quienes afirman que el ajuste de los parámetros de *kernel* SVM controla la complejidad de la hipótesis resultante y evita el sobreajuste del modelo.

La evaluación de los modelos fue realizada aplicando la técnica de técnica de CV con repetición. Esto de acuerdo a Garcia y Filzmoser (2015) posibilita un método adecuado para escoger el mejor modelo al analizar la media y desviación estándar de los resultados de las repeticiones. Estos resultados corresponden al conjunto de datos de test, es decir son datos que no fueron utilizados para la calibración, lo que permite estimar cómo se comportaría a futuro el modelo con datos nuevos (Kuhn y Johnson, 2013; Mitchell, 2014).

La Tabla 4.8 muestra los resultados del modelo global propuesto para ° Brix comparándolo con los referidos por Tange et al (2015), se aprecia que el modelo global propuesto tiene un RMSE de 0.305 mientras que el modelo global publicado llega a 0.59. Esto evidencia la optimización del modelo propuesto respecto al publicado. Lo mismo sucede en los cuatro pasos del proceso, en los que el modelo global propuesto mejora al modelo global publicado, y también a los cuatro modelos locales (individuales) publicados.

Tabla 4.8. Resultados comparativos de RMSE de los modelos referidos* por Tange et al (2015) con el modelo optimizado de predicción de ° Brix propuesto por el autor.

Modelo	Jugo	Jarabe	Masa cocida	Melaza	Global
Referencia *	0.1	0.2	0.5	0.5	0.5
SVM Local *	0.08	0.22	0.39	0.75	-
SVM Global *	0.16	0.25	0.47	0.79	0.59
SVM Optimizada	0.040±0.018	0.084±0.063	0.1702±0.073	0.576±0.183	0.305±0.076

La Tabla 4.9 muestra los resultados del modelo global propuesto para Sacarosa, con dos alternativas de valores que puede tomar ε, La SVM1 Optimizada, con un valor de 0.07 y la SVM 2 Optimizada, con un valor de 0.51, los que se comparan con los publicados por Tange et al (2015). El modelo global propuesto tiene un RMSE de 0.486 (SVM1 Optimizada) y 0.485 (SVM2 Optimizada) mientras que el modelo global publicado llega a 0.64, lo que evidencia una optimización del modelo con ambos valores de ε.

El modelo para Sacarosa propuesto supera al publicado, en tres de los cuatro pasos del proceso. En el paso del proceso Melaza, el modelo propuesto se aproxima al publicado, sin embargo no lo supera, esto se debe a que en el conjunto de espectros de Melaza se encontró valores con absorbancia mayor a dos, los que de acuerdo a Tange et at (2015) en su publicación fueron eliminados de la base de datos original, sin embargo a criterio del autor, estos espectros no son considerados *outliers*, por lo que en la presente investigación se mantuvieron para dar mayor robustez al modelo.

Tabla 4.9. Resultados comparativos de RMSE de los modelos referidos* por Tange et al (2015) con los modelos optimizados de predicción de Sacarosa propuestos por el autor.

Modelo	Jugo	Jarabe	Masa cocida	Melaza	Global
Referencia *	0.1	0.2	0.5	0.5	0.5
SVM Local *	0.11	0.22	0.56	0.62	-
SVM Global *	0.20	0.24	0.72	0.72	0.64
SVM 1 Optimizada	0.016±0.008	0.062±0.023	0.369±0.161	0.858±1.586	0.485±0.631
SVM 2 Optimizada	0.108±0.036	0.094±0.035	0.431± 0.180	0.765±1.396	0.486±0.559

Capítulo:
V. CONCLUSIONES Y FUTUROS DESARROLLOS

En este capítulo se presenta las principales conclusiones a las que se llegó durante el desarrollo de la Tesis Doctoral. Además, se resalta la contribución de esta investigación a la aplicación del conocimiento científico de las técnicas de ML en el mejoramiento de los procesos productivos en el sector agropecuario. Finalmente, se sugieren varias líneas de investigación para trabajos futuros, así como una reflexión sobre las potencialidades de la aplicación de estas técnicas en el análisis de otros tipos de datos derivados del sector agropecuario.

5.1 Conclusiones

El desarrollo de este trabajo, permite llegar a la conclusión de que el uso de técnicas de ML en el sector agropecuario, es aplicable para el apoyo en la toma las decisiones basadas en el procesamiento de la información disponible.

El uso de estas técnicas representa un potencial de aplicación en los diversos procesos que actualmente se realizan de forma manual. De esta manera las empresas del sector agropecuario, pueden mejorar su competitividad, al reducir los costos de producción, maximizar sus indicadores productivos y evitar pérdidas previsibles.

No obstante, es necesario que, en los programas de formación de los profesionales del sector agropecuario, se amplíe el desarrollo de competencias en análisis de datos, ya que las publicaciones existentes sobre la aplicación de técnicas de ML en el ámbito del mejoramiento de la productividad agropecuaria, son limitados.

Por lo expuesto anteriormente, se considera que, aún queda un amplio campo por explorar en el ámbito de la aplicación de técnicas de ML en datos del sector agropecuario. Las publicaciones más recientes, dan lugar a pensar que en los próximos años esta temática, se convertirá en un *hot-topic* entre los científicos del área.

En el presente trabajo se realizó el estudio de aplicabilidad de las técnicas de ML en dos bases de datos agropecuarias. Una de ellas relacionada con el ámbito de la producción de huevos comerciales, y la otra relacionada con el ámbito del proceso de control en la

industria azucarera. Sobre estas bases de datos se aplicó varias técnicas y metodologías, cuyos resultados fueron altamente satisfactorios.

En el caso de la aplicación en la industria avícola, la metodología que fue empleada, permitió encontrar los parámetros óptimos de los modelos de SVM y ANN. Se puede afirmar que, con cualquiera de las dos técnicas, es posible desarrollar un sistema de alarma, e incluso un sistema predictivo de las caídas de producción en ponedoras comercial.

Al analizar comparativamente ambos modelos, la principal diferencia del modelo SVM fue su entrenamiento con características extraídas manualmente por un especialista, mientras que el modelo de ANN fue entrenado usando una técnica de selección automática de características. Dicha extracción automática de características, representó una ventaja ya que no se requirió disponer del tiempo de un especialista para definirlas. En ambos modelos, se encontró que muchos de los falsos positivos eran en realidad días previos al evento, que no fueron etiquetados como problema porque aparentemente no existía una reducción significativa que haya sido observada por los expertos. Sin embargo, los modelos desarrollados fueron capaces de detectar de forma anticipada estos patrones.

La alerta temprana de caídas de producción, al ser utilizadas en sistemas informáticos de gestión de la producción avícola, generaría un gran impacto positivo en la industria avícola, ya que les confiere a estos sistemas, la capacidad de detectar y proponer soluciones oportunas a las causas del problema, mitigando el impacto de las pérdidas económicas asociadas a caídas de la producción.

A nivel de finca, un pronóstico con un día de antelación, podría resultar útil para la inspección diagnóstica en finca en busca de síntomas clínicos, u otros hallazgos para la toma de medidas tendientes a la solución inmediata del problema. Esto mejoraría la capacidad preventiva en el sistema de producción avícola, brindando monitorización asistida de manera automática como complemento a la observación humana, lo que resulta especialmente útil, al manejar altas poblaciones de animales.

Por otro lado, en cuanto a la aplicación en una industria de caña de azúcar, los modelos para la estimación de ° Brix y % de Sacarosa que son propuestos en esta Tesis Doctoral, mejoran a los resultados publicados previamente. Esta mejora se aprecia en los cuatro pasos del proceso, con excepción de la predicción del % de Sacarosa en las melazas, la cual es similar al modelo publicado, a pesar de que en el presente trabajo ningún espectro fue separado de la base de datos original.

El modelo de SVR propuesto fue capaz de estimar de mejor manera las no linealidades causadas por la combinación de los espectros NIR de múltiples diferentes etapas del proceso industrial de la caña de azúcar. Además, la selección de características redujo el número de longitudes de onda seleccionadas para la calibración de los modelos, lo que simplifica el modelo final, con la consecuente reducción de cálculos que se necesitan para estimar los parámetros de calidad en la industria azucarera.

El uso de modelos globales con un menor error permite a la industria azucarera tener una mejor estimación de los parámetros de calidad, con un único proceso de calibración

y por lo tanto un seguimiento más sencillo y efectivo del proceso. Contar con una metodología que permite un control de calidad con mayor precisión, abre la puerta hacia la detección de sustancias que se encuentran en menor concentración utilizando espectroscopía NIR.

Queda demostrado, por lo tanto, que la aplicación de técnicas de inteligencia artificial ofrece un importante medio de apoyo a la toma de decisiones en el sector agropecuario. Se considera que esta aplicación tiene un gran futuro en la gestión de la producción, al mejorar significativamente la eficiencia, en las tareas que normalmente son realizadas por expertos. Esta ganancia en eficiencia, además de reducir los costos de producción, incrementa las ganancias económicas, por lo que se espera un amplio nivel de adopción por parte de las empresas del sector agropecuario.

5.2 Futuros desarrollos

Los trabajos a futuro se enfocan en la aplicación de estas técnicas en otras aplicaciones del sector agropecuario, resulta de particular interés el uso de técnicas de ML para el análisis de datos más complejos como son los provenientes de sensores *wearables*, de imágenes, sonidos o videos. En estos conjuntos de datos se puede encontrar información valiosa que permite encontrar patrones para la detección de anomalías, diagnóstico de enfermedades, y el apoyo a la toma de decisiones agropecuarias.

Estas nuevas fuentes de datos son de mayor complejidad y abren la oportunidad de aplicación de las nuevas técnicas de inteligencia artificial como son las redes neuronales profundas. A corto y mediano plazo, el desarrollo de la investigación se orientará en esa dirección, y también en la búsqueda de nuevas técnicas de selección automática de características, con la finalidad de que tener algoritmos más robustos.

En cuanto a las aplicaciones en avicultura, los próximos trabajos se orientan hacia la incorporación de algoritmos de ML en los sistemas de gestión de la producción avícola, lo que supone un gran avance en materia de soporte inteligente a la toma de decisiones.

Otro campo interesante en este ámbito es la detección temprana de enfermedades específicas, lo que mejoraría la brecha entre el momento de la detección y el tratamiento. El resultando sería unas perdidas menores de producción y menos diseminación de la enfermedad. Para esto, la base de datos debe contener más información como, por ejemplo, el momento de la recolección de huevos, el consumo diario de agua y alimentos, patrones de sonidos e imágenes multiespectrales de las aves.

Respecto a la aplicación en la industria de la caña de azúcar, el trabajo a futuro se centra en la simplificación de los modelos mediante la identificación de respuestas redundantes a distintas longitudes de onda. Este aspecto es importante ya que es posible lograr modelos más simples con mayor capacidad para la generalización.

En este ámbito, un área que llama mucho la atención es el uso de espectrómetros portátiles de bajo costo. Estos dispositivos pueden utilizar una metodología similar a la propuesta en este trabajo para determinar la composición bromatológica de los alimentos.

También resulta de especial interés el uso de imágenes hiperespectrales, que resultan de la combinación de varias fotografías capturadas con sensores a distintas longitudes de onda. Este tipo de dispositivos en la actualidad tienen un alto costo, sin embargo, en el futuro cercano podrán ser accesibles, con lo que su estudio denota un avance significativo con potencial de aplicación en el sector agropecuario.

Finalmente, es necesario que se realice una amplia difusión respecto de la potencialidad y las ventajas derivadas del uso de algoritmos de ML en aplicaciones del sector agropecuario. El objetivo debiera ser hacer crecer el ecosistema investigador mediante la involucración de la mayor cantidad de investigadores del área agropecuaria en el desarrollo de soluciones automatizadas basadas en técnicas de inteligencia artificial.

BIBLIOGRAFÍA

Abdi, H., & Williams, L. J. (2010). *Tukey*'s honestly significant difference (HSD) test. *Encyclopedia of Research Design. Thousand Oaks, CA: Sage*.

Acar, U. A., Hudson, B., Miller, G. L., & Phillips, T. (2008). SVR: Practical Engineering of a Fast 3D Meshing Algorithm*. En M. L. Brewer & D. Marcum (Eds.), *Proceedings of the 16th International Meshing Roundtable* (pp. 45-62). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75103-8_3

Ahmad, H. A. (2011). Egg production forecasting: Determining efficient modeling approaches. *The Journal of applied poultry research*, *20*(4), 463-473. https://doi.org/10.3382/japr.2010-00266

Ahmadi, H., & Golian, A. (2008). Neural Network Model for Egg Production Curve. *Journal of Animal and Veterinary Advances*. https://doi.org/javaa.2008.1168.1170

Altman, D. G., & Bland, J. M. (1994). Statistics Notes: Diagnostic tests 2: predictive values. *BMJ* , *309*(6947), 102-102. https://doi.org/10.1136/bmj.309.6947.102

Amiri, A., Niaki, S. T. A., & Moghadam, A. T. (2014). A probabilistic artificial neural network-based procedure for variance change point estimation. *Soft Computing*, *19*(3), 691-700. https://doi.org/10.1007/s00500-014-1293-x

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, *8*(1), 69-80. https://doi.org/10.1016/0169-2070(92)90008-W

Aydin, A. (2016). Automated Monitoring Systems to Improve Health and Welfare of Broiler Chickens by Image and Sound Technology. Recuperado a partir de https://lirias.kuleuven.be/handle/123456789/518765

Azose, J. J., Ševčíková, H., & Raftery, A. E. (2016). Probabilistic population projections with migration uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6460-6465. https://doi.org/10.1073/pnas.1606119113

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, *43*(5), 772-777.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203-224.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1-127. https://doi.org/10.1561/2200000006

Benítez, R., Escudero, G., & Kanaan, S. (2013). *Inteligencia artificial avanzada*. España: Editorial UOC.

Bennett, *K*. P., & Campbell, C. (2000). Support vector machines. *ACM SIGKDD Explorations Newsletter*, *2*(2), 1-13. https://doi.org/10.1145/380995.380999

Bertran, E., Blanco, M., Maspoch, S., Ortiz, M. C., Sánchez, M. S., & Sarabia, L. A. (1999). Handling intrinsic non-linearity in near-infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *49*(2), 215-224. https://doi.org/10.1016/S0169-7439(99)00043-X

Bhanot, G., Alexe, G., Venkataraghavan, B., & Levine, A. J. (2006). A robust meta-classification strategy for cancer detection from MS data. *Proteomics*, *6*(2), 592-604.

Bielza, C., Barreiro, P., Rodríguez-Galiano, M. I., & Martín, J. (2003). Logistic regression for simulating damage occurrence on a fruit grading line. *Computers and Electronics in Agriculture*, *39*(2), 95-113. https://doi.org/10.1016/S0168-1699(03)00021-8

Blagus, R., & Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *11*, 523. https://doi.org/10.1186/1471-2105-11-523

Blanco, M., & Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. *Trends in analytical chemistry: TRAC*, *21*(4), 240-250. https://doi.org/10.1016/S0165-9936(02)00404-1

Brereton, R. G. (2015). Pattern recognition in chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *149, Part B*, 90-96. https://doi.org/10.1016/j.chemolab.2015.06.012

Cameron, A. (2012). *Manual of Basic Animal Disease Surveillance* (p. 110). Interafrican Bureau for Animal Resources.

Cen, H., & He, Y. (2007/2). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, *18*(2), 72-83. https://doi.org/10.1016/j.tifs.2006.09.003

Chen, H. Z., Wen, J. B., Chen, J. C., Li, L. H., & Huo, Y. J. (2014). Near-infrared spectroscopic modeling optimization for quantitative determination of sugar brix in sugarcane initial-pressure jugo. *Int. J. Tech. Res. Applic*, *2*, 6.

Christin, C., Hoefsloot, H. C. J., Smilde, A. *K*., Hoekman, B., Suits, F., Bischoff, R., & Horvatovich, P. (2013). A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics: MCP*, *12*(1), 263-276. https://doi.org/10.1074/mcp.M112.022566

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines (Cambridge. Cambridge Univ. Press.

Dai, X., & Gao, Z. (2013). From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis. *IEEE Transactions on Industrial Informatics*, *9*(4), 2226-2238. https://doi.org/10.1109/TII.2013.2243743

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., & Huvenne, J.-P. (2009). Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems*, *96*(1), 27-33. https://doi.org/10.1016/j.chemolab.2008.11.005

De Vries, A., & Reneau, J. *K*. (2010). Application of statistical process control charts to monitor changes in animal production systems. *Journal of animal science*, *88*(13 elect suppl), E11-E24. https://doi.org/10.2527/jas.2009-2622

Díez, J., Bahamonde, A., Alonso, J., López, S., Del Coz, J. J., Quevedo, J. R., … Goyache, F. (2003). Artificial intelligence techniques point out differences in classification performance between light and standard bovine carcasses. *Meat Science*, *64*(3), 249-258. https://doi.org/10.1016/S0309-1740(02)00185-7

Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of food engineering*, *75*(2), 196-204. https://doi.org/10.1016/j.jfoodeng.2005.03.056

Dos Santos, C. A. T., Lopo, M., Páscoa, R. N. M. J., & Lopes, J. A. (2013). A review on the

applications of portable near-infrared spectrometers in the agro-food industry. *Applied spectroscopy*, *67*(11), 1215-1233. https://doi.org/10.1366/13-07228

Elkan, C. (2001). The foundations of cost-sensitive learning. En *International joint conference on artificial intelligence* (Vol. 17, pp. 973-978). LAWRENCE ERLBAUM ASSOCIATES LTD.

El-Korany, A., El-Azhary, E., & Yehia, M. (2004). An Approach for building generic diagnosis model in agricultural domain. En *Proceedings of the 5th International Workshop on Artificial Intelligent in Agricultural, Mar* (pp. 8-10). claes.sci.eg.

Erenturk, *K*., Erenturk, S., & Tabil, L. G. (2004). A comparative study for the estimation of dynamical drying behavior of Echinacea angustifolia: regression analysis and neural network. *Computers and Electronics in Agriculture*, *45*(1), 71-90. https://doi.org/10.1016/j.compag.2004.06.002

Erho, N., Crisan, A., Vergara, I. A., Mitra, A. P., Ghadessi, M., Buerki, C., … Jenkins, R. B. (2013). Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PloS One*, *8*(6), e66855. https://doi.org/10.1371/journal.pone.0066855

Farkas, I., Reményi, P., & Biró, A. (2000). A neural network topology for modelling grain drying. *Computers and Electronics in Agriculture*, *26*(2), 147-158. https://doi.org/10.1016/S0168-1699(00)00068-5

Fawcett, T. (2006/6). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Felipe, V. P. S., Silva, M. A., Valente, B. D., & Rosa, G. J. M. (2015). Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. *Poultry science*, *94*(4), 772-780. https://doi.org/10.3382/ps/pev031

Fernandez-Lozano, C., Fernández-Blanco, E., Dave, *K*., Pedreira, N., Gestal, M., Dorado, J., & Munteanu, C. R. (2014). Improving enzyme regulatory protein classification by means of SVM-RFE feature selection. *Molecular bioSystems*, *10*(5), 1063-1071. https://doi.org/10.1039/c3mb70489k

Fernández Pierna, J. A., Baeten, V., & Dardenne, P. (2006). Screening of compound feeds using NIR hyperspectral data. *Chemometrics and Intelligent Laboratory Systems*, *84*(1-2), 114-118. https://doi.org/10.1016/j.chemolab.2006.03.012

Flanders, F., & Gillespie, J. R. (2015). *Modern Livestock & Poultry Production*. Cengage Learning.

Florkowski, W. J., Prussia, S. E., Shewfelt, R. L., & Brueckner, B. (2009). *Postharvest Handling: A Systems Approach*. Elsevier Science.

Frank, R. J., Davey, N., & Hunt, S. P. (2001). Time Series Prediction and Neural Networks. *Journal of Intelligent and Robotic Systems*, *31*(1-3), 91-103. https://doi.org/10.1023/A:1012074215150

Frost, A. R., Schofield, C. P., Beaulah, S. A., Mottram, T. T., Lines, J. A., & Wathes, C. M. (1997). A review of livestock monitoring and the need for integrated systems. *Computers and Electronics in Agriculture*, *17*(2), 139-159.

Garcia, H., & Filzmoser, P. (2015). Multivariate Statistical Analysis using the R package chemometrics. *Vienna, Austria*.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14–15), 2627-2636. https://doi.org/10.1016/S1352-2310(97)00447-0

Gates, M. C., Holmstrom, L. *K*., Biggers, *K*. E., & Beckham, T. R. (2015). Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities. *Frontiers in Public Health*, *3*, 74. https://doi.org/10.3389/fpubh.2015.00074

Goel, P. *K*., Prasher, S. O., Patel, R. M., Landry, J. A., Bonnell, R. B., & Viau, A. A. (2003). Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn. *Computers and Electronics in Agriculture*, *39*(2), 67-93. https://doi.org/10.1016/S0168-1699(03)00020-6

Grossman, M., & Koops, W. J. (2001). A Model for Individual Egg Production in Chickens. *Poultry science*, *80*(7), 859-867. https://doi.org/10.1093/ps/80.7.859

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*.

Guo, L., Rivero, D., & Pazos, A. (2010). Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks. *Journal of Neuroscience Methods*, *193*(1), 156-163. https://doi.org/10.1016/j.jneumeth.2010.08.030

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of machine learning research: JMLR*, *3*, 1157-1182.

Guyon, I., & Elisseeff, A. (2006). An Introduction to Feature Extraction. En I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Feature Extraction* (pp. 1-25). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_1

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). *Feature Extraction: Foundations and Applications*. Springer Berlin Heidelberg.

Hadad, G. M., Ra, A. S., & Elkhoudarya, M. M. (2015). Simultaneous Determination of Clarithromycin, Tinidazole and Omeprazole in Helicure Tablets Using Reflectance Near-Infrared Spectroscopy with the Aid of Chemometry. *Pharmaceutica Analytica Acta*, *2015*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Vol. 18, p. 746). New York, NY: Springer New York. https://doi.org/10.1007/b94608

He, H.-J., Wu, D., & Sun, D.-W. (2014). Rapid and non-destructive determination of drip loss and pH distribution in farmed Atlantic salmon (Salmo salar) fillets using visible and near-infrared (Vis–NIR) hyperspectral imaging. *Food chemistry*, *156*, 394-401. https://doi.org/10.1016/j.foodchem.2014.01.118

Hempstalk, *K*., McParland, S., & Berry, D. P. (2015). Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *Journal of dairy science*, *98*(8), 5262-5273. https://doi.org/10.3168/jds.2014-8984

Henry, R., & Kettlewell, P. (2012). *Cereal Grain Quality*. Springer Netherlands.

Hepworth, P. J., Nefedov, a. V., Muchnik, I. B., & Morgan, *K*. L. (2012). Broiler chickens can benefit from machine learning: support vector machine analysis of observational epidemiological data. *Journal of the Royal Society, Interface / the Royal Society*, *9*(73), 1934-1942. https://doi.org/10.1098/rsif.2011.0852

Herrera, F., Hervas, C., Otero, J., & Sánchez, L. (2004). Un estudio empırico preliminar sobre los tests estadısticos más habituales en el aprendizaje automático. *Tendencias de la Mineria de Datos en Espana, Red Espanola de Mineria de Datos y Aprendizaje (TIC2002-11124-E)*, 403-412.

Hornik, *K*., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks: the official journal of the International Neural Network*

Society, *2*(5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Hossain, M. A., Ayodele, B. V., Cheng, C. *K*., & Khan, M. R. (2016). Artificial neural network modeling of hydrogen-rich syngas production from methane dry reforming over novel Ni/CaFe 2 O 4 catalysts. *International journal of hydrogen energy*.

Huang, Y.-M., Hung, C.-M., & Jiau, H. C. (2009). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear analysis. Real world applications*, *7*(4), 720-747. https://doi.org/10.1016/j.nonrwa.2005.04.006

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast ACC. *International journal of forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jacob, J. P., Wilson, H. R., Miles, R. D., Butcher, G. D., & Mather, F. B. (2014). Factors Affecting Egg Production in Backyard Chicken. *University of Florida*, 1-8.

Jafari, P., & Azuaje, F. (2006). An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, *6*, 27. https://doi.org/10.1186/1472-6947-6-27

Jeng, J.-T. (2006). Hybrid approach of selecting hyperparameters of support vector machine for regression. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society*, *36*(3), 699-709. https://doi.org/10.1109/TSMCB.2005.861067

Jie, N.-F., Zhu, M.-H., Ma, X.-Y., Osuch, E. A., Wammes, M., Théberge, J., … Calhoun, V. D. (2015). Discriminating Bipolar Disorder From Major Depression Based on SVM-FoBa: Efficient Feature Selection With Multimodal Brain Imaging Data. *IEEE Transactions on Autonomous Mental Development*, *7*(4), 320-331. https://doi.org/10.1109/TAMD.2015.2440298

Jones, J. W. (1993). Decision support systems for agricultural development. *Systems approaches for agricultural development. Kluwer Academic Press, Boston*, 459-471.

Kalhor, T., Rajabipour, A., Akram, A., & Sharifi, M. (2016). Modeling of energy ratio index in broiler production units using artificial neural networks. *Sustainable Energy Technologies and Assessments*, *17*, 50-55. https://doi.org/10.1016/j.seta.2016.09.002

Kapoor, P., & Bedi, S. S. (2013). Weather Forecasting Using Sliding Window Algorithm. *International Scholarly Research Notices*, *2013*. https://doi.org/10.1155/2013/156540

Karayiannis, N., & Venetsanopoulos, A. N. (2013). *Artificial Neural Networks: Learning Algorithms, Performance Evaluation, and Applications*. Springer US.

Keleş, S., van der Laan, M., & Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics* , *18*(9), 1167-1175.

Koch, P., Bischl, B., Flasch, O., Bartz-Beielstein, T., Weihs, C., & Konen, W. (2012). Tuning and evolution of support vector *kernel*s. *Evolutionary Intelligence*, *5*(3), 153-170. https://doi.org/10.1007/s12065-012-0073-8

Kowalski, B. R. (1980). Chemometrics. *Analytical chemistry*, *52*(5), 112R-122R. https://doi.org/10.1021/ac50055a016

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., & Held, P. (2013). Multi-Layer Perceptrons. En *Computational Intelligence* (pp. 47-81). Springer London. https://doi.org/10.1007/978-1-4471-5013-8_5

Kuhn, M., & Johnson, *K*. (2013). *Applied Predictive Modeling:* Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kumaravelu, C., & Gopal, A. (2015). A review on the applications of Near-Infrared spectrometer and Chemometrics for the agro-food processing industries, 8-12. https://doi.org/10.1109/TIAR.2015.7358523

Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of chemometrics*, *6*(5), 267-281. https://doi.org/10.1002/cem.1180060506

Leiyua, W., Ruizhia, S., & Zhenlib, C. (2012). Study of Monitor and Early Warning System of Livestock Health Culture [J]. *Journal of Agricultural*.

Liu, H., & Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Data Mining*. Springer US.

Lokhorst, C., Fuchs, H., & Visscher, J. (1998). Introduction of the knowledge intensive system Lay-Monitor. *edepot.wur.nl*. Recuperado a partir de http://edepot.wur.nl/108969

Lokhorst, C., & Lamaker, E. J. J. (1996). An expert system for monitoring the daily production process in aviary systems for laying hens. *Computers and Electronics in Agriculture*, *15*(3), 215-231. https://doi.org/10.1016/0168-1699(96)00017-8

Long, A., & Wilcox, S. (2011). Optimizing Egg Revenue for Poultry Farmers, 1-10.

Luypaert, J., Heuerding, S., de Jong, S., & Massart, D. L. (2002). An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. *Journal of pharmaceutical and biomedical analysis*, *30*(3), 453-466.

Magwaza, L. S., Opara, U. L., Nieuwoudt, H., Cronje, P. J. R., Saeys, W., & Nicolaï, B. (2011). NIR Spectroscopy Applications for Internal and External Quality Analysis of Citrus Fruit—A Review. *Food and Bioprocess Technology*, *5*(2), 425-444. https://doi.org/10.1007/s11947-011-0697-1

Mariano, F., Lima, R. R., Alvarenga, R. R., Rodrigues, P. B., & Lacerda, W. S. (2014). Neural network committee to predict the AMEn of poultry feedstuffs. *Neural computing & applications*, *25*(7-8), 1903-1911.

Martelo-Vidal, M. J., & Vázquez, M. (2014). Evaluation of ultraviolet, visible, and near infrared spectroscopy for the analysis of wine compounds. *Czech Journal of Food Science*, *32*, 37.

Martens, D., & Baesens, B. (2010). Building Acceptable Classification Models (pp. 53-74). https://doi.org/10.1007/978-1-4419-1280-0_3

Martens, H., & Naes, T. (1992). *Multivariate Calibration*. Wiley.

Ma, X., Zhang, Y., & Wang, Y. (2015). Performance evaluation of *kernel* functions based on *grid search* for support vector regression. En *2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (pp. 283-288). ieeexplore.ieee.org. https://doi.org/10.1109/ICCIS.2015.7274635

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF02478259

Mench, J. A., van Tienhoven, A., Marsh, J. A., McCormick, C. C., Cunningham, D. L., & Baker, R. C. (1986). Effects of cage and floor pen management on behavior, production, and physiological stress responses of laying hens. *Poultry Science*, *65*(6), 1058-1069.

Mertens, *K*., Vaesen, I., Löffel, J., Kemps, B., Kamers, B., Zoons, J., … De Ketelaere, B. (2009). An intelligent control chart for monitoring of autocorrelated egg production process data based on a synergistic control strategy. *Computers and Electronics in Agriculture*, *69*(1),

100-111. https://doi.org/10.1016/j.compag.2009.07.012

Minsky, M. L., & Papert, S. (1969). *Perceptions: An Introduction to Computational Geomry*. MIT press.

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, *4*(5), 468-481. https://doi.org/10.1002/wcms.1183

Mollazade, *K*., Omid, M., & Arefi, A. (2012). Comparing data mining classifiers for grading raisins based on visual features. *Computers and Electronics in Agriculture*, *84*, 124-131. https://doi.org/10.1016/j.compag.2012.03.004

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks: the official journal of the International Neural Network Society*, *6*(4), 525-533. https://doi.org/10.1016/S0893-6080(05)80056-5

Moody, J., & Utans, J. (1994). Architecture selection strategies for neural networks: Application to corporate bond rating prediction. En *Neural networks in the capital markets* (pp. 277-300). Citeseer.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). *Data Mining in Agriculture* (Vol. 34, pp. 143-160-160). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-88615-2

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. En *Fundamental Issues of Artificial Intelligence* (pp. 555-572). Springer, Cham. https://doi.org/10.1007/978-3-319-26485-1_33

Narinc, D., Uckardes, F., & Aslan, E. (2014). Egg production curve analyses in poultry science. *World's poultry science journal*, *70*(04), 817-828.

Nürnberger, A., Pedrycz, W., & Kruse, R. (2002). Data mining tasks and methods: Classification: neural network approaches. En *Handbook of data mining and knowledge discovery* (pp. 304-317). Oxford University Press, Inc.

Ozaki, Y., McClure, W. F., & Christy, A. A. (2006). *Near-infrared spectroscopy in food science and technology*. John Wiley & Sons.

Palma, J., & Marín, R. (2013). *Inteligencia artificial. Técnicas, métodos y aplicaciones*. (J. L. Garcia Jurado, Ed.). Murcia: McGraw Hill.

Pan, X., Li, Y., Wu, Z., Zhang, Q., Zheng, Z., Shi, X., & Qiao, Y. (2015). A online NIR sensor for the pilot-scale extraction process in Fructus aurantii coupled with single and ensemble methods. *Sensors* , *15*(4), 8749-8763. https://doi.org/10.3390/s150408749

Pazzani, M., Merz, C., Murphy, P., Ali, *K*., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. En *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 217-225).

Pepe, M. S. (2003). *The Statistical Evaluation of Medical* Test*s for Classification and Prediction*. Oxford University Press.

Perini, A., & Susi, A. (2004/9). Developing a decision support system for integrated production in agriculture. *Environmental Modelling & Software*, *19*(9), 821-829.

Polanco, L. S., Day, D. F., Savoie, S., Bergeron, S., Charlet, T., & Legendre, B. L. (2014). Improvements of raw sugar quality using double purge of c-masa cocidas performance comparison. En *LSU AgCenter Audubon Sugar Institute Factory Operations Seminar* (p. 46).

Ponce, P. (2011). *Inteligencia Artificial con aplicaciones a la ingeniería. México DF: Editorial Alfaomega, 2011. 348 p*. ISBN 978-84-267-1706-1.

Prasoona, R. *K*., Jyoti, A., Mukesh, Y., Nishant, S., Anuraj, N. S., & Shobha, J. (2013). Optimization of Gaussian *Kernel* Function in Support Vector Machine aided QSAR studies of C-aryl glucoside SGLT2 inhibitors. *Interdisciplinary Sciences, Computational Life Sciences*, *5*(1), 45-52. https://doi.org/10.1007/s12539-013-0156-y

Quan, Y., Yang, J., Yao, L.-X., & Ye, C.-Z. (2004). Successive overrelaxation for support vector regression. *Journal of Software Maintenance and Evolution: Research and Practice*, *15*(2), 200-206.

Radionov, A. A., Antonov, L. V., Makarov, *K*. V., & Orlov, A. A. (2015). International Conference on Industrial Engineering (ICIE-2015)Development and Experimental Research on Production Data Analysis Algorithm in Livestock Enterprises. *Procedia Engineering*, *129*, 664-669. https://doi.org/10.1016/j.proeng.2015.12.088

Rady, A. M., & Guyer, D. E. (2015/5). Evaluation of sugar content in potatoes using NIR reflectance and wavelength selection techniques. *Postharvest biology and technology*, *103*, 17-26. https://doi.org/10.1016/j.postharvbio.2015.02.012

Raith, S., Vogel, E. P., Anees, N., Keul, C., Güth, J.-F., Edelhoff, D., & Fischer, H. (2017). Artificial Neural Networks as a powerful numerical tool to classify specific features of a tooth based on 3D scan data. *Computers in biology and medicine*, *80*, 65-76. https://doi.org/10.1016/j.compbiomed.2016.11.013

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). *Cross-validation*. En L. Liu & M. Tamer Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 532-538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Rinnan, Å., Berg, F. van D., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in analytical chemistry: TRAC*, *28*(10), 1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

Rivero, D., Fernandez-Blanco, E., Dorado, J., & Pazos, A. (2011). Using recurrent ANNs for the detection of epileptic seizures in EEG signals. En *2011 IEEE Congress of Evolutionary Computation (CEC)* (pp. 587-592). https://doi.org/10.1109/CEC.2011.5949672

Roussel, S., Bellon-Maurel, V., Roger, J.-M., & Grenier, P. (2003). Authenticating white grape must variety with classification models based on aroma sensors, FT-IR and UV spectrometry. *Journal of food engineering*, *60*(4), 407-419. https://doi.org/10.1016/S0260-8774(03)00064-5

Ruck, D. W., Rogers, S. *K*., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, *1*(4), 296-298. https://doi.org/10.1109/72.80266

Saeed, *K*., & Snášel, V. (2014). *Computer Information Systems and Industrial Management: 13th IFIP TC 8 International Conference, CISIM 2014, Ho Chi Minh City, Vietnam, November 5-7, 2014, Proceedings*. Springer.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* , *23*(19), 2507-2517. https://doi.org/10.1093/bioinformatics/btm344

Samborska, I. A., Alexandrov, V., Sieczko, L., Kornatowska, B., Goltsev, V., Cetner, M. D., & Kalaji, H. M. (2014). Artificial neural networks and their application in biological and agricultural research. *NanoPhotoBioSciences*, *2*, 14-30.

Schaefer, A. L., Cook, N., Tessaro, S. V., Deregt, D., Desroches, G., Dubeski, P. L., … Godson, D. L. (2004). Early detection and prediction of infection using infrared thermography. *Canadian journal of animal science*, *84*(1), 73-80. https://doi.org/10.4141/A02-104

Shahin, M. A., Tollner, E. W., & McClendon, R. W. (2001). AE—Automation and Emerging Technologies: Artificial Intelligence Classifiers for sorting Apples based on Watercore. *Journal of Agricultural Engineering Research*, *79*(3), 265-274. https://doi.org/10.1006/jaer.2001.0705

Shannon, C. E. (1950). XXII. Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *41*(314), 256-275. https://doi.org/10.1080/14786445008521796

Singh, P. *K*., Sarkar, R., & Nasipuri, M. (2015). Statistical validation of multiple classifiers over multiple datasets in the field of pattern recognition. *International Journal of Applied Pattern Recognition*, *2*(1), 1-23. https://doi.org/10.1504/IJAPR.2015.068929

Sorol, N., Arancibia, E., Bortolato, S. A., & Olivieri, A. C. (2010). Visible/near infrared-partial least-squares analysis of Brix in sugar cane jugo: A test field for variable selection methods. *Chemometrics and Intelligent Laboratory Systems*, *102*(2), 100-109. https://doi.org/10.1016/j.chemolab.2010.04.009

Stafford, J. V., & Werner, A. (2003). *Precision Agriculture*. Wageningen Academic Pub.

Stark, E. (1996). Near infrared spectroscopy past and future. *Near Infrared Spectroscopy The Future Waves*, 701-713.

Sun, Y., Wong, A. *K*. C., & Kamel, M. S. (2009). Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687-719. https://doi.org/10.1142/S0218001409007326

Suttorp, T., & Igel, C. (2007). *Artificial Neural Networks – ICANN 2007*. (J. M. de Sá, L. A. Alexandre, W. Duch, & D. Mandic, Eds.) (Vol. 4668, pp. 139-148). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74690-4

Szymańska, E., Gerretzen, J., Engel, J., Geurts, B., Blanchet, L., & Buydens, L. M. C. (2015). Chemometrics and qualitative analysis have a vibrant relationship. *Trends in analytical chemistry: TRAC*, *69*, 34-51. https://doi.org/10.1016/j.trac.2015.02.015

Tajammal Munir, M., Yu, W., Young, B. R., & Wilson, D. I. (2015/6). The current status of process analytical technologies in the dairy industry. *Trends in Food Science & Technology*, *43*(2), 205-218. https://doi.org/10.1016/j.tifs.2015.02.010

Tange, R. I., Rasmussen, M. A., Taira, E., & Bro, R. (2015). Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps. *Journal of Near Infrared Spectroscopy*.

Tavares, T. R., Oliveira, A. L. I., Cabral, G. G., Mattos, S. S., & Grigorio, R. (2013). Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children. En *Neural Networks (IJCNN), The 2013 International Joint Conference on* (pp. 1-8). ieeexplore.ieee.org. https://doi.org/10.1109/IJCNN.2013.6706947

Teye, E., Huang, X.-Y., & Afoakwa, N. (2013). Review on the potential use of near infrared spectroscopy (NIRS) for the measurement of chemical residues in food. *American Journal of Food Science and Technology*, *1*(1), 1-8.

Tomazzoli, M. M., Pai Neto, R. D., Moresco, R., Westphal, L., Zeggio, A. R. S., Specht, L., … Maraschin, M. (2015). Discrimination of Brazilian propolis according to the seasoning using chemometrics and machine learning based on UV-Vis scanning data. *Journal of Integrative Bioinformatics*, *12*(4), 279. https://doi.org/10.2390/biecoll-jib-2015-279

Torrione, P., Collins, L. M., & Morton, *K*. D. (2014). Multivariate analysis, chemometrics, and machine learning in laser spectroscopy. En *Elsevier Ltd.*

Valderrama, P., Braga, J. W. B., & Poppi, R. J. (2007). Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy. *Journal of the Brazilian Chemical Society*, *18*(2), 259-266. https://doi.org/10.1590/S0103-50532007000200003

Vannucci, M., & Colla, V. (2016). Smart Under-Sampling for the Detection of Rare Patterns in Unbalanced Datasets. En I. Czarnowski, A. M. Caballero, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent Decision Technologies 2016* (pp. 395-404). Springer International Publishing. https://doi.org/10.1007/978-3-319-39630-9_33

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. En M. I. Jordan & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9* (pp. 281-287). MIT Press.

Vapnik, V. N., & Kotz, S. (1982). *Estimation of dependences based on empirical data* (Vol. 41). Springer-Verlag New York.

Venkatesan, M., Thangavelu, A., & Prabhavathy, P. (2013). *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*. (J. C. Bansal, P. Singh, *K*. Deep, M. Pant, & A. Nagar, Eds.) (Vol. 202, pp. 277-288). India: Springer India. https://doi.org/10.1007/978-81-322-1041-2

Viscarra Rossel, R. A. (2008). ParLeS: Software for chemometric analysis of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, *90*(1), 72-83. https://doi.org/10.1016/j.chemolab.2007.06.006

Walker, D. H. (2002). Decision support, learning and rural resource management. *Agricultural systems*, *73*(1), 113-127.

Wang, L., Sun, D.-W., Pu, H., & Cheng, J.-H. (2016). Quality Analysis and Classification and Authentication of Liquid Foods by Near-infrared Spectroscopy: A Review of Recent Research Developments. *Critical Reviews in Food Science and Nutrition*, 0. https://doi.org/10.1080/10408398.2015.1115954

Wang, X., Ye, H.-J., Li, Q.-T., Xie, J.-C., Lu, J.-J., Xia, A.-L., & Wang, J. (2010). Determination of Brix and POL in Sugar Cane Jugo by Using Near Infrared Spectroscopy Coupled with BP-ANN. *Spectroscopy and Spectral Analysis*, *30*(7), 1759-1762. https://doi.org/10.3964/j.issn.1000-0593(2010)07-1759-04

Woudenberg, S. P. D., Gaag, L. C. V. D., Feelders, A., & Elbers, A. R. W. (2014). Real-time Adaptive Problem Detection in Poultry. En I. Press (Ed.), *ECAI 2014* (pp. 1217-1218). https://doi.org/10.3233/978-1-61499-419-0-1217

Wu, C.-H., Tzeng, G.-H., & Lin, R.-H. (2009/4). A Novel hybrid genetic algorithm for *kernel* function and parameter optimization in support vector regression. *Expert systems with applications*, *36*(3, Part 1), 4725-4735. https://doi.org/10.1016/j.eswa.2008.06.046

Wu, J., & Wei, J. (2007). Combining ICA with SVR for prediction of finance time series. En *2007 IEEE International Conference on Automation and Logistics* (pp. 95-100). ieeexplore.ieee.org. https://doi.org/10.1109/ICAL.2007.4338537

Xiao, J., Wang, H., Shi, L., Lv, M., & Ma, H. (2011). The Development of Decision Support System for Production of Layer. En D. Li & Y. Chen (Eds.), *Computer and Computing Technologies in Agriculture V* (pp. 161-168). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27281-3_21

Xie, L., He, X., Duan, B., Tang, S., Luo, J., Jiao, G., … Hu, P. (2015). Optimization of Near-

Infrared Reflectance Model in Measuring Gelatinization Characteristics of Rice Flour with a Rapid Viscosity Analyzer (RVA) and Differential Scanning Calorimeter (DSC). *Cereal chemistry*, *92*(5), 522-528.

Xu, L., Zhou, Y.-P., Tang, L.-J., Wu, H.-L., Jiang, J.-H., Shen, G.-L., & Yu, R.-Q. (2008). Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta*, *616*(2), 138-143. https://doi.org/10.1016/j.aca.2008.04.031

Zahirnia, *K*., Teimouri, M., Rahmani, R., & Salaq, A. (2015). Diagnosis of type 2 diabetes using cost-sensitive learning. En *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)* (pp. 158-163). https://doi.org/10.1109/ICCKE.2015.7365820

Zamora-Rojas, E., Pérez-Marín, D., De Pedro-Sanz, E., Guerrero-Ginel, J. E., & Garrido-Varo, A. (2012). Handheld NIRS analysis for routine meat quality control: Database transfer from at-line instruments. *Chemometrics and Intelligent Laboratory Systems*, *114*, 30-35. https://doi.org/10.1016/j.chemolab.2012.02.001

Zayas-Ruiz, E. P., Lorenzo-Izquierdo, M., & Fragoso-Concepción, F. O. (2015). La quimiometría y la industria del azúcar y sus derivados. *ICIDCA. Sobre los Derivados de la Caña de Azúcar*, *49*(3), 31-33.

Zhang, L., Zhou, W.-D., Chang, P.-C., Yang, J.-W., & Li, F.-Z. (2013). Iterated time series prediction with multiple support vector regression models. *Neurocomputing*, *99*, 411-422. https://doi.org/10.1016/j.neucom.2012.06.030

Zhao, J., Lin, H., Chen, Q., Huang, X., Sun, Z., & Zhou, F. (2010). Identification of egg's freshness using NIR and support vector data description. *Journal of food engineering*, *98*(4), 408-414. https://doi.org/10.1016/j.jfoodeng.2010.01.018

Zhiliang, L., Zuo, M., Zhao, X., & Xu, H. (2015). An Analytical Approach to Fast Parameter Selection of Gaussian RBF *Kernel* for Support Vector Machine *. *Journal of Information Science & Engineering*, *710*(51375078), 691-710.

Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE transactions on knowledge and data engineering*, *18*(1), 63-77. https://doi.org/10.1109/TKDE.2006.17

Zhu, D., Li, *K*., Terry, D. P., Puente, A. N., Wang, L., Shen, D., … Liu, T. (2014). Connectome-scale assessments of structural and functional connectivity in MCI. *Human Brain Mapping*, *35*(7), 2911-2923. https://doi.org/10.1002/hbm.22373

ANEXOS

Computers and Electronics in Agriculture 121 (2016) 169–179

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Early warning in egg production curves from commercial hens: A SVM approach

Iván Ramírez Morales [a,b,*], Daniel Rivero Cebrián [b], Enrique Fernández Blanco [b], Alejandro Pazos Sierra [b]

[a] Universidad Técnica de Machala, Faculty of Agricultural & Livestock Sciences, 5.5 km Pan-American Av, Machala, El Oro, Ecuador
[b] Universidade A Coruña, Department of Computer Science, 15071 A Coruña A Coruña (03082), A Coruña, Spain

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form 15 December 2015
Accepted 18 December 2015

Keywords:
Early warning
Drop in egg production
Poultry management
Support vector machines
Machine learning

ABSTRACT

Artificial Intelligence allows the improvement of our daily life, for instance, speech and handwritten text recognition, real time translation and weather forecasting are common used applications. In the livestock sector, machine learning algorithms have the potential for early detection and warning of problems, which represents a significant milestone in the poultry industry. Production problems generate economic loss that could be avoided by acting in a timely manner.

In the current study, training and testing of support vector machines are addressed, for an early detection of problems in the production curve of commercial eggs, using farm's egg production data of 478,919 laying hens grouped in 24 flocks.

Experiments using support vector machines with a 5 k-fold cross-validation were performed at different previous time intervals, to alert with up to 5 days of forecasting interval, whether a flock will experience a problem in production curve. Performance metrics such as accuracy, specificity, sensitivity, and positive predictive value were evaluated, reaching 0-day values of 0.9874, 0.9876, 0.9783 and 0.6518 respectively on unseen data (test-set).

The optimal forecasting interval was from zero to three days, performance metrics decreases as the forecasting interval is increased. It should be emphasized that this technique was able to issue an alert a day in advance, achieving an accuracy of 0.9854, a specificity of 0.9865, a sensitivity of 0.9333 and a positive predictive value of 0.6135. This novel application embedded in a computer system of poultry management is able to provide significant improvements in early detection and warning of problems related to the production curve.

1. Introduction

Poultry farmers have used data to monitor health and production of their flocks for over 40 years. Data such as consumption of food, water, growth and mortality have been collected in order to monitor and improve yields, and these data and statistics are plotted on a graph and used as early indicators of the health and welfare of poultry (Hepworth et al., 2012).

Egg producers usually know and record the number of eggs produced, frequently a production curve is plotted and monitored in order to detect problems in the production curve indicating a possible disease, or any other issues (Long and Wilcox, 2011).

The curve of egg production can be affected by various factors such as food intake (quality and quantity), water consumption, intensity and duration of the light received, vermin infestation, diseases and other handling or environmental causes (Jacob et al., 2014).

When it comes to a disease, having early detection tools is of vital importance. That is, before it is spreading to other animals and/or becoming entrenched in the environment. The early detection of a problem means acting in a timely manner; reducing the cost and increasing the effectiveness of the treatment or control of a disease are directly related to the time it takes to detect it (Schaefer et al., 2004; Cameron, 2012).

The machine learning algorithms are present in various activities of our daily life, and they allow discovering rules and patterns in data sets. For example, in epidemiology, the supervised machine learning has the potential to classify, diagnose and identify risks. Support vector machines, are one of this algorithms, the main feature is that they can learn how to classify data from examples (McQueen et al., 1995; Hepworth et al., 2012).

* Corresponding author at: Universidad Técnica de Machala, Faculty of Agricultural & Livestock Sciences, 5.5 km Pan-American Av, Machala, El Oro, Ecuador.
E-mail address: iramirez@utmachala.edu.ec (I.R. Morales).

http://dx.doi.org/10.1016/j.compag.2015.12.009
0168-1699/

References to studies that used machine learning techniques in livestock have been found, for example, various algorithms were employed to predict the rate of pregnancy, or weight in cattle, from routine production data (Hempstalk et al., 2015; Alonso et al., 2015).

Support vector regression and neural networks to predict the body and carcass characteristics of broilers (Faridi et al., 2012). Support vector machines to predict hock burn in chickens (Hepworth et al., 2012). Artificial intelligence and images to detect the avian smallpox (Hemalatha et al., 2014).

Lokhorst and Lamaker (1996) reported an expert system for monitoring the daily production process in aviary systems for laying hens, however, no information has been found regarding the early detection of problems using farm's data which are normally recorded in poultry production.

To the best of the authors' knowledge, there are no prior studies on using machine learning algorithms for early detection of problems in the egg production curve from commercial hens. Although, since the early 1980s there are similar works in the mathematical study of the production curve of laying hens. Nonlinear models have been widely used to adjust the curves of egg production in laying hens (Adams and Bell, 1980; Grossman and Koops, 2001; Savegnago et al., 2012).

Moreover, a vast amount of literature has been compiled, for over 30 years, on the use of control charts to monitor animal farming, but its practical use does not seem to be widespread (De Vries and Reneau, 2010).

Studies such as those carried out by Grossman et al. (2000) and Narinc et al. (2014) have been found, who developed mathematical models to describe the production curve and the persistence of the curve in laying hens. Other works, such as those published by Long and Wilcox (2011), studied the production curve of laying hens to determine whether the economic use of flocks of laying hens was optimal.

Some learning techniques have been used to model the production curve, especially artificial neural networks, showing that they are able to successfully replace traditional mathematical and statistical models when predicting egg production in laying hens. These models, which are easier to use, require fewer variables and can be more efficiently compared with their mathematical counterparts (Ahmadi and Golian, 2008; Ahmad, 2011; Felipe et al., 2015).

There is general agreement on the need to monitor the production yield of farm animals, that is why mathematical methods (Dohoo, 1993), recursive algorithms (Roush et al., 1992), data display systems and statistical techniques (Woodall and Tech, 2006) have been used. Significant differences indicating an alteration in the productive indicators of farm animals are sought (De Vries and Reneau, 2010).

The real-time monitoring is a major challenge because data collection includes natural variability; Woudenberg et al. (2014) developed a method for early detection of problems based on the calculation of waste, which allows identifying potential problems in egg production from 10 production flocks.

The concept of control charts as part of the statistical process control is commonly used to monitor industrial processes; several authors demonstrated their use in the context of animal husbandry, although the statistical properties of data regarding animals often do not meet the basic principles of these control charts (Mertens et al., 2011).

In the above-mentioned cases, computer-aided detection methods were presented, but no publications were found on the use of machine learning algorithms aimed at developing models that allow partial automation of this task.

This study is aimed at developing and testing an early warning model based on support vector machines algorithms, in order to detect problems in egg production curve from commercial hens.

2. Materials and methods

2.1. Data description

A farm database of egg production of laying hens of the ISA Brown, Lohmann Brown and H&N layer lines were used, collected over a period of seven years (January 2008 to December 2014) from a poultry company. Data correspond to 24 flocks, of approximately 20,000 birds at the beginning of the production cycle, using the "all-in all-out" replacement system, i.e. each flock contains only birds of the same age at the beginning, during the entire production period and when the production cycle is completed.

Data are recorded once a day, at the end of the day, but not always at same time, it is done when counting and sorting of eggs and dead birds have been carried out, and it also depends on the weekday. The production period used for the experiments encompassed 60 weeks (from age 19 to 79 weeks), for each day in which there was a production problem was labeled as positive by an experts' panel formed by the farm's production manager (veterinarian), the owner who has been poultry farmer for 30 years, and a local poultry veterinarian.

The average number of days labeled as positive for each flock is 8 days, however, it is observed that there are flocks, which present no problem, and there are others, which present up to 33 days labeled as positive. In total, the 24 flocks, throughout the 7 years of study, presented 188 positive labels, representing only 1.85% of the 10,142 records. That is, the classifier has a lot of negative patterns (days when there are no problems) and few positive patterns, this fact unbalances the expected outputs and adds difficulty to the task of classification and forecasting.

Table 1 describes each flock with its corresponding general indicators: production time; birds housed at the beginning of the production cycle; dead birds during the production time; total eggs produced by the flock during the production time; average number of eggs produced per day; daily eggs per hen housed; production maximum% (peak) reached and the number of positive labels of each flock.

Fig. 1 shows three flocks which are representative for the database: the solid line represents flock 11, which has a characteristic curve, without any problems throughout the production time; the dotted line represents flock 21, which has small drops and delays in the production curve, but they are not significant; the dashed line represents flock 1, which has two significant production drops, the first one begins near 31 weeks and the second at 72 weeks old.

From the numerous meetings with the poultry farmers, it is found that on farms where the collection of eggs is done at a specific time, there is greater data consistency than on those where it is carried out at different times each day; on the farm where no standard time routine was established for the collection of eggs, in either house, the number of eggs produced per day varies. This variability can be observed in Fig. 2, representing the daily egg production per bird. This fact represents an additional challenge for the early warning model, because it should be able to distinguish between a real problem and these drops due to weekly cyclical variations related to routine and time of collection

Fig. 3 shows an example of problem zone tagged using poultry experts' judgment, each production day of each flock was labeled with values of 0 in the absence of a problem and 1 otherwise.

2.2. Support vector machines algorithms

Basically there are two types of machine learning algorithms: supervised and unsupervised; the former is used when there is knowledge about the desired outputs, and it is trained to obtain

Table 1
Main production indicators of the flocks under study.

Flock	Production time (d)	Housed birds	Dead birds	Total amount of eggs	Average eggs per day	Eggs/housed bird/day	Peak of the production% per bird/day	Positive labels
1	473	20,300	2929	7,211,252	15,246	0.7510	98.14%	30
2	429	20,361	3022	6,951,132	16,203	0.7958	97.04%	0
3	516	20,137	3630	8,160,013	15,814	0.7853	96.87%	1
4	148	18,874	1430	1,767,577	11,943	0.6328	N/A	0
5	480	19,770	2421	7,185,831	14,970	0.7572	97.25%	0
6	461	20,408	1573	7,145,492	15,500	0.7595	97.11%	33
7	518	20,187	2718	7,974,633	15,395	0.7626	95.97%	14
8	501	20,130	1984	8,093,083	16,154	0.8025	97.03%	0
9	104	19,740	436	1,594,527	15,332	0.7767	97.03%	0
10	389	19,668	2153	6,078,320	15,626	0.7945	95.86%	0
11	543	19,920	2409	7,900,793	14,550	0.7304	97.32%	0
12	491	19,934	1969	7,230,558	14,726	0.7387	98.70%	17
13	431	19,492	1382	6,787,937	15,749	0.8080	96.30%	13
14	419	19,920	1600	7,147,832	17,059	0.8564	97.17%	0
15	468	20,120	1549	7,172,119	15,325	0.7617	98.91%	0
16	517	20,234	2865	7,692,698	14,879	0.7354	97.42%	12
17	498	19,971	2051	7,744,766	15,552	0.7787	96.83%	24
18	391	20,104	1238	6,463,994	16,532	0.8223	97.72%	13
19	307	20,094	693	5,301,566	17,269	0.8594	98.52%	0
20	450	19,895	1984	6,905,452	15,345	0.7713	98.16%	13
21	480	19,910	2702	7,590,784	15,814	0.7943	96.94%	0
22	529	19,950	2973	8,429,271	15,934	0.7987	98.20%	10
23	374	19,907	2050	6,023,519	16,106	0.8090	98.30%	8
24	202	19,893	814	3,407,626	16,869	0.8480	97.63%	0

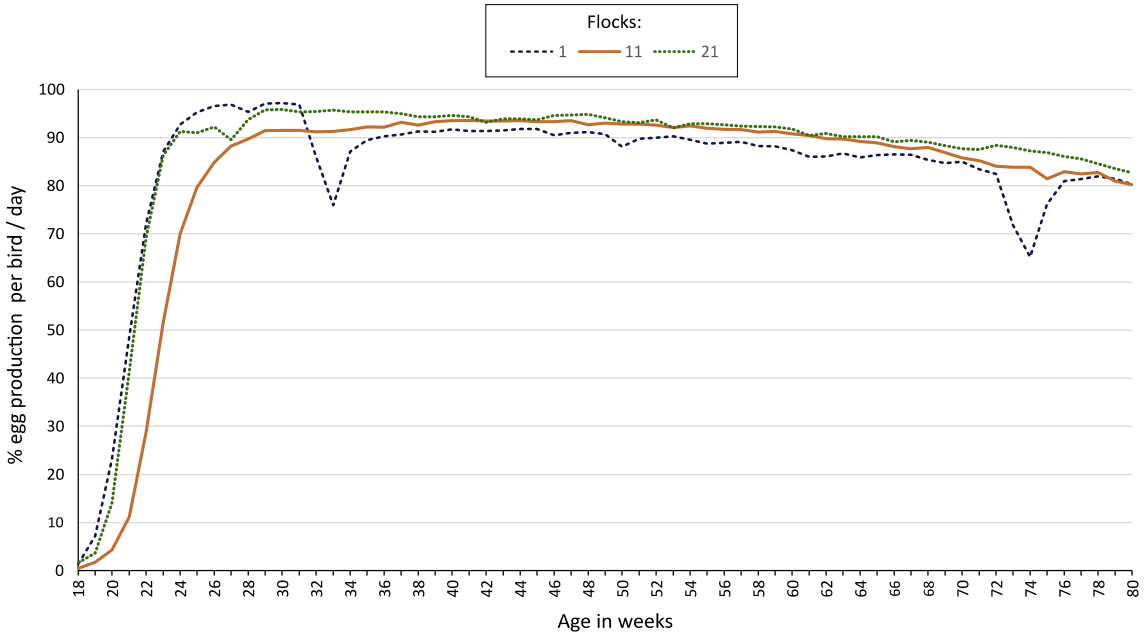


Fig. 1. Weekly average production per bird in three representative flocks.

them, whereas the latter generates a grouping (clusters) without information on the expected outputs (Mucherino et al., 2009).

Once an algorithm has been trained, it is able to transfer the learned dependence between the input patterns (features) and expected outputs (targets) into new data. The quality of a classifier can be measured by the proportion of correctly classified patterns in the test set, i.e. in new data that were not used during training, this set allows for assessing the error in the generalization of the final model chosen (Hastie et al., 2009).

Among the most commonly used techniques for data mining, are the support vector machines (SVM), which are supervised machine learning algorithms used to classify data sets into two different classes, separated by a hyperplane defined in an appropriate space (Mucherino et al., 2009).

They can be used in classification and regression problems, as their functioning starts from a set of training samples whose classes are labeled, and they also train an SVM to build a model that predicts the class of a new sample, different from the original one (Palma and Marín, 2013; Benítez et al., 2013).

The basics of SVM were developed by Vapnik and Chervonenkins in 1963, in a study on the theories of statistical learning that was aimed at narrowing down the generalization error according to the complexity of the search space. In 1992 Vapnik, Boser and Guyon proposed a method to create non-linear classifiers (Boser et al., 1992), and the current standard of SVM was proposed by Cortes and Vapnik (1995). The purpose of SVM is for obtaining models which structurally have little risk of error regarding future data. Although originally they were designed to solve binary (two

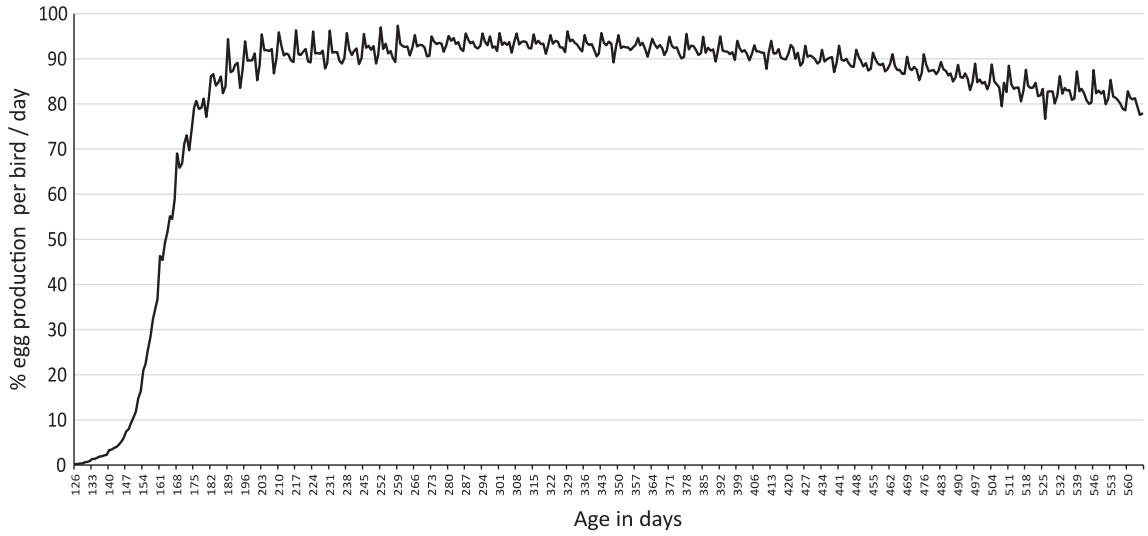


Fig. 2. Daily records of egg production per bird in flock 11.

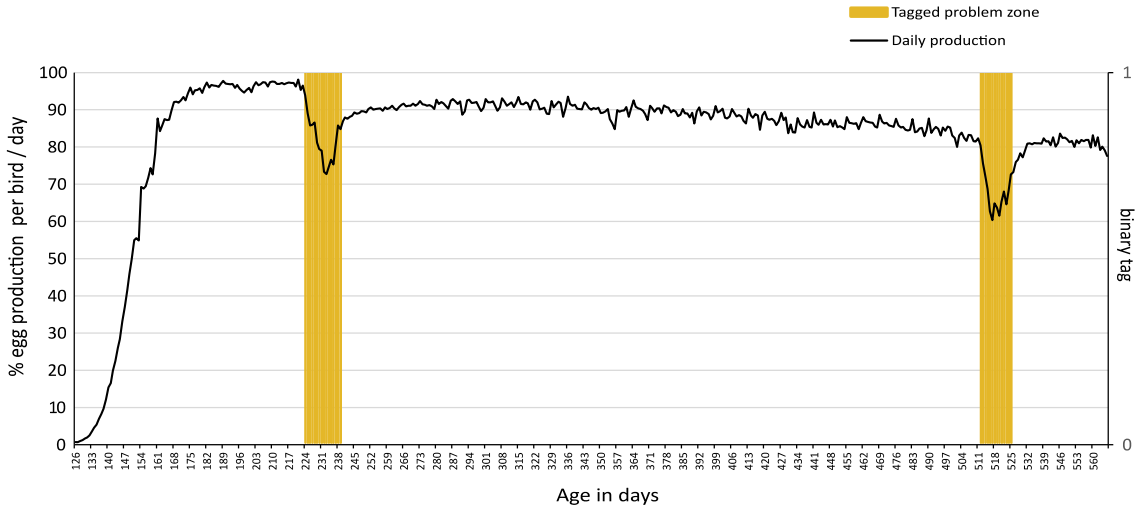


Fig. 3. Labeling example of problems in flock 1.

classes) classification problems, their application has been extended to regression, multiclassification, clustering and other tasks (Palma and Marín, 2013).

This technique is intended to find an optimal hyperplane able to distribute data into the classes to which they belong. Intuitively, it seems obvious to conclude that when facing a problem of linear classification there is a high probability of obtaining several solutions which successfully classify data (Fernandez-Lozano et al., 2013).

The optimal hyperplane used to separate the two classes can be defined from a small amount of data from the training set called support vectors, which determine the margin (Cortes and Vapnik, 1995; Mucherino et al., 2009). Fig. 4 shows the above-mentioned concepts.

The choice of the best hyperplane was solved in 1965 (Vapnik and Kotz, 1982) with the approach that the optimal hyperplane is defined as the linear decision function with the maximum margin between the vectors of the two classes.

However, in most problems, the data are not linearly separable and it is required to use strategies such as the identification of other separation dimensions. The kernel functions are used to transform the original multidimensional space into another, where

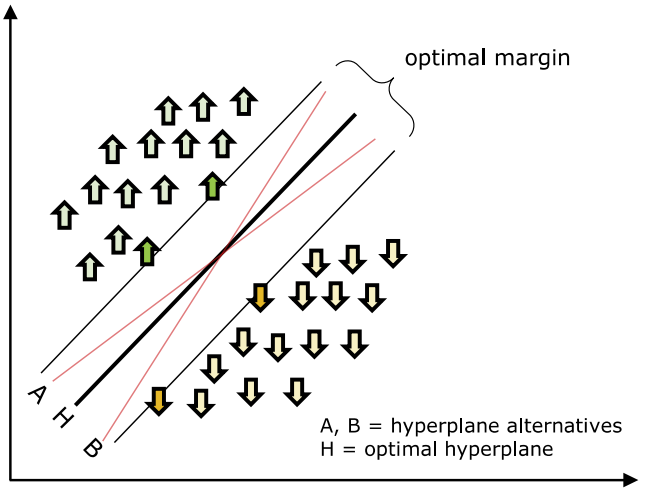


Fig. 4. A problem separable in a two-dimensional space. Support vectors define the margin of greatest separation between classes.

classes are linearly separable. In practice, support vector machines are trained using different kernels to select the one with the best performance for the problem raised (Mucherino et al., 2009).

Some preliminary tests were performed by trial and error on the test set (Mollazade et al., 2012), using most common available kernels, however only four had an acceptable accuracy to the opinion of the authors. The research focused on these kernels: polynomial, radial basis function – RBF (Gaussian), quadratic and linear, in order to perform an exhaustive evaluation.

The polynomial and RBF kernels are among the most commonly used ones; the latter has a sigma (σ) parameter which can be tuned to adjusts the size of the kernel (Bennett and Campbell, 2000). Preliminary tests were performed to select the sigma tuning best range, which was between one and six; this range was used for exhaustive evaluation.

SVM has a compensation parameter *C*, which can be modified and affects the classification quality, since it determines how severely any misclassification should be penalized; generally, very high *C* values may lead to overfitting problems, reducing the SVM ability to generalize (Mucherino et al., 2009). In order to evaluate this parameter without overfitting the classifier, values below 0.25 were selected.

2.3. Data processing

Starting from the production data, two sets of patterns were created: the inputs, which had SVM and the desired outputs for them. The input patterns are made up by taking data from a sliding window (Lindsay and Cox, 2005), with a sample of current day and some previous and consecutive samples, according to the windows size.

During the preliminary determination of optimum window size, several trials were performed, finding out that numbers which are multiples of seven, had better performance than other values, it could be due to the weekly cyclical variations related to routine and time of collection, referred previously on Fig. 2.

From a collection of more than 30 initial features, preliminary testing was conducted, in which six relevant features were selected. It was determined that features like the genetic line of birds, stochastic variations in egg production, daily and cumulative mortality, weekly slope of the curve, and many others, don't provide a significant improvement to the model, and were discarded (Mollazade et al., 2012).

The feature selection for the input patterns of the SVM was defined as follows:

A. The production percentage over a day (number of eggs produced over a day/number of existing birds) minus the percentage of historical production for a similar day.
B. The production percentage over the day at the end of the sliding window, minus the production percentage over the day at the beginning of the sliding window.
C. The production over the day minus the production from seven days earlier.
D. The coefficient of variation (standard deviation/mean $*$ 100) of the second half of the sliding window.

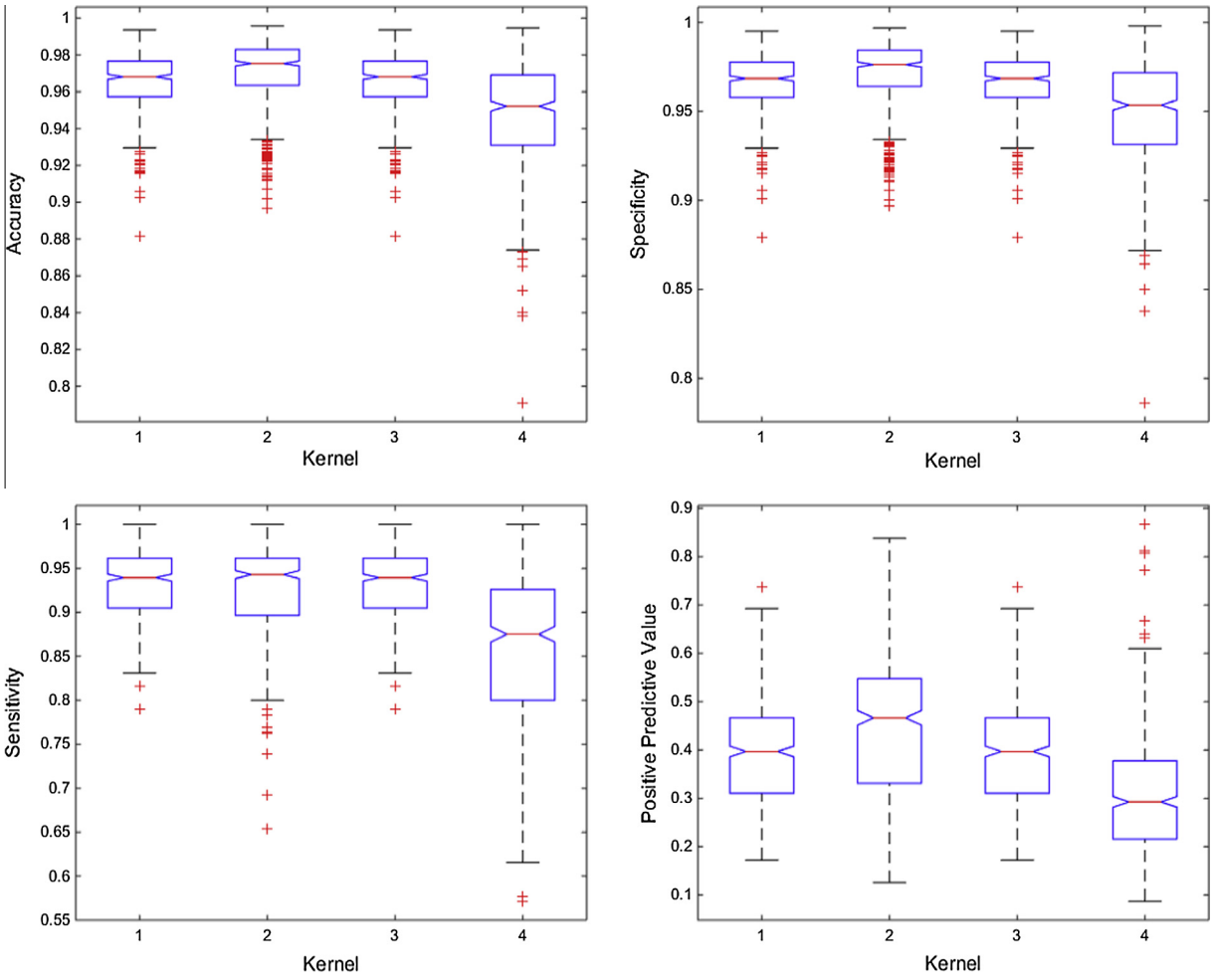


Fig. 5. Diagrams of boxes: performance metrics of the kernels evaluated.

Table 2
Multiple comparison (MC) of kernels for each performance metric.

	Kernel			
	1 polynomial	2 radial basis function	3 quadratic	4 linear
Accuracy	0.9654[a]	0.9687[a]	0.9654[a]	0.9475[b]
Specificity	0.9661[a]	0.9696[a]	0.9661[a]	0.9492[b]
Sensitivity	0.9289[a]	0.9203[a]	0.9289[a]	0.8546[b]
Positive Predictive Value	0.3932[b]	0.4445[a]	0.3932[b]	0.3045[c]

Rows with different letters differ significantly according to Tukey's Honest Significant Difference method for a value of $p < 0.01$.

E. The standard deviation of the first half of the sliding window minus the standard deviation of the second half of the sliding window.
F. Age of birds in weeks.

Relevant features, as determined by the authors, were selected from the sliding window, each input pattern is having a corresponding pattern in the output set, which were zero or one, depending on whether the label of the day at the end of the forecasting interval was positive or negative regarding the presence of a problem in the curve. This procedure is performed for each day during the study period, always extracting the same fixed features.

To assess the forecasting interval, expected outputs for each sliding window has been taken from corresponding pattern in output set (zero-day forecasting interval), and a time shift (Lindsay and Cox, 2005) of one to five days later, that way SVM leaning is based next days expected outputs, and thus SVM trained could be able to detect problems prior to experts' criteria.

A k-fold cross-validation technique was used in order to ensure that the results were independent of the partition between the training and test data, also cross-validation prevents an overfitting problem (Hsu et al., 2003), thereby the subsets of each fold were a representative sample containing flocks which presented problems and flocks which did not, in a random and stratified manner.

During the k-fold cross-validation process, the data are divided into k subsets; one is used as a test subset and the others $(k-1)$ as training subsets (Mucherino et al., 2009). The cross-validation process is repeated for k folds, with each of the possible subsets, and finally an arithmetic mean of the results for each fold is performed to obtain a single result, which is passed on to the SVM.

Thus, 100 repetitions of k-fold 5 cross-validation were performed. For this study, 12,500 support vector machines were evaluated, 500 for each factor of variation.

2.4. Performance analysis

The first performance requirement for a classification model is that the model generalizes well, in the sense that it provides the correct predictions for new, unseen data instances (generalization). This behavior is typically measured by percentage correctly classified test instances (accuracy), other measures include sensitivity and specificity, which are generated from a confusion matrix (Martens and Baesens, 2010).

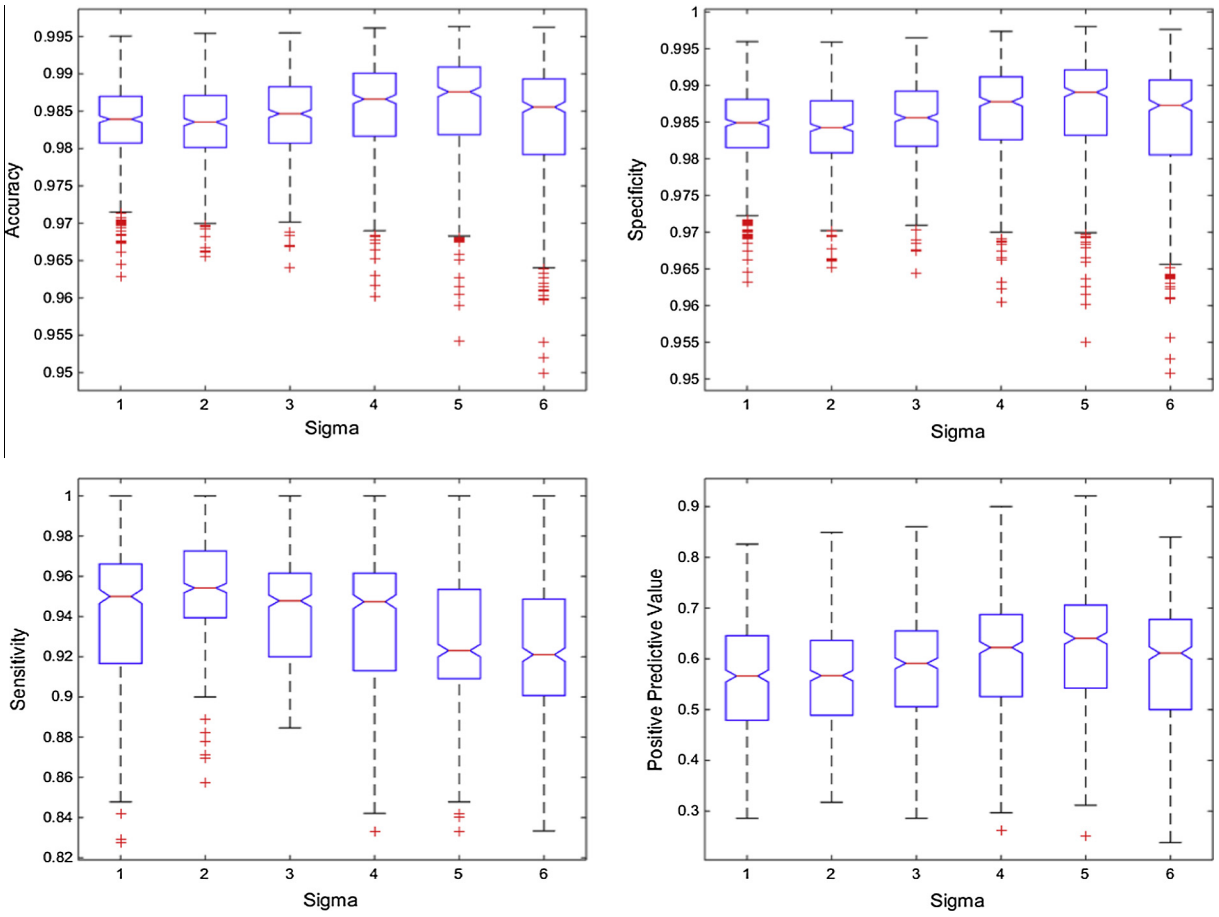


Fig. 6. Diagrams of boxes: performance metrics according to the value of sigma.

Table 3
MC of different sigma (σ) values for each performance metric.

	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$	$\sigma = 6$
Accuracy	0.9833^{b}	0.9833^{b}	0.9842^{ab}	0.9853^{a}	0.9856^{a}	0.9833^{b}
Specificity	0.9843^{b}	0.9840^{b}	0.9851^{b}	0.9864^{ab}	0.9869^{a}	0.9847^{b}
Sensitivity	0.9406^{b}	0.9520^{a}	0.9419^{b}	0.9365^{b}	0.9259^{c}	0.9199^{c}
Positive predictive value	0.5657^{b}	0.5647^{b}	0.5836^{b}	0.6100^{ab}	0.6223^{a}	0.5858^{b}

Rows with different letters differ significantly according to Tukey's Honest Significant Difference method for a value of $p < 0.01$.

The accuracy value is usually the only performance requirement used for evaluating the performance of machine learning techniques; this accuracy value is a statistical measure used to determine whether a binary (true or false) classification test is able to correctly identify or exclude a condition (Martens and Baesens, 2010; Venkatesan et al., 2013).

Considering that there are only 188 positive labels and 9954 negative labels in the database, samples Tang et al. (2009) states it is required to evaluate other metrics such as specificity and sensitivity, to avoid misinterpretations when having rare positive labels. A common used example to support this statement is that a classifier, which predicts all samples as negative, has high accuracy, but it is useless to detect rare positive.

The aim of this study is related to detection of problems in egg production. Therefore it is very important to achieve a highly effective detection ability for positive labels, for this, Tang et al. (2009) suggests another metric, called precision or positive predictive value.

Specificity is the ability to detect the absence of problems as false; sensitivity is the ability to detect the presence of problems as true; and positive predictive value is the probability that a problem actually occurs when the test is positive (Altman and Bland, 1994; Tang et al., 2009; Hastie et al., 2009; Venkatesan et al., 2013).

Analysis of Variance (ANOVA) and Multiple Range Tests (MRT) with Tukey's Honest Significant Difference (HSD) method for a value of $p < 0.01$, were performed to select the optimal model configuration; a positive selection of those parameter that provided the best performance metrics was carried out. Metrics were calculated from the confusion matrix of the test subset, that is, data different from those used for training, reducing the possibility of overtraining and improving its ability to generalize.

3. Results

3.1. Kernel selection

With fixed values of parameter C to 0.1, windows size to seven and forecasting interval, to one, until they are assessed respectively, four kernels were exhaustively evaluated: (1) polynomial, (2) radial basis function (RBF), (3) quadratic and (4) linear. The

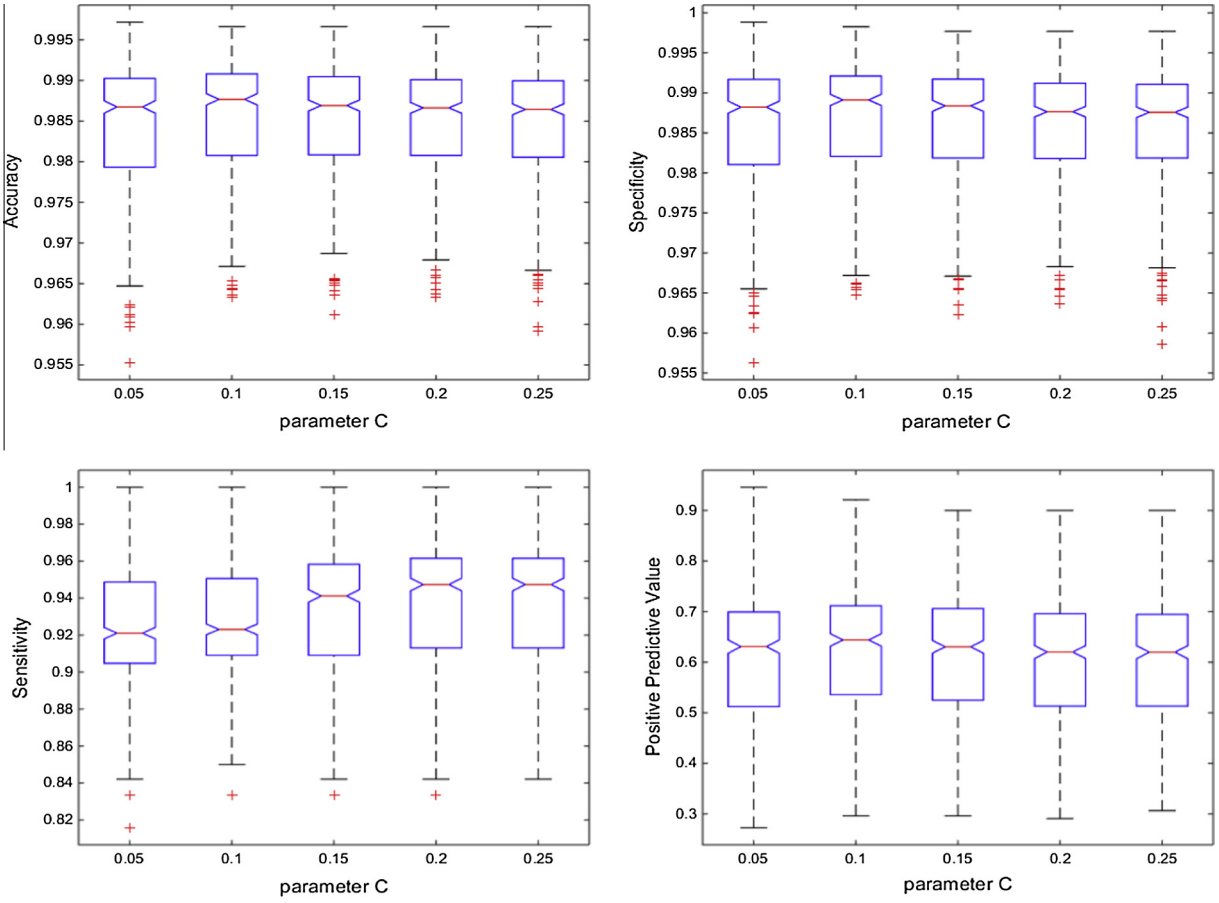


Fig. 7. Diagrams of boxes: performance metrics according to the value of the parameter C.

Table 4
MC of different values of the parameter C for each performance metric.

	$C = 0.01$	$C = 0.1$	$C = 0.15$	$C = 0.2$	$C = 0.25$
Accuracy	0.9845[a]	0.9856[a]	0.9852[a]	0.9849[a]	0.9847[a]
Specificity	0.9859[a]	0.9869[a]	0.9864[a]	0.9860[a]	0.9858[a]
Sensitivity	0.9185[b]	0.9256[b]	0.9328[a]	0.9343[a]	0.9358[a]
Positive Predictive Value	0.6063[a]	0.6222[a]	0.6131[a]	0.6057[a]	0.6025[a]

Rows with different letters differ significantly according to Tukey's Honest Significant Difference method for a value of $p < 0.01$.

results of the kernel evaluation is shown in a diagram of boxes on Fig. 5.

As shown, kernel 4 (linear) produces the worst results, whereas kernels 1, 2 and 3 (polynomial, RBF and quadratic) obtain similar results between them. An ANOVA statistical test was performed using the multiple comparison procedure, Tukey HSD, which is shown in Table 2.

From the above-mentioned analysis, it is selected radial basis function kernel, as it is statistically better in all four parameters evaluated, polynomial and quadratic kernels have similar performance, but have a statistically significant lower positive predictive value.

From preliminary tests performed, the sigma tuning best range was between one and six; exhaustive test through this range was performed in a gradient ascent optimization, seeking optimal performance of the model. Fig. 6 shows the results of performance metrics according to values of tuned sigma.

As noted, in terms of accuracy, specificity and positive predictive value, the values tend to improve as sigma is higher (up to five), however, the sensitivity values tend to worsen as sigma is higher. In order to produce the best decision-making tool, an ANOVA statistical test was performed using the multiple comparison procedure, Tukey HSD, which is shown in Table 3.

A value of sigma equal to two performed the best sensitivity; however, the accuracy, specificity and positive predictive value metrics, are on group b according to Tukey's test. On the other hand, a value of sigma equal to five performed the best accuracy, specificity and positive predictive value, with the lowest sensitivity among those evaluated. A sigma value equal to five is set since it improves most of the performance metrics.

3.2. Parameter C

Once it was decided to set the RBF kernel, with a sigma value equal to five, an evaluation was conducted by varying the parameter C; the initial values of the window size and the forecasting interval remained constant, and only values of C below 0.25 were tested. Fig. 7 shows the results from the evaluation of different values of the parameter C.

As noted, for accuracy, specificity and positive predictive value, minor modifications of the parameter C do not generate significant differences, whereas, for sensitivity, the modification of this parameter does generate slight increases. In order to make the best decision, an ANOVA statistical test was performed using the multiple comparison procedure, Tukey HSD, which is shown in Table 4.

Based on the results, any value of the parameter C could be set to 0.15 or higher. Recognizing that by setting a lower value there

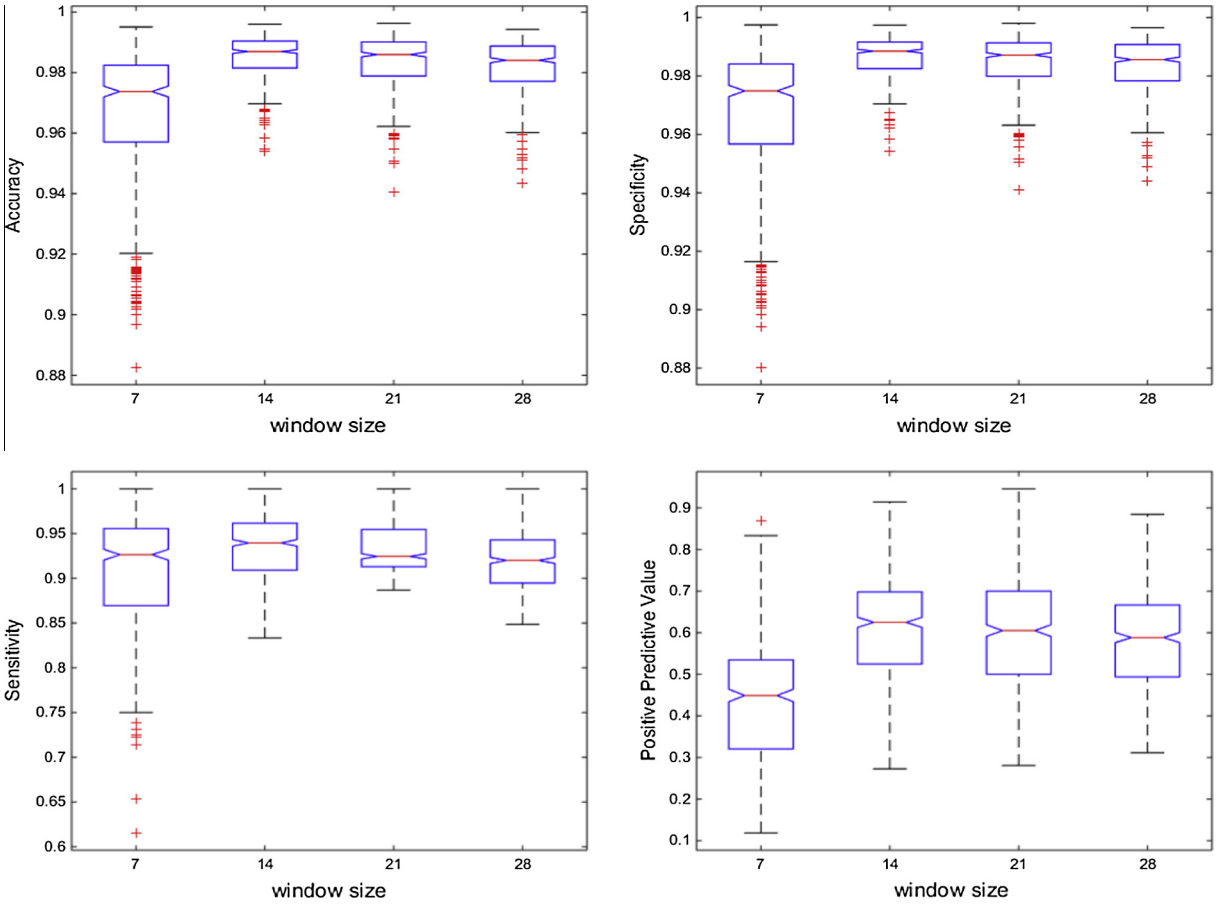


Fig. 8. Diagrams of boxes: performance metrics according to the window size.

Table 5
MC of different values of window size for each performance metric.

	WS = 7	WS = 14	WS = 21	WS = 28
Accuracy	0.9659^{c}	0.9852^{a}	0.9838^{ab}	0.9821^{b}
Specificity	0.9670^{c}	0.9864^{a}	0.9850^{ab}	0.9836^{b}
Sensitivity	0.9020^{c}	0.9318^{a}	0.9320^{a}	0.9168^{b}
Positive Predictive Value	0.4300^{c}	0.6122^{a}	0.5972^{ab}	0.5808^{b}

Rows with different letters differ significantly according to Tukey's Honest Significant Difference method for a value of $p < 0.01$.

was less possibility of overfitting, it was decided to select the parameter C value at 0.15.

3.3. Window size

The window size expresses the amount of data in the days before the event, which are supplied to the model in order to configure the input patterns, multiples of 7 values from 7 to 28 days were evaluated, and the results are shown below in Fig. 8.

It is clearly noted that a window size equal to 7 generates the worst results in all the performance metrics, whereas the values among those for 14, 21 and 28 produce similar results. Therefore, an ANOVA statistical test was performed using the multiple comparison procedure, Tukey HSD, which is shown in Table 5.

The Tukey's test results unequivocally indicate that the optimal window size for this type of problem is 14 days.

3.4. Forecasting interval

The forecasting interval can be adjusted to suit the specific demands, a value equal to zero implies that the model works as an early warning; values higher or equal to one imply that it works as a forecasting model. Forecasting interval values were evaluated between zero and five, the results of experiments performed are shown in Fig. 9.

As expected, the shorter the forecasting interval, better performance is obtained for all performance metrics. Table 6 shows performance metrics of the assessed forecasting intervals, an ANOVA statistical test using the multiple comparison procedure, Tukey HSD was performed.

4. Discussion

For an early warning of problems in egg production curve, SVM classifier is proposed by authors not to classify but to detect abnormal instances, as stated by Bennett and Campbell (2000) about novelty or abnormality detection potential applications in many problem domains. Lindsay and Cox (2005) state that traditional machine learning techniques, like SVM, can be a viable alternative to the classical time-series analysis technique. In this study, different settings of SVM parameters were assessed using ANOVA statistical tests and Tukey Multiple Comparison tests for a value of $p < 0.01$.

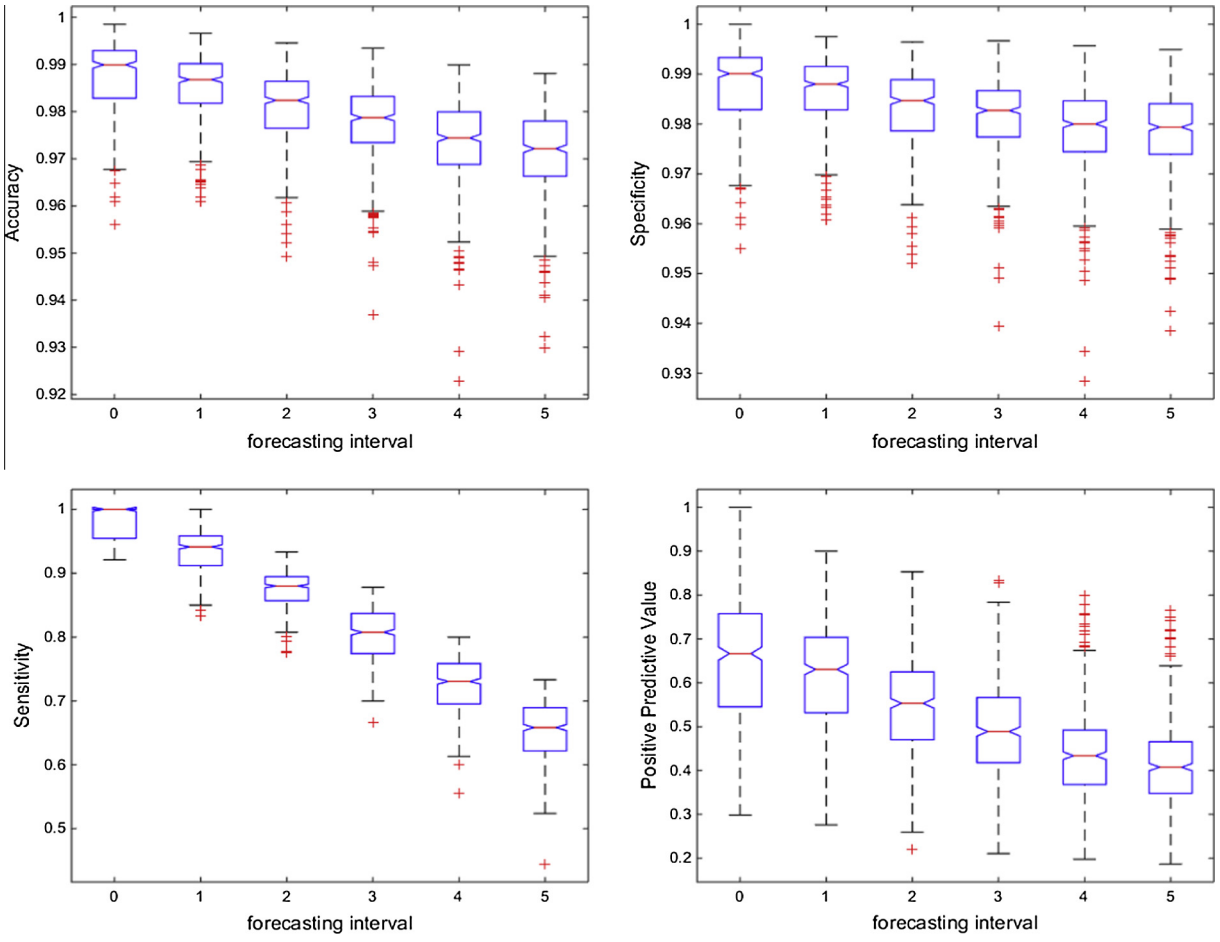


Fig. 9. Diagrams of boxes: performance metrics according to the forecasting interval.

Table 6
MC of different values of forecasting interval for each performance metrics.

	FI = 0	FI = 1	FI = 2	FI = 3	FI = 4	FI = 5
Accuracy	0.9874[a]	0.9854[b]	0.9811[c]	0.9776[d]	0.9735[e]	0.9713[f]
Specificity	0.9876[a]	0.9865[b]	0.9835[b]	0.9814[c]	0.9789[d]	0.9783[d]
Sensitivity	0.9783[a]	0.9333[b]	0.8738[c]	0.8030[d]	0.7229[e]	0.6483[f]
Positive Predictive Value	0.6518[a]	0.6135[b]	0.5480[c]	0.4940[d]	0.4419[e]	0.4090[f]

Rows with different letters differ significantly according to Tukey's Honest Significant Difference method for a value of $p < 0.01$.

Since kernel is arguably the most important component of SVM algorithm (Suttorp and Igel, 2007; Zhao et al., 2010; Mollazade et al., 2012), exhaustive tests with four kernels were assessed in order to select the one with the best performance as proposed by Mucherino et al. (2009).

RBF, polynomial and quadratic kernels had similar performance on accuracy, specificity and sensitivity, the positive predictive value achieved by RBF kernel was better than the other kernels evaluated. The authors selected RBF kernel, which has been proved to be an excellent kernel function for several applications, agreeing with Fernández Pierna et al. (2006), Han et al., (2007), Zhao et al., (2010) and Zhiliang et al. (2015).

According to Bennett and Campbell (2000) and Zhao et al. (2010) when RBF kernel is used, sigma parameter must be optimized, in order to obtain better performance. A common technique for this is stepping through a range of values for sigma, in a gradient ascent optimization (Suttorp and Igel, 2007). The selected range to evaluate the model was one to six.

A value of sigma equal to five performed the best accuracy, specificity and positive predictive value, 0.9856, 0.9869, 0.6223 respectively, nevertheless, performed a sensitivity value of 0.9259, the worst among those evaluated; the best sensitivity value was reached when sigma is equal to two, but in this case, the specificity value was of 0.9840.

Since the database of production of eggs, has much more negative labels than positive ones, the specificity metric has more impact on misclassifications; from this approach, a value of sigma equal to five is better. Another approach to support this decision is stated by Fernández Pierna et al. (2006) who argue that the generalization ability increases while sigma gets higher values.

Modification of the parameter C generates slight increases for sensitivity, and minor changes for the rest of metrics. Given that high values of parameter C, can cause overfitting problems (Mucherino et al., 2009), a value of 0.15 was selected since it is the lowest value with higher sensitivity performance, among the evaluated.

Window size refers to the amount of data needed by the model to perform the classification task. Besides relevant features B, D and E, depends on the amount of data provided in order to calculate a single value for each feature, which constitutes a part of a pattern.

Our results showed that a window size equal to 14 generates the best results in all the performance metrics. A windows size of 7 days, did not provide enough data, consequently patterns differ among same labels. A windows size of over 28 days, grouped excessive data, thus patterns become similar between positive and negative labels.

Forecasting interval was assessed, in a value range from zero to five, the model performed an accuracy of 0.9874, specificity of 0.9876, sensitivity of 0.9783 and a positive predictive value of 0.6518, at a forecasting interval of zero, in this case, the model works as an early warning.

As the forecasting interval increases, the performance metrics decreases, in the case of the sensitivity, the forecasting interval affects it more intensely than to other metrics. In the authors' opinion, sensitivity values above 0.8 are acceptable. Therefore, the optimal forecasting interval is considered to be from zero to three days.

At optimal forecasting interval values, the model is able to identify the problem before it became apparent to the experts' judgement. The selection of either value will depend on how accurate, sensitive and specific the model is expected to perform.

In some instances, it was found that the model was able to detect as false positives, some days prior to an event occurring. Yet those days remained overlooked by the experts as no significant reduction had been observed.

5. Conclusions

In this work, optimal parameter configuration of an SVM classifier model is assessed by performance metrics, results clearly indicate that it is achievable to early warn problems in the curve of commercial laying hens.

Radial basis function kernel with a sigma value equal to 5, and a parameter C value of 0.15 is the one which achieved the best performance, that is 0.9874 for accuracy, 0.9876 for specificity, 0.9783 for sensitivity and 0.6518 for positive predictive value, as early warning at 0-day forecasting interval.

For this application, a window size equal to 14 generates the best results in all the performance metrics, by the modification of computed values of relevant features B, D and E, been part of input patterns.

It should be pointed out that the model has the ability to issue an alert with a sensitivity of 0.9333, 0.8738, and 0.8030, for one, two and three days respectively, before experts realized the drop of the production, the sensitivity decreases below 0.8 for greater forecasting intervals.

At farm level, an alert a day in advance, could be very helpful to decide performing a preventive diagnosis looking for clinical symptoms, or any other related issue in order to take actions for solving immediately.

6. Future developments

Future work is focusing on the use of these techniques to identify features that allow for early warning of specific poultry diseases, for which a new field with confirmed diagnosis can be included in the database. Time of egg collection, daily water and food consumption, sound patterns and thermal infrared images of the birds, could be added as fields to the database in order to improve the accuracy over longer intervals of time.

The early warning model, could be embedded in hardware or production management information software, and may have a major positive impact on the poultry industry, as it allows detecting and acting in time, and could reduce economic losses related to delayed treatments.

Acknowledgements

We gratefully acknowledge to DINTA-UTMACH, RNASA-UDC and Agrolomas CL, for providing all the resources for this research; our special thanks to the two anonymous reviewers whose suggestions helped to improve and clarify this manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compag.2015.12.009.

References

Adams, C.J., Bell, D.D., 1980. Predicting poultry egg production. Poult. Sci. 59, 937–938. http://dx.doi.org/10.3382/ps.0590937.

Ahmad, H.A., 2011. Egg production forecasting: determining efficient modeling approaches. J. Appl. Poult. Res. 20, 463–473. http://dx.doi.org/10.3382/japr.2010-00266.

Ahmadi, H., Golian, A., 2008. Neural network model for egg production curve. J. Anim. Vet. Adv., doi: javaa2008.1168.1170.

Alonso, J., Villa, A., Bahamonde, A., 2015. Improved estimation of bovine weight trajectories using support vector machine classification. Comput. Electron. Agric. 110, 36–41. http://dx.doi.org/10.1016/j.compag.2014.10.001.

Altman, D.G., Bland, J.M., 1994. Statistics notes: diagnostic tests 2: predictive values. BMJ 309. http://dx.doi.org/10.1136/bmj.309.6947.102, 102–102.

Benítez, R., Escudero, G., Kanaan, S., 2013. Inteligencia artificial avanzada. Editorial UOC, España.

Bennett, K.P., Campbell, C., 2000. Support vector machines. ACM SIGKDD Explor. Newsl. 2, 1–13. http://dx.doi.org/10.1145/380995.380999.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A Training Algorithm for Optimal Margin Classifiers. In: Proc. 5th Annu. ACM Work. Comput. Learn. Theory, pp. 144–152. http://dx.doi.org/10.1.1.21.3818..

Cameron, A., 2012. Manual of basic animal disease surveillance. Interafrican Bureau Anim. Resour.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. http://dx.doi.org/10.1007/BF00994018.

De Vries, A., Reneau, J.K., 2010. Application of statistical process control charts to monitor changes in animal production systems. J. Anim. Sci. 88, E11–E24. http://dx.doi.org/10.2527/jas.2009-2622.

Dohoo, I.R., 1993. Monitoring livestock health and production: service—epidemiology's last frontier? Prev. Vet. Med. 18, 43–52. http://dx.doi.org/10.1016/0167-5877(93)90092-8.

Faridi, a., Sakomura, N.K., Golian, A., Marcato, S.M., 2012. Predicting body and carcass characteristics of 2 broiler chicken strains using support vector regression and neural network models. Poult. Sci. 91, 3286–3294. http://dx.doi.org/10.3382/ps.2012-02491.

Felipe, V.P.S., Silva, M.A., Valente, B.D., Rosa, G.J.M., 2015. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. Poult. Sci. 94, 772–780. http://dx.doi.org/10.3382/ps/pev031.

Fernández Pierna, J.A., Baeten, V., Dardenne, P., 2006. Screening of compound feeds using NIR hyperspectral data. Chemom. Intell. Lab. Syst. 84, 114–118. http://dx.doi.org/10.1016/j.chemolab.2006.03.012.

Fernandez-Lozano, C., Canto, C., Gestal, M., Andrade-Garda, J.M., Rabuñal, J.R., Dorado, J., Pazos, A., 2013. Hybrid model based on genetic algorithms and SVM applied to variable selection within fruit juice classification. Sci. World J. 2013, 1–13. http://dx.doi.org/10.1155/2013/982438.

Grossman, M., Gossman, T.N., Koops, W.J., 2000. A model for persistency of egg production. Poult. Sci. 79, 1715–1724. http://dx.doi.org/10.1093/ps/79.12.1715.

Grossman, M., Koops, W.J., 2001. A model for individual egg production in chickens. Poult. Sci. 80, 859–867. http://dx.doi.org/10.1093/ps/80.7.859.

Han, D., Chan, L., Zhu, N., 2007. Flood forecasting using support vector machines. J. Hydroinformatics 9, 267. http://dx.doi.org/10.2166/hydro.2007.027.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer 2001. Springer Series in Statistics. Springer New York, New York, NY. http://dx.doi.org/10.1007/b94608.

Hemalatha, Muruganand, S., Maheswaran, R., 2014. Recognition of Poultry Disease in Real Time. In: Proceedings Int. Conf. Inter-Disciplinary Res. Eng. Technol. 2014.

Hempstalk, K., McParland, S., Berry, D.P., 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. J. Dairy Sci. 98, 5262–5273. http://dx.doi.org/10.3168/jds.2014-8984.

Hepworth, P.J., Nefedov, a.V., Muchnik, I.B., Morgan, K.L., 2012. Broiler chickens can benefit from machine learning: support vector machine analysis of observational epidemiological data. J. R. Soc. Interf. 9, 1934–1942. http://dx.doi.org/10.1098/rsif.2011.0852.

Hsu, C.-W., Chang, C.-C., Chih-Jen, L., 2003. A practical guide to support vector classification, pp. 1–16. http://dx.doi.org/10.1177/02632760022050997.

Jacob, J.P., Wilson, H.R., Miles, R.D., Butcher, G.D., Mather, F.B., 2014. Factors affecting egg production in backyard chicken. Univ. Florida, 1–8.

Lindsay, D., Cox, S., 2005. Effective probability forecasting for time series data using standard machine learning techniques, 35–44. http://dx.doi.org/10.1007/11551188_4.

Lokhorst, C., Lamaker, E.J.J., 1996. An expert system for monitoring the daily production process in aviary systems for laying hens. Comput. Electron. Agric. 15, 215–231. http://dx.doi.org/10.1016/0168-1699(96)00017-8.

Long, A., Wilcox, S., 2011. Optimizing Egg Revenue for Poultry Farmers, 1–10.

Martens, D., Baesens, B., 2010. Building Acceptable Classification Models, 53–74. http://dx.doi.org/10.1007/978-1-4419-1280-0_3.

McQueen, R.J., Garner, S.R., Nevill-Manning, C.G., Witten, I.H., 1995. Applying machine learning to agricultural data. Comput. Electron. Agric. 12, 275–293. http://dx.doi.org/10.1016/0168-1699(95)98601-9.

Mertens, K., Decuypere, E., De Baerdemaeker, J., De Ketelaere, B., 2011. Statistical control charts as a support tool for the management of livestock production. J. Agric. Sci. 149, 369–384. http://dx.doi.org/10.1017/S0021859610001164.

Mollazade, K., Omid, M., Arefi, A., 2012. Comparing data mining classifiers for grading raisins based on visual features. Comput. Electron. Agric. 84, 124–131. http://dx.doi.org/10.1016/j.compag.2012.03.004.

Mucherino, A., Papajorgji, P.J., Pardalos, P.M., 2009. Data Mining in Agriculture, Media, Springer Optimization and Its Applications. Springer New York, New York, NY. http://dx.doi.org/10.1007/978-0-387-88615-2.

Narinc, D., Uckardes, F., Aslan, E., 2014. Egg production curve analyses in poultry science. Worlds. Poult. Sci. J. 70, 817–828. http://dx.doi.org/10.1017/S0043933914000877.

Palma, J., Marín, R., 2013. Inteligencia artificial. Técnicas, métodos y aplicaciones. McGraw Hill, Murcia.

Roush, W.B., Tomiyama, K., Garnaoui, K.H., D'Alfonso, T.H., Cravener, T.L., 1992. Kalman filter and an example of its use to detect changes in poultry production responses. Comput. Electron. Agric. 6, 347–356. http://dx.doi.org/10.1016/0168-1699(92)90005-8.

Savegnago, R.P., Cruz, V.a.R., Ramos, S.B., Caetano, S.L., Schmidt, G.S., Ledur, M.C., El Faro, L., Munari, D.P., 2012. Egg production curve fitting using nonlinear models for selected and nonselected lines of White Leghorn hens. Poult. Sci. 91, 2977–2987. http://dx.doi.org/10.3382/ps.2012-02277.

Schaefer, A.L., Cook, N., Tessaro, S.V., Deregt, D., Desroches, G., Dubeski, P.L., Tong, a. K.W., Godson, D.L., 2004. Early detection and prediction of infection using infrared thermography. Can. J. Anim. Sci. 84, 73–80. http://dx.doi.org/10.4141/A02-104.

Suttorp, T., Igel, C., 2007. Artificial Neural Networks – ICANN 2007, Artificial Neural Networks – ICANN 2007, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-74690-4.

Tang, Y., Zhang, Y.Q., Chawla, N.V., 2009. SVMs modeling for highly imbalanced classification. IEEE Trans. Syst. Man, Cybern. Part B Cybern. 1, 1–9. http://dx.doi.org/10.1109/TSMCB.2008.2002909.

Vapnik, V.N., Kotz, S., 1982. Estimation of Dependences Based on Empirical Data. Springer-Verlag, New York.

Venkatesan, M., Thangavelu, A., Prabhavathy, P., 2013. Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), Advances in Intelligent Systems and Computing, Advances in Intelligent Systems and Computing. Springer India, India. http://dx.doi.org/10.1007/978-81-322-1041-2.

Woodall, W.H., Tech, V., 2006. The Use of Control Charts in Surveillance. J. Qual. Technol. 38, 89–104.

Woudenberg, S.P.D., Gaag, L.C., Van Der, Feelders, A., Elbers, A.R.W., 2014. Real-time Adaptive Problem Detection in Poultry. In: Press, I. (Ed.), ECAI 2014, pp. 1217–1218. http://dx.doi.org/10.3233/978-1-61499-419-0-1217.

Zhao, J., Lin, H., Chen, Q., Huang, X., Sun, Z., Zhou, F., 2010. Identification of egg's freshness using NIR and support vector data description. J. Food Eng. 98, 408–414. http://dx.doi.org/10.1016/j.jfoodeng.2010.01.018.

Zhiliang, L., Zuo, M., Zhao, X., Xu, H., 2015. An analytical approach to fast parameter selection of Gaussian RBF Kernel for support vector machine ∗. J. Inf. Sci. Eng. 710, 691–710.

Chemometrics and Intelligent Laboratory Systems 159 (2016) 45–57

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Optimization of NIR calibration models for multiple processes in the sugar industry

Iván Ramírez-Morales[a,b,*], Daniel Rivero[b], Enrique Fernández-Blanco[b], Alejandro Pazos[b]

[a] Universidad Técnica de Machala, Faculty of Agricultural & Livestock Sciences, 5.5 km Pan-American Av, Machala, El Oro, Ecuador
[b] Universidade A Coruña, Department of Computer Science, 15071 A Coruña (03082), Spain

ARTICLE INFO

Keywords:
NIR
Chemometrics
Calibration models
Machine learning
Support vector machines
Agro-industry

ABSTRACT

The measurements of Near-Infrared (NIR) Spectroscopy, combined with data analysis techniques, are widely used for quality control in food production processes.

This paper presents a methodology to optimize the calibration models of NIR spectra in four different stages in a sugar factory. The models were designed for quality monitoring, particularly °Brix and Sucrose, both common parameters in the sugar industry.

A three stage optimization methodology, including pre-processing selection, feature selection and support vector machines regression metaparameters tuning, were applied to the spectral data divided by repeated cross-validation. Global models were optimized while endeavoring to ensure they are able to estimate both quality parameters with a single calibration, for the four steps of the process.

The proposed models improve the prediction for the test set (unseen data) compared to previously published models, resulting in a more accurate quality assessment of the intermediate products of the process in the sugar industry.

1. Introduction

The production flow in the sugar industry encompasses several processes and subprocesses that need to be analyzed in order to maintain a quality standard [1]. The agro-industrial plants require cost-efficient and non-destructive systems to monitor the quality of their production process, food safety and compliance with the technical specifications [2].

One of these non-destructive systems aimed at ensuring quality is chemometrics, which has been developing since the 1970s as an interdisciplinary field of study. This field covers a wide and varied range of mathematical and statistical techniques for analyzing the chemical composition of materials [3].

To analyze the quality of organic raw materials, a commonly-used technique is the Near-Infrared Reflectance (NIR) spectroscopy, associated with chemometrics; however, the relationship between the absorption in the spectral region of the near infrared and the analyte is frequently of a non-linear type [4].

The origin of these non-linear relationships is diverse and difficult to identify, in some cases due to the differences in viscosity, temperature, pH, particle size and the chemical nature of the analyte. For this reason, calibration is commonly performed using non-linear methods and multivariate analysis [5]. A proper selection of the variables aimed at gathering a small subgroup with lower sensitivity to non-linearities or at discarding the most pronounced wavelengths is usually effective to improve the performance of the models [6,7].

Chemometrics is an essential part of NIR spectroscopy in the food sector [28], recently, with the development of information technologies, the applications of NIR spectroscopy have become increasingly popular and arousing great interest among researchers, considering that the technique is able to detect analyte concentrations of 0.1% w/w [8].

Common chemometric techniques use multivariate analysis methods, such as PCA as a qualitative analysis technique of the spectral data, and PLS regression analysis as a technique for quantitative prediction of the parameters of interest in the sample [2]. Currently, the literature that applies machine learning techniques in chemometrics is in constant expansion [9–12], the use of artificial neural networks, support vector regression (SVR) is also reported [5,24,30] as these techniques are based on pattern recognition [11].

The food industry has widely used NIR spectroscopy [26] since it is a rapid, accurate, minimally invasive, non-destructive quality analysis technique [2]. NIR has been used to analyze quality of dairy products [13,14], oils [14], meat products [15], fish [16], cereals [17] and fruit

* Corresponding author at: Universidad Técnica de Machala, Faculty of Agricultural & Livestock Sciences, 5.5 km Pan-American Av, Machala, El Oro, Ecuador.
E-mail address: iramirez@utmachala.edu.ec (I. Ramírez-Morales).

http://dx.doi.org/10.1016/j.chemolab.2016.10.003
Received 1 July 2016; Received in revised form 20 September 2016; Accepted 5 October 2016
Available online 14 October 2016
0169-7439/

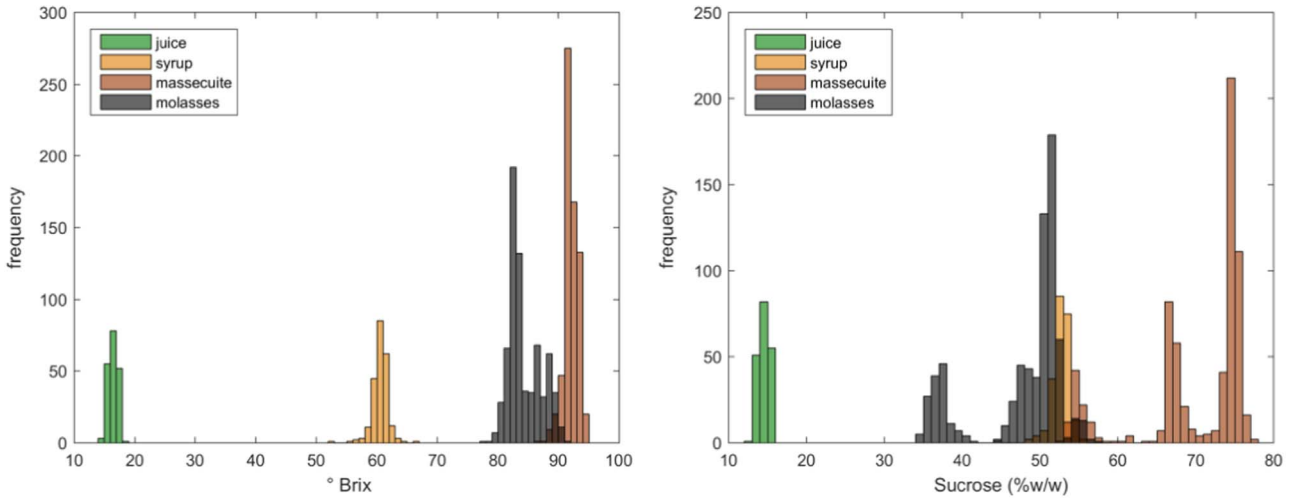


Fig. 1. Histograms of °Brix and Sucrose content for each process step.

[18].

In the sugarcane industry, studies were found which proved a good correlation between the NIR spectra and quality indicators of sugarcane [19]. The use of NIR spectra preprocessing techniques in sugar cane was analyzed [20], along with the selection of features [21] and chemometric algorithms in order to improve prediction of target analytes [22,23].

A recent study conducted by Tange et al. [24] has shown that the use of calibration models with Support Vector Machines (SVM) for regression is efficient in order to predict °Brix and Sucrose values, the quality parameters of the industrial process of sugar. The use of SVM improves in terms of *RMSE* compared to the technique of Partial Least Square (PLS); however, the proposed model uses the entire NIR spectrum, which leads us to believe that the optimization of the model is still possible by implementing an appropriate preprocessing technique, feature selection and optimization of the parameters of the machine support vectors.

The aim of this study is to optimize global calibration NIR models in order to improve quality control of the °Brix and Sucrose parameters. A global calibration NIR model is capable of predict a value in the four steps of the sugar production process.

2. Materials and methods

Generally, the sample processing consists of the following steps: acquisition of spectral data, preprocessing of the data to reduce the noise, thereby increasing the signal-to-noise ratio (S/N) [25], selection of relevant features, and development of the calibration model using a set of spectra from which the values of target analytes obtained by reference techniques are known, and, finally, the model validation using data different to those of calibration [8].

2.1. Data description

The database was published by Tange et al. [24], and the employed data were obtained in a Japanese sugar factory (Daito Togyo Co), where sugar cane is processed. The samples were obtained throughout three months in the harvest season, and during each of the process steps: after grinding (juice), after the process of evaporation (syrup), after crystallization (massecuite) and after centrifugation (molasses), three cycles of crystallization and centrifugation were carried out, resulting in a higher number of samples of massecuite and molasses than those obtained in the other two stages. Immediately after sampling, the NIR signals were extracted, along with the reference technique in relation to process temperature.

The °Brix quality parameter expresses all the dissolved solids (sugar and non-sugar) as a percentage of total weight, its scale reflecting the percentage of sucrose in pure solutions. In any sugar materials (juice, honey, etc.) the °Brix parameters are always higher than those of Sucrose, whereas in high-purity materials, such as spirits from a refinery, the difference between these indicators is minimal. For the current work, the °Brix quality parameter was measured using an Abbemat-WR refractometer, developed by Anton Paar GmbH in Germany.

The Sucrose quality parameter (POL) refers to the amount of sucrose contained in a solution, expressed as percent of the weight; in pure solutions, the POL percentage is equivalent to the percentage of Sucrose, while in other impure solutions, like cane juice and honey, there is a difference between these two values, the more impure is the solution, the higher the difference. For this reason, the POL value is internationally accepted as apparent sucrose. In the current paper, Sucrose was measured using a MCP500 polarimeter, developed by Anton Paar GmbH in Germany.

2.2. Near infrared reflectance

NIR spectra are obtained as a result of vibrational transitions primarily associated with chemical bonds containing hydrogen, C-H, N-H, S-H and O-H, and which are present in most organic compounds; the NIR spectrum region spans within the wavelength range between 780 nm and 2500 nm [2,27,28].

The NIR spectrum, at all wavelengths, may contain information regarding several analytes [30] and other physical features as temperature, viscosity, crystals and pH. This implies that the resulting spectrum is the consequence of the modifications carried out simultaneously in all the analytes in the sample, making the calibration process more complicated [5,27,30].

A total of 1797 NIR spectra ranging between 400.0 nm and 1888 nm, with an increase of 2 nm, were obtained using an NIR DS2500 spectrometer, developed by FOSS AB in Denmark. In the work presented by Tange et al. [24], the spectral signals with an absorbance value greater than two were excluded from the analysis; in contrast, the current study used the full dataset.

Fig. 1. shows the histograms of spectra according to their content of °Brix and Sucrose content for each process step, for the °Brix data, a pronounced separation in the frequency distributions is observed, with a slight overlapping between massecuite and molasses in the range around 90 °Brix. On the other hand, the sucrose measurements show overlapping between syrup, massecuite and molasses, in the range around 50% w/w, of sucrose.

2.3. Preprocessing techniques

The near-infrared (NIR) spectral data preprocessing is an integral part of chemometric modeling. The purpose of preprocessing is to remove physical phenomena from the spectra [27,29,31].

The NIR spectra of solid samples are influenced by the physical properties of the samples, making preprocessing increasingly important to minimize the contributions which contain irrelevant information, therefore, simpler and more robust models could be developed [27,29]. The proper choice of the preprocessing technique is difficult to assess before the validation of the model, therefore, NIR spectra preprocessing is still carried out through trial and error [31,32].

NIR spectroscopy has led to both a greater number and diversity of preprocessing techniques, primarily because the spectra may be significantly influenced by the non-linearities introduced by the light scattering [31]. This study applies four basic techniques for the preprocessing combinations of NIR spectra: the Beer-Lambert law, the First Spectral Derivative, Standard Normal Variate, and detrending.

2.3.1. Beer-Lambert law

is empirical for NIR spectra and suggests a linear relationship between the absorbance of spectra and concentration(s) of the constituent; this law is valid only for systems of pure transmittance without scattering; in reflectance measurements, the law can be expressed as follows [31]:

$$A\lambda = -\log_{10}(R) \cong \epsilon\lambda \times l \times c$$

Where $A\lambda$ is the absorbance dependent on wavelength, R is the reflectance detected, ελ is the molar absorptivity dependent on wavelength, l is the effective length of the light path through the sample matrix, and c is the concentration of the component of interest.

According to this law, the spectra is processed by taking the base 10 negative logarithmic calculation of the sample's reflectance, which results in the linear relationship to the concentration of the analyte.

2.3.2. First Spectral Derivative (FSD)

First Spectral Derivative (FSD) have the ability to remove both additive and multiplicative effects from the spectra and have been used in analytical spectroscopy for decades [27]. The most basic method for derivation is finite differences [31]; the first order derivative is calculated as the difference between two subsequent spectral measurement points:

$$X_{i,fsd} = X_i - X_{i-1}$$

Where $X_{i,fsd}$ denotes the first order derivative at wavelength i. This technique removes only the baseline of the spectra.

2.3.3. Standard Normal Variate (SNV)

is possibly the second most applied method for scattering correction of NIR data [33]. The basic format for SNV, correction and normalization is as follows:

$$X_{i,snv} = \frac{X_i - \bar{x}}{S}$$

Where $X_{i,snv}$ denotes the SNV at a wavelength i, $\bar{x}$ is the spectrum average of the sample to be corrected, and S is the standard deviation of the spectrum sample.

2.3.4. Detrending technique

is applied to the spectra to remove the effects of the baseline and curvilinearity changes, it is typical of NIR spectra to which the preprocessing technique based on the Beer-Lambert law was applied. This effect is generally linear [34]. The method consists of modeling the baseline as a linear function of the wavelength and this function is subsequently subtracted from each spectrum value independently. The expression is as follows:

$$X_{i,dt} = X_i - b_i$$

Where $X_{i,dt}$ denotes the detrended spectrum at a wavelength i, b_i is the linear model baseline of the spectrum in the wavelength i.

In general, the combination of techniques is not advisable, and, as a minimum requirement, preprocessing should maintain or reduce the complexity of the effective model [31]. In contrast to the above-stated criteria, Xu et al. [32] showed that the combination of preprocessing methods improved the stability of the models and the results in terms of *RMSE*, as it takes advantage of the complementary information given by each preprocessing method, therefore, the stability of the models and the results are improved in terms of *RMSE*. Other recent studies have obtained good results by combining preprocessing techniques [35–38]. In our work, preliminary tests with commonly used preprocessing methods were performed, and based on the results of these tests, nine combinations were defined, as explained below:

1. Raw: unprocessed reflectance signal (as extracted from the instrument).
2. SVN to Raw: the Standard Normal Variate was calculated for the raw signal.
3. SNN to Raw & detrend: a technique of detrending is applied to the processed signal in Combination 2.
4. Beer Lambert (BL): the Beer-Lambert law is applied to the raw signal.
5. SNV to BL: calculation of SNV for the processed signal in Combination 4.
6. SNV to BL & detrend: a technique of detrending is applied to the processed signal in Combination 5.
7. First spectral derivative (FSD): the calculation of the first spectral derivative from the raw signal.
8. SNV to FSD: calculation of SNV for the processed signal in Combination 7.

SNV to FSD & detrend: a technique of detrending is applied to the processed signal in Combination 8.

Fig. 2 shows nine alternative NIR spectra transformations that result from combining the above-mentioned basic preprocessing techniques, which will be evaluated in the first stage. In order to better illustrate this section for the reader, spectra were colored according to their respective °Brix values; it should also be noted how the differences are softened and accentuated between the different shades in the NIR spectrum, when applying preprocessing techniques.

2.4. Feature selection techniques

Due to the large amount of spectral information provided by NIR spectrophotometers, a substantial reduction of the number of samples needed to build the classification and calibration models is required [29]. Over the past decade, in the construction of models, feature selection (FS) has passed from illustrative examples regarding its operation to becoming a requirement, particularly due to the nature of the high-dimensionality problems, such as microarray and spectral analyses [7].

Considering that many of the pattern recognition techniques were not originally designed to deal with large amounts of information with little relevance, the application of FS techniques has currently become a necessity in many applications [39]. Its application prevents overfitting of the model, improves performance and reduces the computation time, obtaining a deeper understanding of the data [7,40,41]. FS does not alter the original representation of the variables, since it simply consists of selecting a subset of the best features, preserving their original nature, allowing them to be easily interpretable by an expert in the field [7].

A group of techniques, commonly used for FS, are filters which

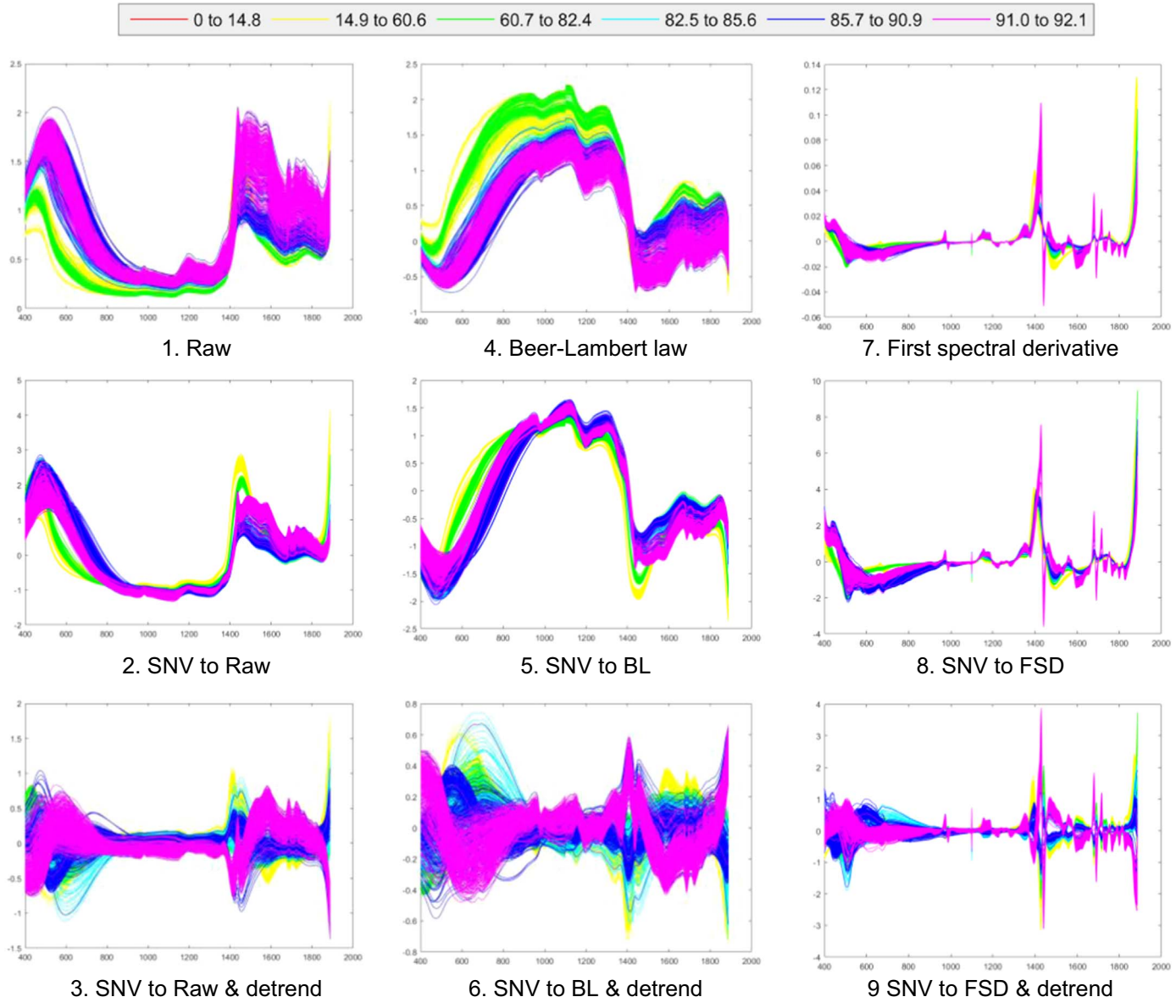


Fig. 2. Changes in the NIR spectra when applying the preprocessing techniques.

evaluate the relevance of a feature, analyzing the intrinsic properties of the data; generally, a score of relevance is calculated and the characteristics with lower score are eliminated [7,42]. There are two types, univariate and multivariate filters [43]. The univariate filter is a simple but efficient paradigm, which is fast, scalable and independent of the classification/regression technique [44]. A threshold method is usually defined to select the filters which meet a condition above or below the threshold. Filters work regardless of the model and use the intrinsic properties of the data [42].

The univariate filter method can be applied using a t-test [45], an F-test [46] or the Wilcoxon signed-rank test [42], by calculating a p-value which represents the statistical significance of each variable in the model, thus variables are organized depending upon their p-value [7].

2.5. Support vector regression

There are two types of machine learning algorithms: supervised and unsupervised; the former are used when there is prior knowledge about the desired outputs of the model, while the latter generate similar groups without having prior information [47]. Once a machine learning algorithm is trained, it is able to transfer what it learned to new data [48], that is, to new data that are not used during training, this set allows evaluating the generalization error of the final model chosen.

Support vector machines (SVM) are supervised machine learning algorithms [49], based on the structural risk minimization, and can be used in classification and regression (SVR) problems. Their operation starts from a set of training patterns whose outputs are known, and which allow making predictions about new patterns [50,51].

The principles of SVM were developed by Vapnik and Chervonenkins in 1963, in a study on statistical learning theories that aimed to reduce the error generalization according to the complexity of the search space [52]. The current standard of SVM was proposed by Cortes and Vapnik [53]. The purpose of SVM is to obtain models which structurally present little risk of error with respect to future data. Although originally designed to solve binary (two classes) classification problems, their application was extended to regression, multiclass classification, clustering and other tasks [51].

A version of an SVM for regression was proposed in 1997 by [54]. This method is called Support Vector Regression (SVR). The model depends only on a subset of data (support vectors), because the cost function for the construction of the model does not consider the points that are beyond the margin; in addition, the cost function ignores any

data that are close to the prediction model, within a threshold ε [50].

SVRs have been applied in several fields, such as time series [55], finances [56], engineering approaches in complex analyses [57], and convex quadratic programming [58], among others [50].

SVRs are learning methods based on kernel, which makes it possible to perform a transformation of data space, so that data could be described by linear models and the problem could be simplified [59]. The kernel is one of the most important SVR parameters; in the current study, some preliminary tests were carried out by trial and error using the most common kernels for SVR, in accordance with [60]. However, only the radial basis function kernel generated acceptable results, thus it was used to optimize its parameters.

Searching for optimal parameters of an SVR is essential in building a prediction model that is accurate and stable [61,62]. Kernel parameters are adjustable in the SVRs to control the complexity of the resulting hypothesis and to avoid the overfitting of the model [63,64]. The optimization of the parameters C (regularization parameter), γ (gamma RBF kernel parameter) and ε (epsilon parameter) is a key step in an SVR, since their combined values determine the complexity of the limits and therefore the performance of the model. Different techniques may be used to optimize these parameters [64,65].

2.6. Experimental optimization

The spectral data were divided into training (calibration) and test subsets using repeated cross-validation [66,67]. A 10-fold repeated cross-validation technique was chosen, in which data were divided into 10 groups, 9 being used as calibration sets and the remaining one as a test set; the test set was changed until all groups have been tested. Cross-validation was repeated 100 times. Support Vector Regression (SVR) was used with a Radial Basis Function (RBF) kernel [68].

There are many approaches to optimize simultaneously the preprocessing and the feature selection in order to improve the prediction ability, one of them is genetic algorithm approach. Devos and Duponchel, [69] found that all cooptimizations converge to the same kind of solutions; another new strategy is described for the combined implementation of variable, pre-processing and sample selection using algorithms like ant optimization colony and genetic algorithms [70]. The evaluation of the parameters in this work was performed using a t-test univariate filter [45], and a grid search method [59,71]. The evaluation was conducted in three stages:

2.6.1. First stage

The aim of this phase was to define what processing technique is most appropriate, and which wavelengths (features) should be selected. To this end, the percentile p-value threshold used in the FS process was evaluated for the nine preprocessing techniques, with fixed values of parameter C equal to 100, and both γ and ε parameters equal to 0.1.

2.6.2. Second stage

The objective of this phase was to find an optimal combination of the C and γ parameters. To achieve this, both were simultaneously optimized, using the grid search method [59,71] on a logarithmic scale. Parameter C was evaluated within the range from 10^1 to 10^6 and γ within the range of 10^{-1} to 10^{-3}, whereas the evaluation interval was 0.25 in the exponent, with a fixed value for ε equal to 0.1.

Grid search is a brute-force approach for parameter optimization which is widely used in machine learning techniques. Once the ranks and measures of the parameters are defined, each combination of parameters is tested in order to find the best one, based on a performance measure [71].

2.6.3. Third stage

The objective of this phase was to tune the performance of SVR. For this, the ε parameter of SVR was optimized within the range of 0 to 1 on linear scale, whereas the evaluation interval was 0.01.

At each stage, 100 repetitions for each CV were performed. In total, 478,000 SVRs were evaluated in order to determine the optimal configuration of the calibration model.

All models were evaluated using as performance measures the r-squared coefficient of determination (R^2) and root mean square error (*RMSE*) in the cross-validation test data as indicators of the accuracy of the model based on NIR [27,72]. The equation for *RMSE* is:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{i=n}(y_i-\hat{y}_i)^2}{n}}$$

Where y_i is the prediction value of the i-th observation, $\hat{y}_i$ is the measured value of the observation, and n is the number of observations.

The *RMSE* consists of the differences between the values predicted by a model and the values actually observed. This technique is often preferred over others, since its interpretation is on the same scale as the data, and it gained popularity because of its theoretical relevance in statistical modeling [73,74]. Mean and standard deviation of 100 repetitions in 10 fold cross-validation, were calculated in order to present the results in graphics and tables.

3. Results

This study presents a methodology which optimizes NIR calibration models for multiple processes in the sugar industry. This consists of a sequence of steps described below:

As shown in Fig. 3, the choice of the most appropriate preprocessing technique and the feature selection which best models the non-linearities of the spectrum is performed in parallel. At this point, the model is evaluated and its optimal values are set, in order that the C and γ parameters of SVR can be optimized using a grid search technique. Using these optimized values, a tuning of the parameter ε of SVR which defines the final model is carried out, and each evaluation (e) point corresponds to the evaluation phases outlined in Section 3.

3.1. First stage: processing technique and feature selection

The nine alternative preprocessing techniques proposed were evaluated at different intensities of feature selection, which was performed by changing the values of threshold percentile for the p-value of the T-test technique for feature selection.

Fig. 4 shows the results of *RMSE* in CV obtained with nine preprocessing techniques at different values of percentile p-value for the T-test feature selection, in which it is observed that both in the case of °Brix and Sucrose, the preprocessing technique number 9 (listed in Fig. 2) is the one providing the best results. The lowest values of *RMSE* are recorded within the percentile range between 1 and 17 for °Brix,

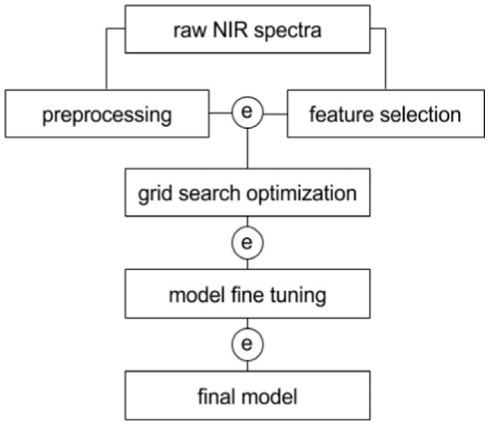


Fig. 3. Methodology for the optimization of NIR calibration models used in the current study.

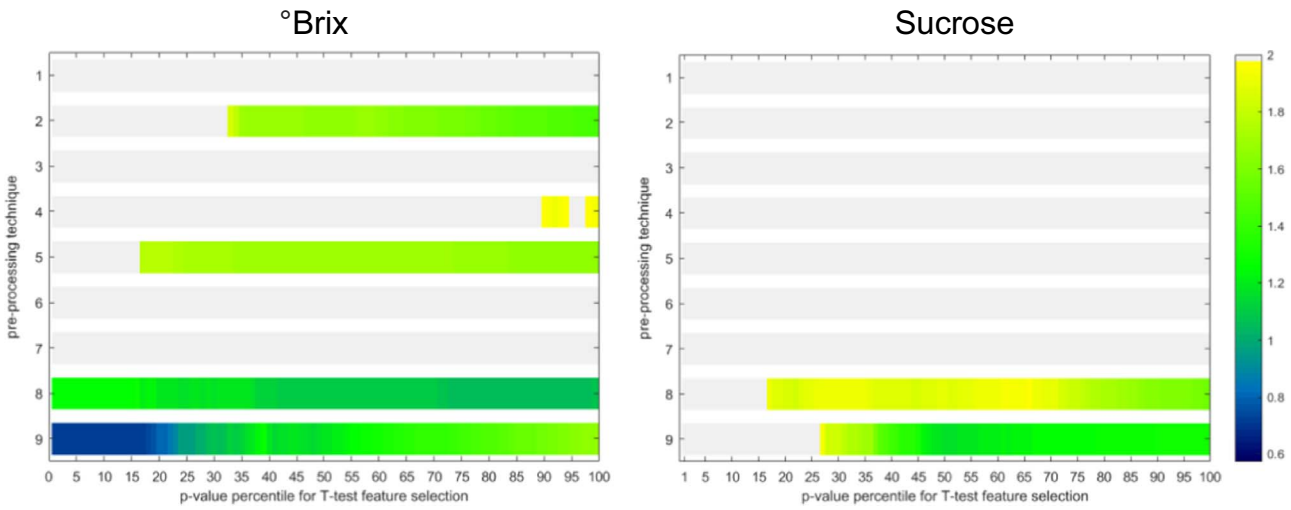


Fig. 4. RMSE in CV obtained with the nine preprocessing techniques at different values of percentile p-value for the T-test feature selection.

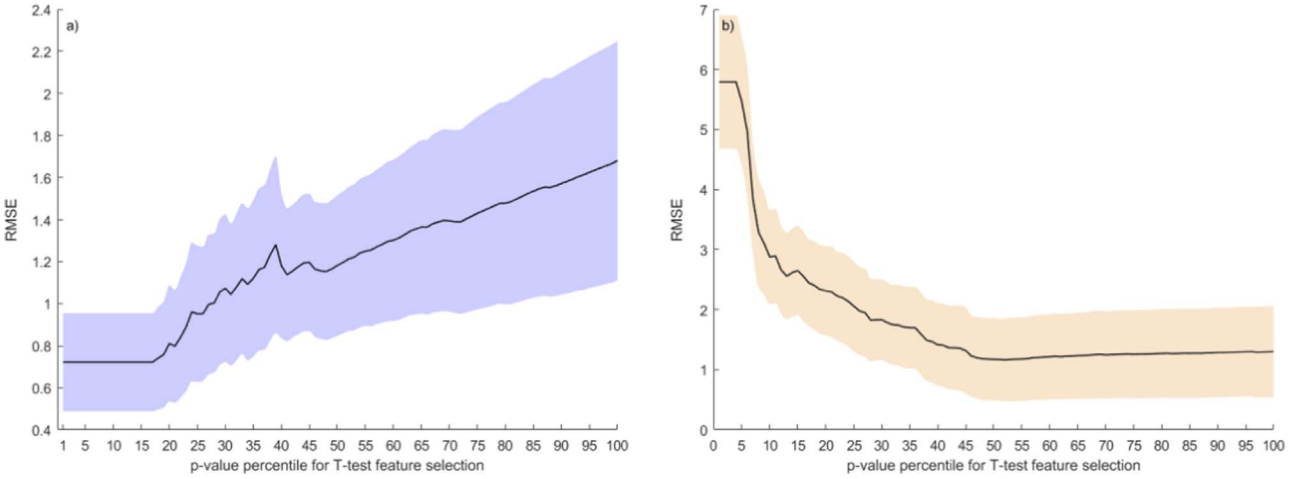


Fig. 5. *RMSE* in CV obtained in the model a) °Brix and b) Sucrose, with the preprocessing technique number 9, at different values of percentile p-value for the T-test feature selection.

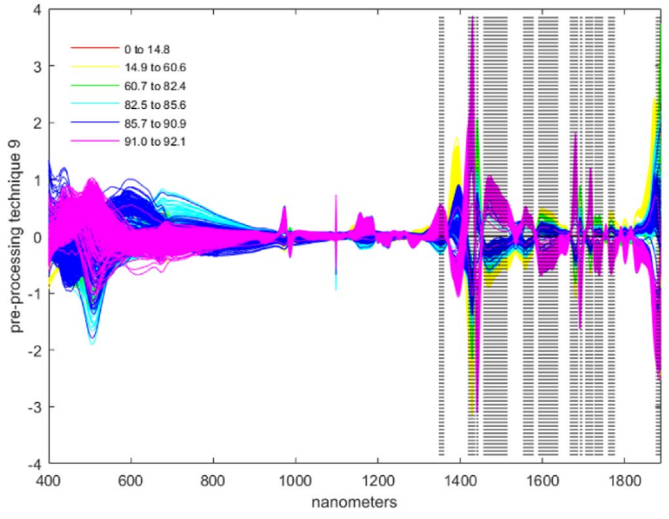


Fig. 6. Spectral bands (features) selected to build the prediction model for °Brix.

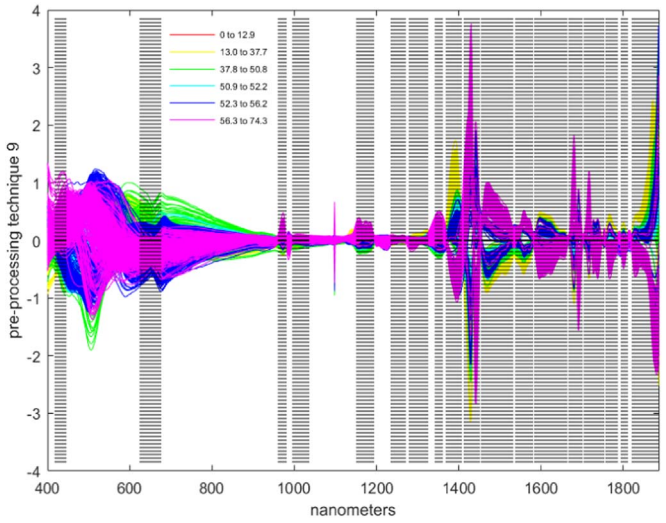


Fig. 7. Spectral bands (features) selected to build the prediction model for Sucrose.

while for Sucrose, the lowest *RMSE* values were recorded within the percentile range between 45 and 55. The preprocessing technique number 9 consists of combining the calculation of the first derivative, which is subsequently normalized with the SNV technique and finally the trend is extracted.

Fig. 5 shows the mean and standard deviation *RMSE* in CV obtained with the preprocessing technique number 9, at different values of percentile p-value for the T-test feature selection. Note that in the case of the °Brix model, the optimal values of percentile p-value for the T-test feature selection are between 1 and 17, a value equal to

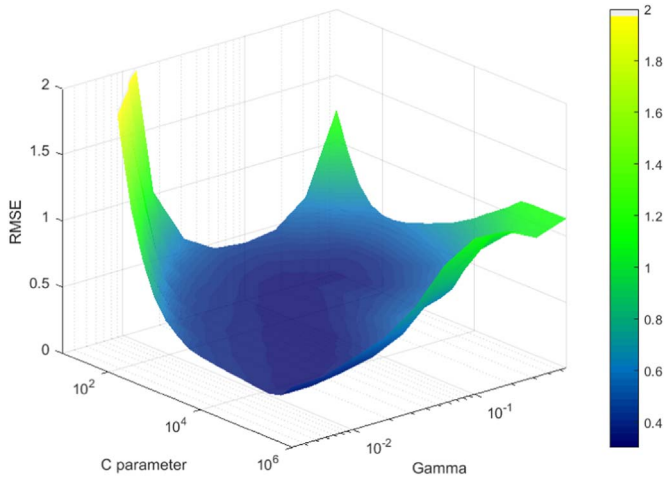


Fig. 8. Surface chart of *RMSE* in cross-validation of the prediction model for °Brix according to the values of parameters *C* and γ in all steps of the process.

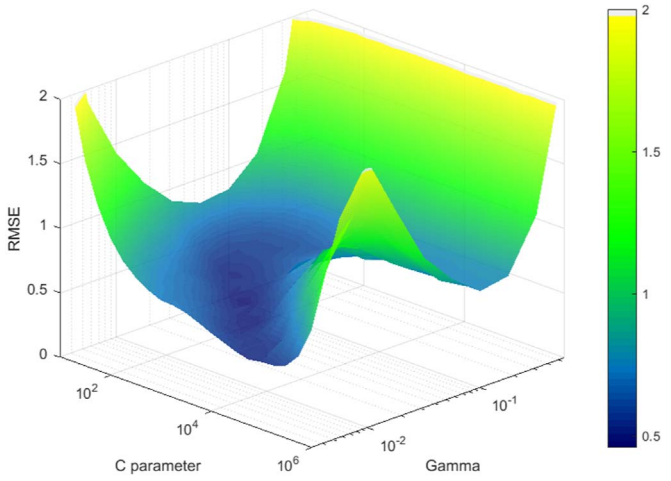


Fig. 10. Surface chart of *RMSE* in cross-validation of the prediction model for Sucrose according to the values of parameters C and γ in all steps of the process.

10 is selected, since it has a good mean value of *RMSE*; the Sucrose model has an optimal value of percentile p-value for the T-test feature selection equal to 52.

The selected features with a percentile p-value for the T-test feature selection equal to 10 are shown in Fig. 6, where the most relevant features for the °Brix model are highlighted in vertical, shaded bands.

The selected features with a percentile p-value for the T-test feature selection equal to 52 are shown in Fig. 7, where the most relevant features for the Sucrose model are highlighted in vertical, shaded bands.

3.2. Second stage: Optimization of C and γ parameters of SVR

After applying the preprocessing technique and selecting the

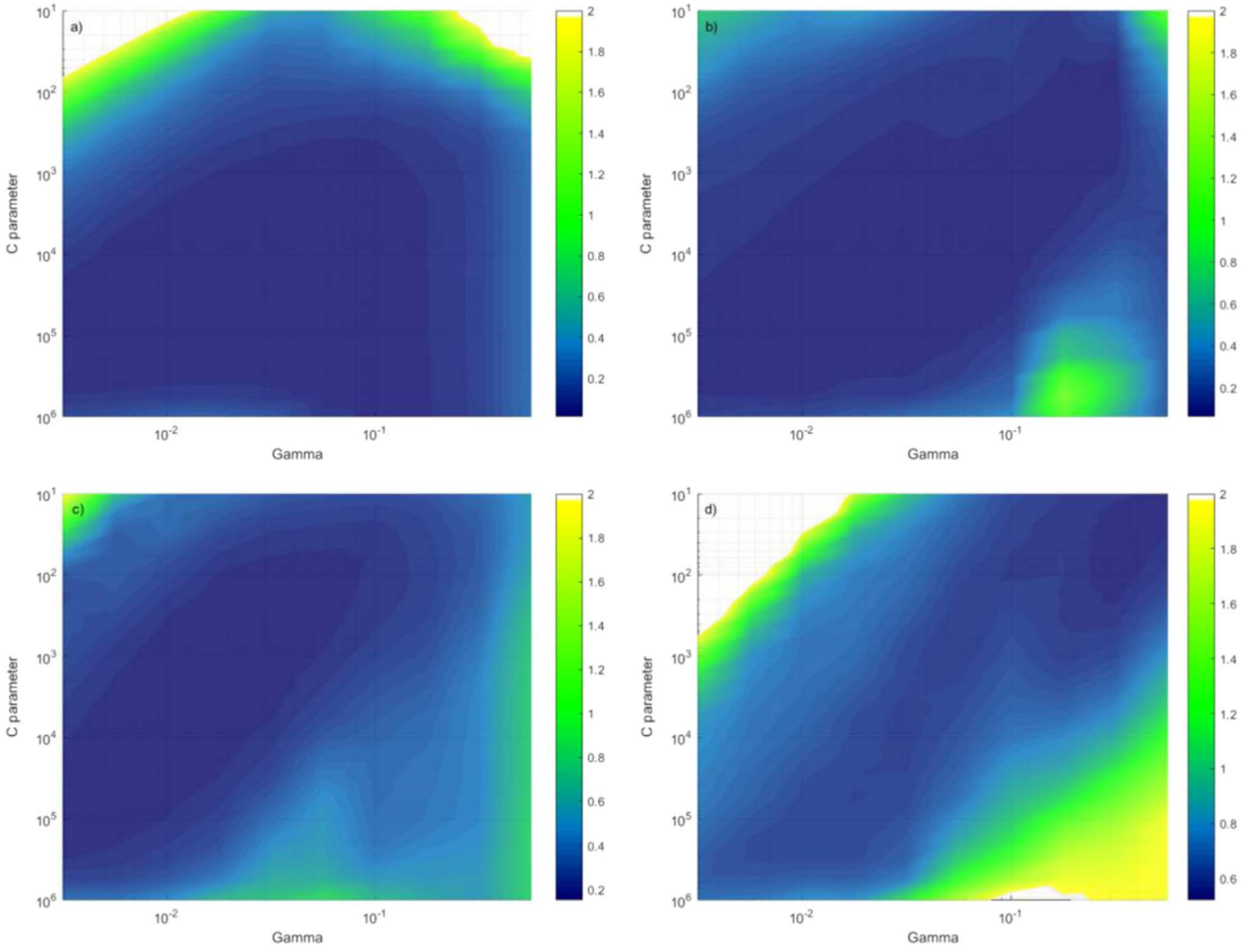


Fig. 9. Heat maps of *RMSE* in cross-validation of the prediction model for °Brix according to the values of parameters C and γ in the 4 steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

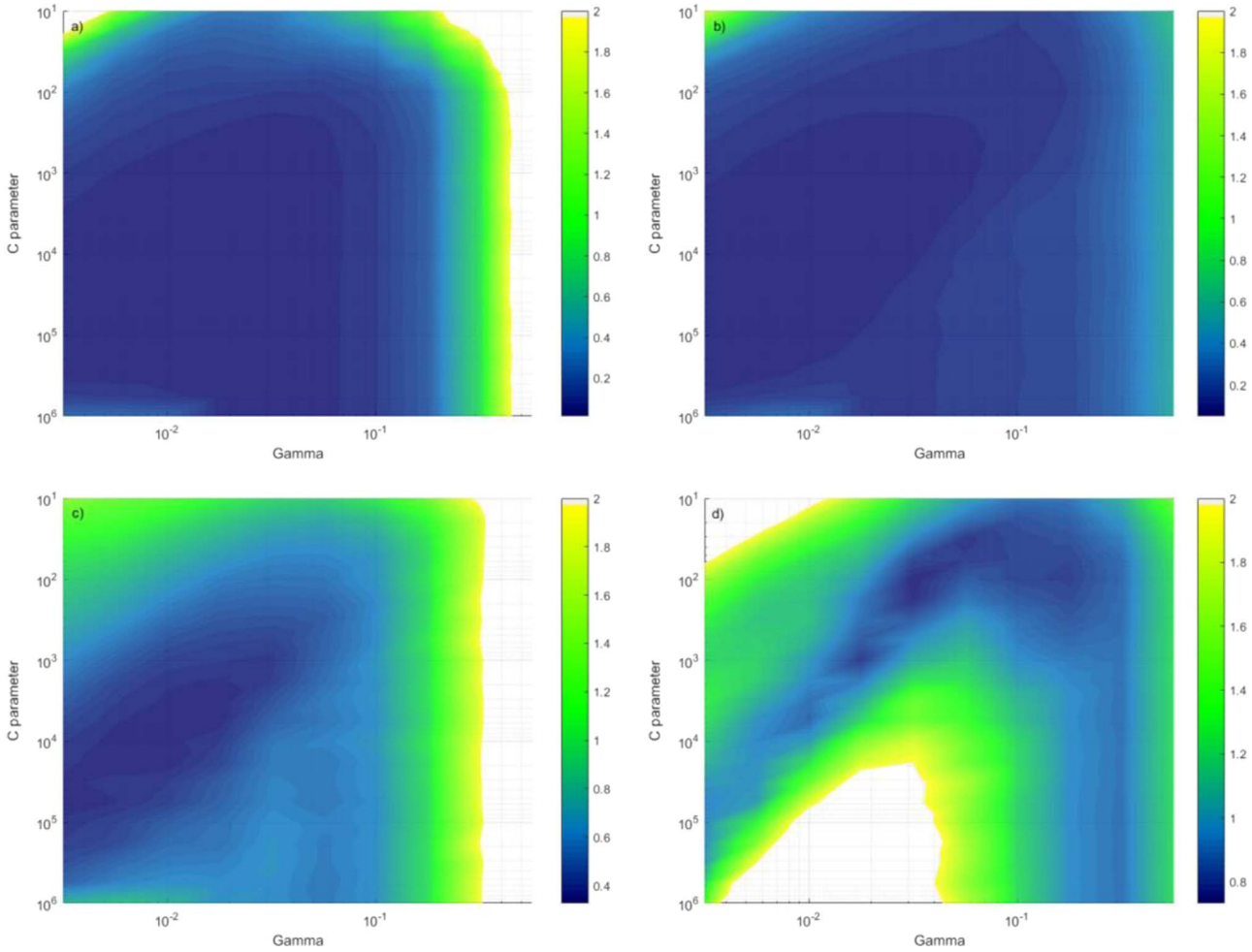


Fig. 11. Heat maps of *RMSE* in cross-validation of the prediction model for Sucrose according to the values of parameters C and γ in the 4 steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

wavelengths (features), the calibration models were evaluated in order to find the optimal combination of parameters C and γ, for a fixed ε value equal to 0.1.

The grid search method [43,50] was applied in logarithmic scale, parameter C was evaluated within the range from 10^1 to 10^1, γ within the range from 10^{-1} to 10^{-3}, and the evaluation interval was 0.25 in the exponent.

Fig. 8 shows a surface chart with the *RMSE* values obtained by the global model of °Brix for different combinations of parameters *C* and γ; the optimal parameters are 10e+2.75 and 10e-1.25 respectively, with those that together generate the lowest *RMSE* being evaluated in the average test results of the repeated cross-validation.

Fig. 9 details the performance of the global model at each step of the process, and it confirms that the optimal parameter values are those presented above.

Fig. 10 shows a surface chart with the *RMSE* values obtained by the global model of Sucrose for different combinations of parameters *C* and γ; the optimal parameters are 10e+3 and 10e-1.75 respectively, with those that together generate the lowest *RMSE* being evaluated in the average test results of the repeated cross-validation.

Fig. 11 details the performance of the global model at each step of the process, and it confirms that the optimal parameter values are those presented above.

3.3. Third phase: optimization of parameter ε

Once the preprocessing technique, the wavelengths (features) and the optimal combination of parameters C and γ were selected, the parameter ε of SVR was evaluated within the range of 0–1 in the linear scale, and the evaluation interval was 0.01.

Fig. 12 shows the curves of mean and standard deviation of *RMSE* for the optimization of the ε parameter in the prediction model for °Brix; it can be observed that a value of ε equal to 0.16 optimizes *RMSE* in the four steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

Fig. 13 shows the curves of mean and standard deviation for the optimization of the parameter ε in the prediction model for Sucrose; it can be observed that the optimal ε value is 0.07. However, in the case of molasses, the optimal ε value is 0.51; for the discussion section, the models optimized with both values will be evaluated, to consider an improvement in the estimation of Sucrose in the process step of molasses.

Finally, to statistically evaluate the performance of models whose parameters were optimized, R-squared, Adjusted R-squared and p-value were calculated. The model for °Brix was analyzed with its optimal parameters, in the case of Sucrose, the model was analyzed with the option of ε equal to 0.07.

Fig. 14 shows two regression graphs (actual vs predicted); the calibration model for °Brix obtained an R-squared of 0.99, an Adjusted R-Squared of 0.99 with a p-value < 0.01, whereas the calibration model for Sucrose obtained an R-squared of 0.99, an Adjusted R-Squared of 0.99 with a p-value < 0.01. For both models, it was determined that there was a highly statistically significant correlation between the actual and the predicted values, strengthening the importance of our

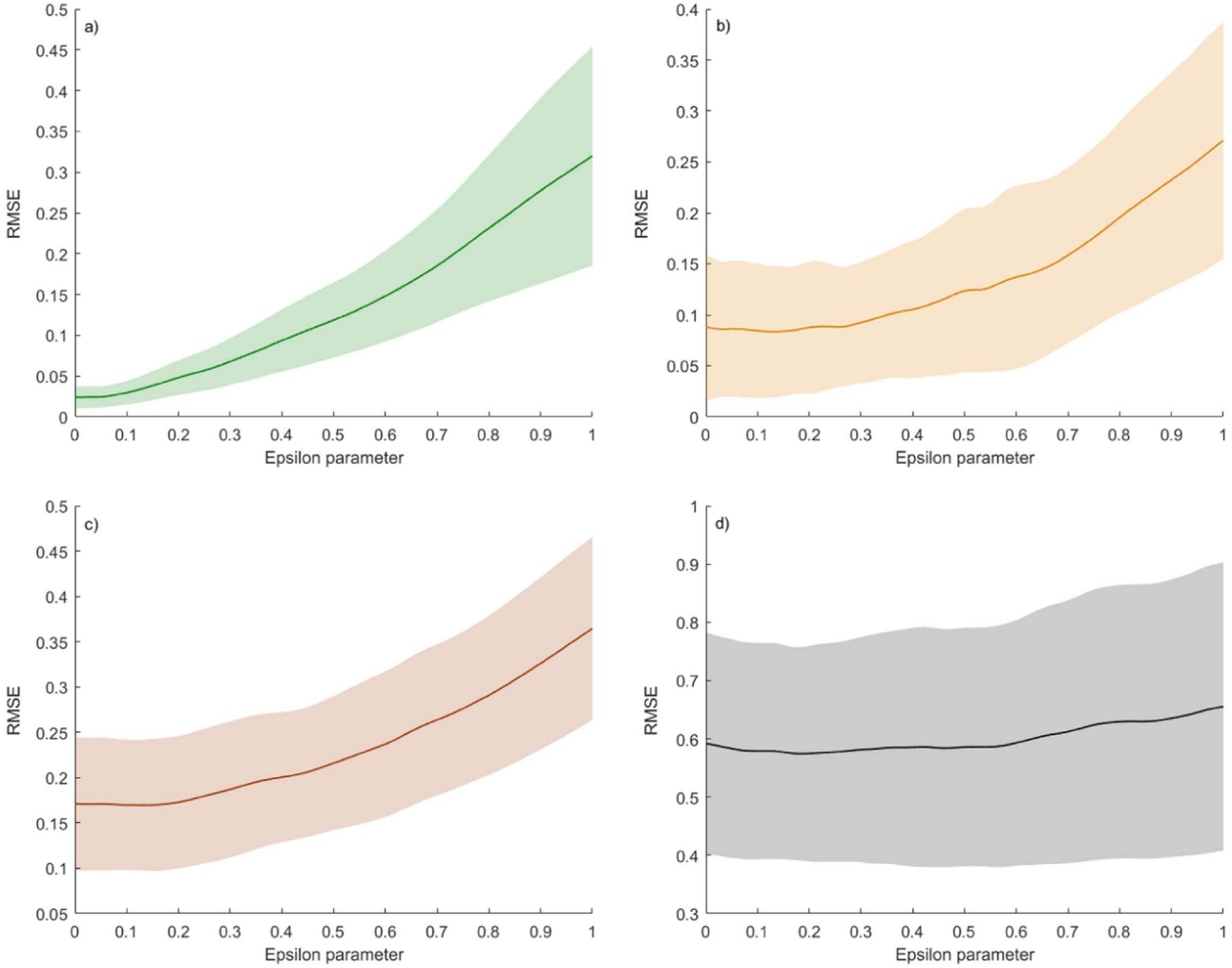


Fig. 12. Curves and bands of a standard deviation of *RMSE* in cross-validation of the prediction model for °Brix according to the value of ε in the 4 steps of the process.

findings.

4. Discussion

The results of this study were compared to those published by Tange et al. [24], who used SVR to carry out the model calibration for °Brix and Sucrose in the manufacturing process of sugar cane. The results presented herein obtained lower *RMSE* values than Tange et al. [24], who presented more accurate estimates of the quality parameters. This is explained by the fact that in the present work techniques of preprocessing and selection of features were used, and therefore noise spectra were removed, along with the wavelengths which did not contribute significantly to the model.

Rinnan et al. [31] argued that performing various preprocessing stages was not advisable in NIR spectra; however, the results of this study showed an improvement with the preprocessing technique number 9, which consisted of calculating the first spectral derivative, then apply SNV and finally extract the trend. Other recent studies on NIR spectroscopy have combined several preprocessing techniques with very good results [32,35–38]. Xu et al. [32] stated that by combining preprocessing techniques, the model takes advantage of the complementary information given by each preprocessing method, therefore, the stability of the models and the results are improved in terms of *RMSE*.

In the proposed model, the T-test feature selection technique allowed recognizing the most relevant wavelengths, removing noise generated by the other wavelengths and obtaining more accurate estimates, which is consistent with other research studies in which T-test feature selection was used to remove the irrelevant variables, proving its efficacy in obtaining more accurate results [41,75–78].

The optimization of parameters of SVR proved its importance to obtain minimum *RMSE* for the model. In our case, it was performed by using a search grid technique, providing the optimal combination of parameters C, γ and ε, which is consistent with the results obtained by Jeng [65] and by Devos et al. [64] who claimed that the combined values of the parameters of SVM determined the complexity of the limits and therefore the performance of the model.

In this regard, one can deduce that the proposed model is accurate and stable, due to the parameter optimization, which is consistent with the results obtained by Cristianini and Shawe [63] and Devos et al. [64] who stated that the adjustment of the SVM kernel parameters controlled the complexity of the resulting hypothesis and avoided the overfitting of the model.

The evaluation of the models was performed using the repeated cross-validation technique, which, according to Garcia and Filzmoser [67] leads to a suitable method aimed at choosing the best model to analyze the mean and standard deviation of the results of repetitions; these results correspond to the test data set, that is, they are data which were not used for calibration, which allows estimating how the model would behave in the future with new data [9,66].

Table 1 shows the results of the proposed global model for °Brix compared to those reported by Tange et al. [24]; note that the proposed global model has a *RMSE* of 0.305, whereas the previously published global model reaches 0.59, showing an optimization of the model. The

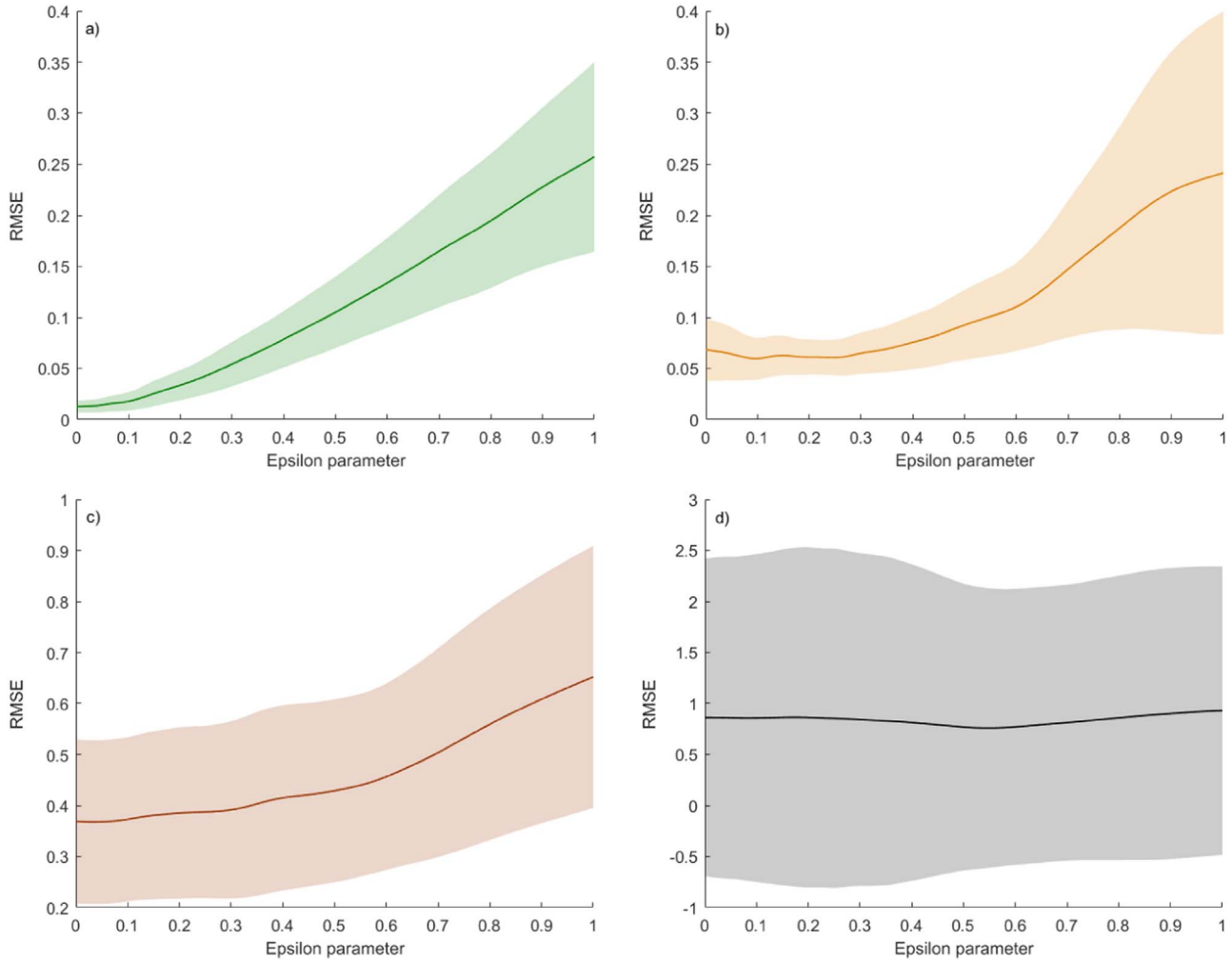


Fig. 13. Curves and bands of a standard deviation of *RMSE* in cross-validation of the prediction model for Sucrose according to the value of ε in the 4 steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

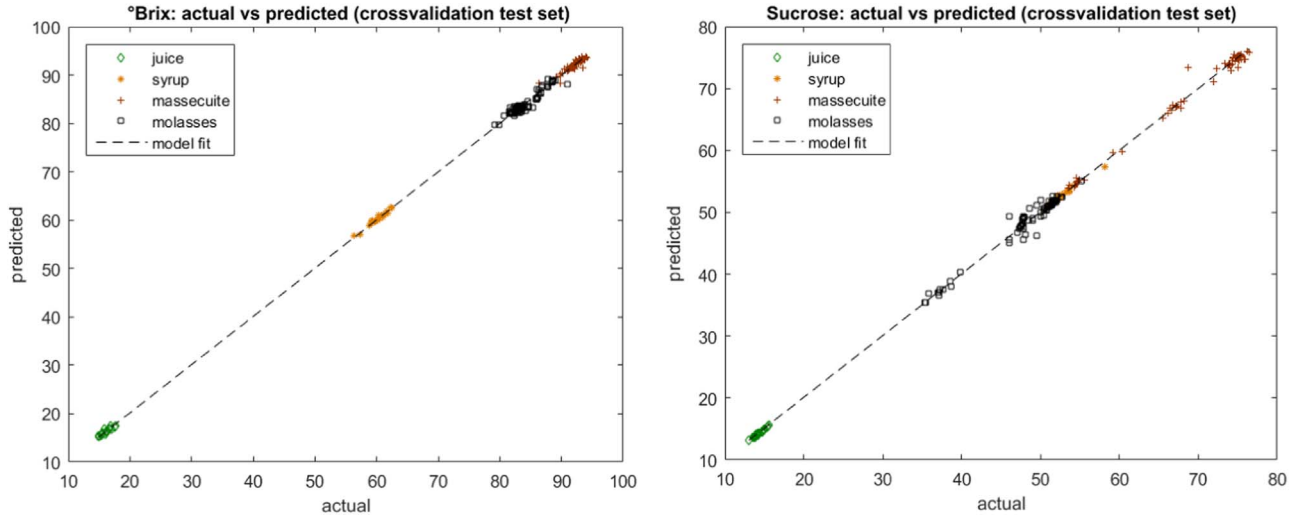


Fig. 14. Regression plot (actual vs predicted) in cross-validation global models for °Brix and Sucrose.

same applies to the four steps of the process, in which the proposed global model improves the previously published global model, and also the four published local (individual) models.

Table 2 shows the results of the proposed global model for Sucrose, with two alternative values that ε may take, Optimized SVM 1 with a value of 0.07 and Optimized SVM 2 with a value of 0.51, which are compared to those published by Tange et al. [24]. The proposed global model has a *RMSE* of 0.486 (Optimized SVM 1) and 0.485 (Optimized SVM 2), whereas the published global model reaches 0.64, showing an optimization of the model with both values of ε.

Table 1
Comparative results of *RMSE* of the models referred* by Tange et al. [24] with an optimized prediction model for °Brix proposed by the authors.

Model	Juice	Syrup	Massecuite	Molasses	Global
Reference*	0.1	0.2	0.5	0.5	0.5
Local SVM*	0.08	0.22	0.39	0.75	–
Global SVM*	0.16	0.25	0.47	0.79	0.59
Optimized SVM	0.040 ± 0.018	0.084 ± 0.063	01702 ± 0.073	0.576 ± 0.183	0.305 ± 0.076

Table 2
Comparative results of *RMSE* of the models referred* by Tange et al. [24] with the optimized prediction models for Sucrose proposed by the authors.

Model	Juice	Syrup	Massecuite	Molasses	Global
Reference*	0.1	0.2	0.5	0.5	0.5
Local SVM*	0.11	0.22	0.56	0.62	–
Global SVM*	0.20	0.24	0.72	0.72	0.64
Optimized SVM	0.016 ± 0.008	0.062 ± 0.023	0.369 ± 0.161	0.858 ± 1.586	0.485 ± 0.631
Optimized SVM 2	0.108 ± 0.036	0.094 ± 0.035	0.431 ± 0.180	0.765 ± 1.396	0.486 ± 0.559

The proposed model for Sucrose exceeds the one published in three out of the four steps of the process; in the process step of Molasses, the proposed model is similar to the one published, however it does not exceed it; this is due to the fact that in the set of Molasses spectra, values with absorbance over two were found, which, according to Tange et al., [24] in their publication were removed from the original dataset. However, in the opinion of the authors, these spectra are not considered outliers, thus in the current study they were maintained to provide more robustness to the model.

5. Conclusions

This study evaluates the application of feature selection techniques and the determination of the optimal configuration of the parameters of a chemometric calibration model based on support vector regression, a technique commonly used in machine learning. Compared to the published models, the models proposed herein were able to better estimate the non-linearities caused by the combination of the NIR spectra from multiple stages of the manufacturing process of sugar cane.

The proposed models for Brix and Sucrose were improved compared to those published by Tange et al. [24] in the four steps of the process, except for the prediction of Sucrose in molasses, which is similar to the published model, although in the present work no spectrum was separated from the original dataset.

Calculating the first spectral derivative from the raw signal, performing SNV and finally extracting the trend was the best combination of preprocessing techniques for the case study. Its implementation improved the stability of the models and the results in terms of *RMSE* by taking supplementary information from each individual technique.

Feature selection reduced the number of wavelengths selected for the calibration of the models, which simplifies the final model, with the corresponding calculation reduction needed to estimate the quality parameters in the sugar industry.

The use of global models with a lower *RMSE* allows a better estimate of the quality parameters, with a single calibration process and, therefore, a simpler and more effective monitoring of the process in the sugar industry. Relying on a methodology that allows a more accurate quality control paves the way for the detection of substances which are found in lower concentrations using NIR spectroscopy.

6. Future developments

Future work focuses on the application of these feature selection techniques such as multivariate filters and wrappers, in addition to the identification of redundant wavelengths by means of analysis of covariance to simplify the models. Other algorithms, such as decision tree regression, artificial neural networks, optimized PLS, and knn regression, can be studied comparatively in order to determine their performance measures. Additionally, authors consider it is important to study the origin of noises are generated in raw spectra.

The proposed methodology can be applied to other models in the food industry, such as ingredients for feed, whose chemical composition could be determined by applying a similar process. Additionally, it could be applied to other signals, such as voltammetry, Raman spectroscopy or electrical impedance, whereby employing a supervised machine learning approach, application models beneficial to the industry can be calibrated.

Acknowledgement

The authors would like to express their gratitude to Professor Ramus Bro and the University of Copenhagen, who kindly provided the spectral dataset for this research study. They also gratefully acknowledge the support from CEDIA National Research and Education Network, and CESGA Supercomputing Center. This work is part of DINTA-UTMACH and RNASA-UDC research groups. Our special thanks to the three anonymous reviewers whose suggestions helped to improve and clarify this manuscript.

References

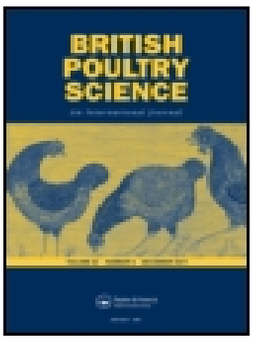
[1] L.S. Polanco, D.F. Day, S. Savoie, S. Bergeron, T. Charlet, B.L. Legendre, Improvements of Raw Sugar Quality Using Double Purge of C-Massecuites Performance Comparison, in: lsu AgCenter Audubon Sugar Institute Factory Operations Seminar, lsuagcenter.com, 2014, p. 46.
[2] C. Kumaravelu, A. Gopal, A review on the applications of Near-Infrared spectrometer and Chemometrics for the agro-food processing industries, 2015, pp. 8–12.
[3] B.R. Kowalski, Chemom., Anal. Chem. 52 (1980) 112R–122R.
[4] E. Bertran, M. Blanco, S. Maspoch, M.C. Ortiz, M.S. Sánchez, L.A. Sarabia, Handling intrinsic non-linearity in near-infrared reflectance spectroscopy, Chemom. Intell. Lab. Syst. 49 (1999) 215–224.
[5] H. Martens, T. Naes, Multivariate Calibration, Wiley, 1992.
[6] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemom. 6 (1992) 267–281.
[7] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.
[8] H. Cen, Y. He, Theory and application of near infrared reflectance spectroscopy in determination of food quality, Trends Food Sci. Technol. 18 (2) (2007) 72–83.
[9] J.B.O. Mitchell, Machine learning methods in chemoinformatics, Wiley Interdiscip. Rev. Comput. Mol. Sci. 4 (2014) 468–481.
[10] P. Torrione, L.M. Collins, K.D. Morton, Multivariate analysis, chemometrics, and machine learning in laser spectroscopy.
[11] R.G. Brereton, Pattern recognition in chemometrics, Chemom. Intell. Lab. Syst. 149 (Part B) (2015) 90–96.
[12] M.M. Tomazzoli, R.D. Pai Neto, R. Moresco, L. Westphal, A.R.S. Zeggio, L. Specht, C. Costa, M. Rocha, M. Maraschin, Discrimination of Brazilian propolis according to the seasoning using chemometrics and machine learning based on UV–vis

scanning data, J. Integr. Bioinform. 12 (2015) 279.
[13] M. Tajammal Munir, W. Yu, B.R. Young, D.I. Wilson, The current status of process analytical technologies in the dairy industry, Trends Food Sci. Technol. 43 (6) (2015) 205–218.
[14] L. Wang, D.-W. Sun, H. Pu, J.-H. Cheng, Quality analysis and classification and authentication of liquid foods by near-infrared spectroscopy: a review of recent research developments, Crit. Rev. Food Sci. Nutr. (2016) 0.
[15] E. Zamora-Rojas, D. Pérez-Marín, E. De Pedro-Sanz, J.E. Guerrero-Ginel, A. Garrido-Varo, Handheld NIRS analysis for routine meat quality control: Database transfer from at-line instruments, Chemom. Intell. Lab. Syst. 114 (2012) 30–35.
[16] H.-J. He, D. Wu, D.-W. Sun, Rapid and non-destructive determination of drip loss and pH distribution in farmed Atlantic salmon (Salmo salar) fillets using visible and near-infrared (Vis–NIR) hyperspectral imaging, Food Chem. 156 (2014) 394–401.
[17] R. Henry, P. Kettlewell, Cereal Grain Quality, Springer, Netherlands, 2012.
[18] L.S. Magwaza, U.L. Opara, H. Nieuwoudt, P.J.R. Cronje, W. Saeys, B. Nicolaï, NIR Spectroscopy Applications for Internal and External Quality Analysis of Citrus Fruit—A Review, Food Bioprocess Technol. 5 (2011) 425–444.
[19] P. Valderrama, J.W.B. Braga, R.J. Poppi, Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy, J. Braz. Chem. Soc. 18 (2007) 259–266.
[20] E.P. Zayas-Ruiz, M. Lorenzo-Izquierdo, F.O. Fragoso-Concepción, La quimiometría y la industria del azúcar y sus derivados, ICIDCA. Sobre Los Derivados de La Caña de Azúcar., vol. 49, 2015, pp. 31–33.
[21] N. Sorol, E. Arancibia, S.A. Bortolato, A.C. Olivieri, Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice: a test field for variable selection methods, Chemom. Intell. Lab. Syst. 102 (2010) 100–109.
[22] X. Wang, H.-J. Ye, Q.-T. Li, J.-C. Xie, J.-J. Lu, A.-L. Xia, J. Wang, Determination of Brix and POL in Sugar Cane Juice by Using Near Infrared Spectroscopy Coupled with BP-ANN, Spectrosc. Spectr. Anal. 30 (2010) 1759–1762.
[23] H.Z. Chen, J.B. Wen, J.C. Chen, L.H. Li, Y.J. Huo, Near-infrared spectroscopic modeling optimization for quantitative determination of sugar brix in sugarcane initial-pressure juice, Int. J. Technol. Res. Appl. 2 (2014) 6.
[24] R.I. Tange, M.A. Rasmussen, E. Taira, R. Bro, Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps, J. Infrared Spectrosc. (2015) ⟨http://www.forskningsdatabasen.dk/en/catalog/2266200601⟩.
[25] C.D. Brown, P.D. Wentzell, Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration, J. Chemom. 13 (1999) 133–152.
[26] E. Stark, Near infrared spectroscopy past and future, Near Infrared Spectroscopy, Future Waves (1996) 701–713.
[27] C.A.T. dos Santos, M. Lopo, R.N.M.J. Páscoa, J.A. Lopes, A review on the applications of portable near-infrared spectrometers in the agro-food industry, Appl. Spectrosc. 67 (2013) 1215–1233.
[28] E. Teye, X.-Y. Huang, N. Afoakwa, Review on the potential use of near infrared spectroscopy (NIRS) for the measurement of chemical residues in food, Am. J. Food Sci. Technol. 1 (2013) 1–8.
[29] M. Blanco, I. Villarroya, NIR spectroscopy: a rapid-response analytical tool, Trends Anal. Chem. 21 (4) (2002) 240–250.
[30] W.J. Florkowski, S.E. Prussia, R.L. Shewfelt, B. Brueckner, Postharvest Handling: a Systems Approach, Elsevier Science, 2009.
[31] Å. Rinnan, F. van, D. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends Anal. Chem. 28 (2009) 1201–1222.
[32] L. Xu, Y.-P. Zhou, L.-J. Tang, H.-L. Wu, J.-H. Jiang, G.-L. Shen, R.-Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta. 616 (2008) 138–143.
[33] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, Appl. Spectrosc. 43 (1989) 772–777.
[34] J. Luypaert, S. Heuerding, S. de Jong, D.L. Massart, An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream, J. Pharm. Biomed. Anal. 30 (2002) 453–466.
[35] M.J. Martelo-Vidal, M. Vázquez, Evaluation of ultraviolet, visible, and near infrared spectroscopy for the analysis of wine compounds, Czech J. Food Sci. 32 (2014) 37.
[36] L. Xie, X. He, B. Duan, S. Tang, J. Luo, G. Jiao, G. Shao, X. Wei, Z. Sheng, P. Hu, Optimization of Near-Infrared Reflectance Model in Measuring Gelatinization Characteristics of Rice Flour with a Rapid Viscosity Analyzer (RVA) and Differential Scanning Calorimeter (DSC), Cereal Chem. 92 (2015) 522–528.
[37] X. Pan, Y. Li, Z. Wu, Q. Zhang, Z. Zheng, X. Shi, Y. Qiao, A online NIR sensor for the pilot-scale extraction process in Fructus aurantii coupled with single and ensemble methods, Sensors 15 (2015) 8749–8763.
[38] G.M. Hadad, A.S. Ra, M.M. Elkhoudarya, Simultaneous determination of clarithromycin, tinidazole and omeprazole in helicure tablets using reflectance near-infrared spectroscopy with the aid of chemometry, Pharm. Anal. Acta (2015).
[39] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[40] S. Keleş, M. van der Laan, M.B. Eisen, Identification of regulatory elements using a feature selection method, Bioinformatics 18 (2002) 1167–1175.
[41] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer, Berlin Heidelberg, 2008.
[42] E. Szymańska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L.M.C. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, Trends Anal. Chem. 69 (6) (2015) 34–51.
[43] I. Guyon, A. Elisseeff, An introduction to seature extraction, in: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh (Eds.), Feature Extraction, Springer, Berlin Heidelberg, 2006, pp. 1–25.
[44] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer, US, 2012.
[45] P. Jafari, F. Azuaje, An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors, BMC Med. Inform. Decis. Mak. 6 (2006) 27.
[46] G. Bhanot, G. Alexe, B. Venkataraghavan, A.J. Levine, A robust meta-classification strategy for cancer detection from MS data, Proteomics 6 (2006) 592–604.
[47] A. Mucherino, G. Ruß, Recent Developments in Data Mining and Agriculture, in: Industrial Conference on Data Mining- ..., 2011: pp. 1–9.
[48] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, New York, NY, 2009.
[49] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, Data Mining in Agriculture, Springer, New York, New York, NY, 2009.
[50] D. Basak, S. Pal, D.C. Patranabis, Support vector regression, Neural Information Processing, Lett. Rev. 11 (2007) 203–224.
[51] J. Palma, R. Marín, Inteligencia artificial. Técnicas, métodos y aplicaciones, McGraw Hill, Murcia, 2013.
[52] B.E. Boser, I.M. Guyon, V.N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. (1992) 144–152.
[53] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[54] V. Vapnik, S.E. Golowich, A.J. Smola, Support vector method for function approximation, regression estimation and signal processing, in: M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, MIT Press, 1997, pp. 281–287.
[55] L. Zhang, W.-D. Zhou, P.-C. Chang, J.-W. Yang, F.-Z. Li, Iterated time series prediction with multiple support vector regression models, Neurocomputing 99 (2013) 411–422.
[56] J. Wu, J. Wei, Combining ICA with SVR for prediction of finance time series, in: 2007 IEEE International Conference on Automation and Logistics, ieeexplore.ieee.org, 2007, pp. 95–100.
[57] U.A. Acar, B. Hudson, G.L. Miller, T. Phillips, SVR: practical Engineering of a Fast 3D Meshing Algorithm*, in: M.L. Brewer, D. Marcum (Eds.), Proceedings of the 16th International Meshing Roundtable, Springer, Berlin Heidelberg, 2008, pp. 45–62.
[58] Y. Quan, J. Yang, L.-X. Yao, C.-Z. Ye, Successive overrelaxation for support vector regression, J. Softw. Maint. Evol.: Res. Pract. 15 (2004) 200–206.
[59] P. Koch, B. Bischl, O. Flasch, T. Bartz-Beielstein, C. Weihs, W. Konen, Tuning and evolution of support vector kernels, Evol. Intell. 5 (2012) 153–170.
[60] K. Mollazade, M. Omid, A. Arefi, Comparing data mining classifiers for grading raisins based on visual features, Comput. Electron. Agric. 84 (2012) 124–131.
[61] C.-H. Wu, G.-H. Tzeng, R.-H. Lin, A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression, Expert Syst. Appl. 36 (4) (2009) 4725–4735.
[62] R.K. Prasoona, A. Jyoti, Y. Mukesh, S. Nishant, N.S. Anuraj, J. Shobha, Optimization of gaussian kernel function in support vector machine aided QSAR studies of C-aryl glucoside SGLT2 inhibitors, Interdiscip. Sci. 5 (2013) 45–52.
[63] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge, 2000.
[64] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.-P. Huvenne, Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation, Chemom. Intell. Lab. Syst. 96 (2009) 27–33.
[65] J.-T. Jeng, Hybrid approach of selecting hyperparameters of support vector machine for regression, IEEE Trans. Syst. Man Cybern. B Cybern. 36 (2006) 699–709.
[66] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, 2013.
[67] H. Garcia, P. Filzmoser, Multivariate Statistical Analysis using the R package chemometrics, Vienna, Austria, 2015. ⟨ftp://155.232.191.229/cran/web/packages/chemometrics/vignettes/chemometrics-vignette.pdf⟩.
[68] S.R. Gunn, Support vector machines for classification and regression, ISIS Technical Report. 14, 1998. ⟨http://ce.sharif.ir/courses/85-86/2/ce725/resources/root/LECTURES/SVM.pdf⟩.
[69] O. Devos, L. Duponchel, Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression, Chemom. Intell. Lab. Syst. 107 (2011) 50–58.
[70] F. Allegrini, A.C. Olivieri, An integrated approach to the simultaneous selection of variables, mathematical pre-processing and calibration samples in partial least-squares multivariate calibration, Talanta 115 (2013) 755–760.
[71] X. Ma, Y. Zhang, Y. Wang, Performance evaluation of kernel functions based on grid search for support vector regression, in: 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), ieeexplore.ieee.org, 2015: pp. 283–288.
[72] R.A. Viscarra Rossel, ParLeS: software for chemometric analysis of spectroscopic data, Chemom. Intell. Lab. Syst. 90 (2008) 72–83.
[73] J.S. Armstrong, F. Collopy, Error measures for generalizing about forecasting methods: empirical comparisons, Int. J. Forecast. 8 (1992) 69–80.
[74] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688.
[75] C. Christin, H.C.J. Hoefsloot, A.K. Smilde, B. Hoekman, F. Suits, R. Bischoff, P. Horvatovich, A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, Mol. Cell. Proteom. 12 (2013) 263–276.
[76] N. Erho, A. Crisan, I.A. Vergara, A.P. Mitra, M. Ghadessi, C. Buerki, E.J. Bergstralh,

T. Kollmeyer, S. Fink, Z. Haddad, B. Zimmermann, T. Sierocinski, K.V. Ballman, T.J. Triche, P.C. Black, R.J. Karnes, G. Klee, E. Davicioni, R.B. Jenkins, Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy, PLoS One 8 (2013) e66855.

[77] D. Zhu, K. Li, D.P. Terry, A.N. Puente, L. Wang, D. Shen, L.S. Miller, T. Liu, Connectome-scale assessments of structural and functional connectivity in MCI, Hum. Brain Mapp. 35 (2014) 2911–2923.

[78] N.-F. Jie, M.-H. Zhu, X.-Y. Ma, E.A. Osuch, M. Wammes, J. Théberge, H.-D. Li, Y. Zhang, T.-Z. Jiang, J. Sui, V.D. Calhoun, Discriminating bipolar disorder from major depression based on SVM-FoBa: efficient feature selection with multimodal brain imaging data, IEEE Trans. Auton. Ment. Dev. 7 (2015) 320–331.



British Poultry Science

ISSN: 0007-1668 (Print) 1466-1799 (Online) Journal homepage: http://www.tandfonline.com/loi/cbps20

Automated early detection of drops in commercial egg production using neural networks

I. Ramírez-Morales, E. Fernández-Blanco, D. Rivero & A. Pazos

To cite this article: I. Ramírez-Morales, E. Fernández-Blanco, D. Rivero & A. Pazos (2017): Automated early detection of drops in commercial egg production using neural networks, British Poultry Science, DOI: 10.1080/00071668.2017.1379051

To link to this article: http://dx.doi.org/10.1080/00071668.2017.1379051

Accepted author version posted online: 20 Sep 2017.
Published online: 17 Oct 2017.

BRITISH POULTRY SCIENCE, 2017
https://doi.org/10.1080/00071668.2017.1379051

Taylor & Francis
Taylor & Francis Group

Automated early detection of drops in commercial egg production using neural networks

I. Ramírez-Morales [a,b], E. Fernández-Blanco [b], D. Rivero [b] and A. Pazos [b]

[a]Universidad Técnica de Machala, Faculty of Agricultural & Livestock Sciences, Machala, Ecuador; [b]Universidade A Coruña, Department of Computer Science, A Coruña, España

ABSTRACT

1. The purpose of this work was to support decision-making in poultry farms by performing automatic early detection of anomalies in egg production.
2. Unprocessed data were collected from a commercial egg farm on a daily basis over 7 years. Records from a total of 24 flocks, each with approximately 20 000 laying hens, were studied.
3. Other similar works have required a prior feature extraction by a poultry expert, and this method is dependent on time and expert knowledge.
4. The present approach reduces the dependency on time and expert knowledge because of the automatic selection of relevant features and the use of artificial neural networks capable of cost-sensitive learning.
5. The optimum configuration of features and parameters in the proposed model was evaluated on unseen test data obtained by a repeated cross-validation technique.
6. The accuracy, sensitivity, specificity and positive predictive value are presented and discussed at 5 forecasting intervals. The accuracy of the proposed model was 0.9896 for the day before a problem occurs.

ARTICLE HISTORY
Received 3 June 2017
Accepted 9 August 2017

KEYWORDS
Farming systems; laying hens; machine learning; modelling; production drops

Introduction

Poultry are monitored based on the producer's experience and expertise in managing and evaluating the productive process (Frost et al. 1997). A current tendency to manage larger populations in poultry farms has motivated the development and use of automatic monitoring systems as a complement to human observations. These techniques aim to increase the company's net income (Antonov et al. 2015).

Although collecting high-quality data under field conditions is a challenging situation (Pica-Ciamarra et al. 2014), it is regularly required for poultry keepers to monitor the health of hens and the production of eggs. Therefore, certain animal-related variables, such as the number of eggs, death rate, food and water consumption, weight, and environment variables are regularly tracked (Wheeler et al. 2003; Long and Wilcox 2011; Hepworth et al. 2012).

Data analysis in poultry systems has mainly been performed using mathematical methods (Frost et al. 1997), statistical techniques (Narinc et al. 2014) and visual graph analyses (Mertens et al. 2009). These methods allow for the identification of anomalies in production by pointing out important differences among the production indices (De Vries and Reneau 2010).

Schaefer et al. (2004) and Cameron (2012) suggested that the detection of problems in animal production is one of the key aspects in data-analysis. Early detection enables the implementation of appropriate actions to adequately correct problems. Therefore, the impact of diseases, the potential number of infected animals, and the costs associated with treatment and production losses are reduced (Gates et al. 2015). Lokhorst and Lamaker (1996) proposed an expert system for supporting decision-making in egg production. Xiao et al. (2011) performed a statistical analysis of the production curve to identify possible problems. More recently, Ramírez-Morales et al. (2016) have shown the suitability of support vector machines with a sliding window for the development of early warning models.

The high complexity of such analyses leads to the use of newer techniques, such as artificial neural networks (ANNs), which allow for the development of more robust systems against unexpected conditions. These techniques could provide insights into the relationship among data from examples through a process called training (Mucherino et al. 2009).

The behaviour of ANNs is inspired by the ability of the human brain to identify patterns. One of the best-known types of ANNs is the multilayer perceptron (MLP), which organises the processing units called neurons in layers with forwarding connections only between consecutive layers. Therefore, the general scheme includes an input layer, zero or more hidden layers, and an output layer (Kruse et al. 2013), as shown in Figure 1. ANNs, and particularly MLPs, have been used in different knowledge areas and shown remarkable results (Guo et al. 2010; Samborska et al. 2014; Kalhor et al. 2016).

The main challenge of warning systems is detecting unexpected values because of the low frequency of these events in datasets. Therefore, cost-sensitive learning ANNs have attracted the attention of experts because they are able to model the occurrence of rare events (Zhi-Hua and Xu-Ying 2006; Zahirnia et al. 2015).

The objective of this study was to optimise the model's parameters to perform automatic early detection of anomalies in egg production. To accomplish this, an automatic

CONTACT Ivan Ramírez-Morales iramirez@utmachala.edu.ec Universidade A Coruña, Department of Computer Science, 15071 A Coruña (03082), A Coruña, España

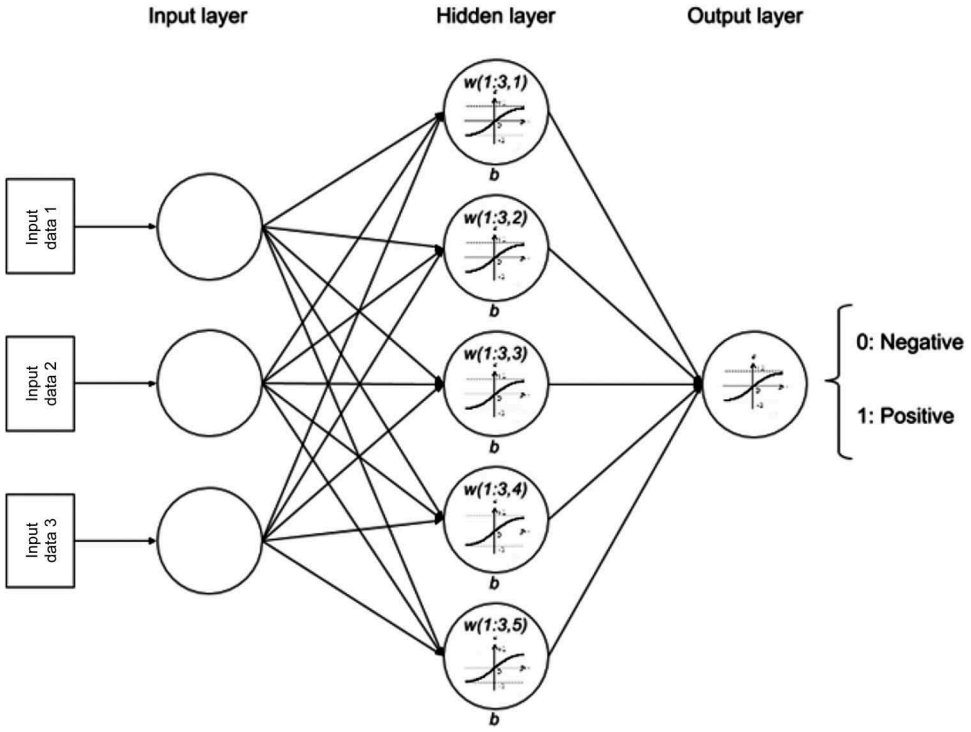


Figure 1. Multilayer perceptron with one hidden layer representation.

feature selection technique and a cost-sensitive ANN model are used.

Materials and methods

Data description

Daily field data from 24 flocks were collected from January 2008 to December 2015 from a commercial egg production farm located in Ecuador. The birds were housed in automatic cage systems using the all-in-all-out replacement method (Flanders and Gillespie 2015). Pullets of the same age were moved at 16 weeks of age from a rearing house to open sided poultry laying houses, where they stayed in small groups with 605 cm^2 per hen until the end of the production cycle.

In this study, data were analysed from 19 weeks to 79 weeks. Because of the management and internal logistics of the farm, the produced eggs were collected at a different time each day. The interval between daily collections ran from 20 to 28 h. This variability over different temporal frames of the data was a challenge for the model because it had to discriminate between the abnormal values and those generated because of a different collection hour.

A total of 188 problematic days were manually labelled for the 24 studied flocks, which represent 1.85% of the total 10 142 records.

Data partition

Data partitioning can be used to determine the accuracy of a model's estimates of new data. In this work, the authors have chosen a 5-fold repeated cross-validation method (Refaeilzadeh et al. 2009; Kuhn and Johnson 2013). This subtype of cross-validation method randomly breaks the dataset into 5 subsets, which are sequentially used as test sets, while the other 4 are used for training. This cross-validation method was repeated 100 times to obtain the information necessary to perform a statistically significant evaluation.

Optimisation methodology

The early detection model was developed using a cost-sensitive learning ANN. This network had to classify raw data in a sliding time window (Kapoor and Bedi 2013; Saeed and Václav 2014). The following methodology focused on setting the optimal values of model's parameters to offer the best possible prediction. This optimisation process can be summarised in three main steps:

First step: input patterns

In this step, the window size and the feature selection threshold based on a t-test (Saeys et al. 2007) were simultaneously optimised. This was performed using a grid search technique (Ma et al. 2015) to conform to the input patterns, which are the sections that results from applying a sliding time window.

The window size determines the number of previous days' data that should be considered as inputs to the model. This parameter was set to one as a default value (Saeed and Václav 2014), and a range from 1 to 30 was tested as a window size.

The upper limit value resulted from a practical situation. If higher values had been used, the flock would have been in production for an extended period before the model could be employed, which would have resulted in a limited usability of the system in field conditions. Similarly, the range for feature selection threshold was set from 0 to 100 for the percentile values of the *P*-value for a single t-test.

Second step: ANN architecture selection

The architecture of an ANN is determined by the number of neurons, layers and their connections. There is no general rule in choosing the best architecture, and optimisation is based on testing several architectures to find one that offers satisfactory results (Herrera et al. 2004; Rivero et al. 2011).

Several architectures of the ANN were tested to choose the best one. One-layer and two-layer architectures with combinations of 0, 25 and 50 neurons in each hidden layer were evaluated. The optimisation algorithm used to train these ANNs is a variation of the gradient descent known as scaled conjugate gradient backpropagation (Møller 1993).

The results were compared using an analysis of variance (ANOVA) and Tukey's Honest Significance Test (Abdi and Williams 2010), and a *P*-value <0.01 was used to determine whether there were statistically significant differences between the proposed architectures.

Third step: optimisation of cost-sensitive learning parameters

The database defines an imbalance between positive and negative targets, which is the main reason to focus attention on the cost-sensitive ANN technique. The last step focused on tuning a weight parameter for modelling imbalanced data. This parameter aimed to reduce overfitting to the most common class (Pazzani et al. 1994; Elkan 2001).

This parameter (S) measures the importance of a wrong prediction of a drop in egg production. S can take values between zero and one as the minimum and maximum importance of the misclassification of positive patterns. Values of the parameter S ranging from 0.05 to 0.95 were tested with steps of 0.05.

Performance analysis

A performance analysis is usually conducted by measuring the accuracy (ACC), which is the proportion of correct classifications made by the model (Martens and Baesens 2010; Venkatesan et al. 2013).

The ACC is the only requirement presented in many machine learning works. However, according to Sun et al. (2009), imbalanced data sets require additional information, such as specificity (SPC), sensitivity (SEN) and positive predictive value (PPV).

In this work, a performance analysis was tested at 6 prediction intervals from time delays between 0 and 5 d (Lindsay and Cox 2005). A value of zero means that the model detects a drop in the production on the current day, whereas values from one to 5 indicate the number of days before the model-predicted anomalies occur.

Results and discussion

The studied flocks had an average of 8 d labelled as problematic because of drops in production. Certain flocks did not present any decrease in egg production, whereas others presented 33 d labelled as drops in production because of different problems.

The summary statistics for egg production in each flock along with their respective mean and dispersion values visualised in a box and whisker plot are provided in Figure 2. A careful examination indicates that each flock has a different data distribution, which is a favourable situation during the training process because the variability of inputs patterns tends to improve the model generalisations.

The production of two specific flocks as representative examples of the presence and absence of problems is shown in Figure 3. Flock 8 does not present any anomalies in the

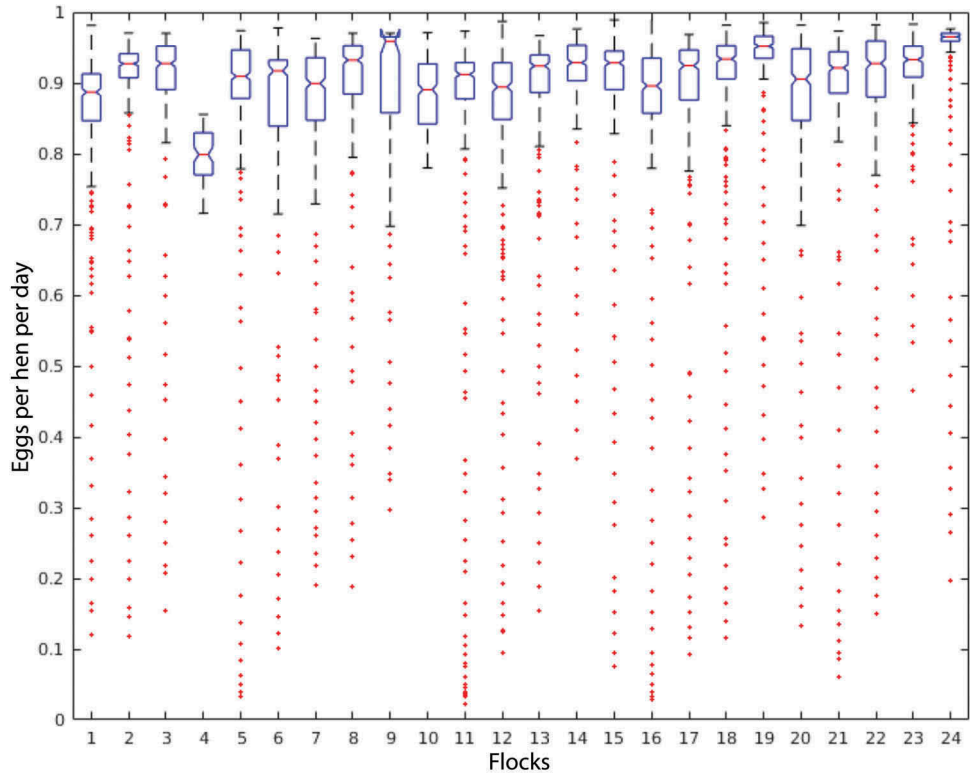


Figure 2. Egg production boxplot representing the daily average per bird for each of the 24 flocks analysed in this work.

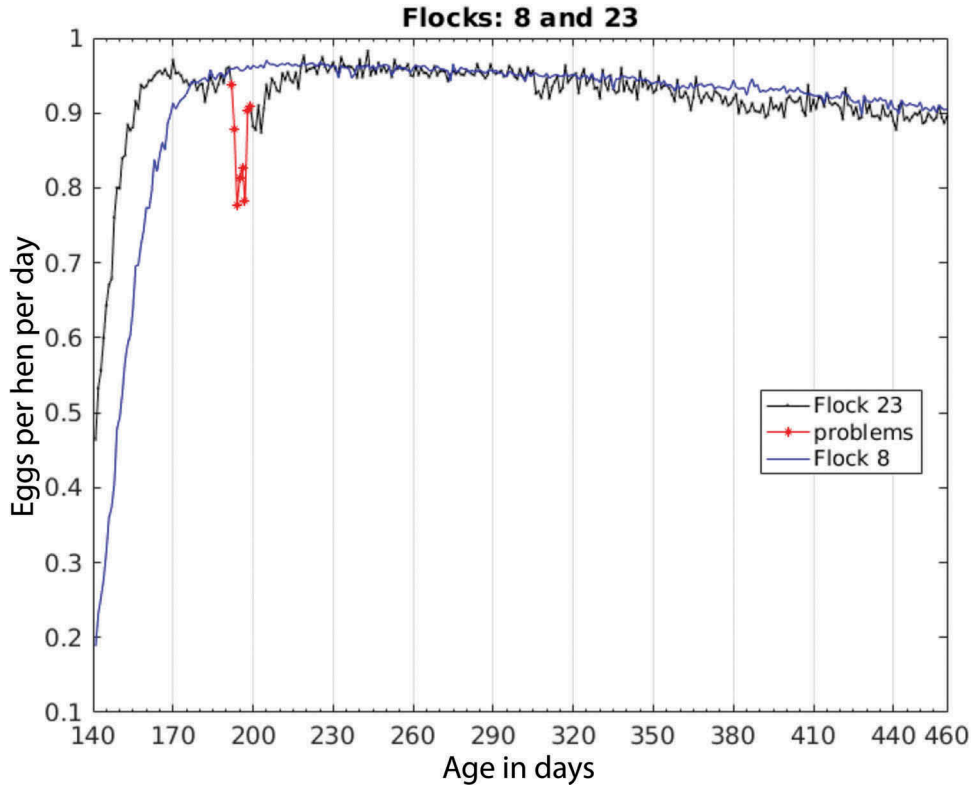


Figure 3. Daily production per bird in representative flocks number 8 and 23.

production curve, whereas Flock 23 starts production earlier but has a significant drop in egg production between d 191 and 199. This interval was labelled a problem because after d 199, egg production was back to normal.

The proposed approach for automatic early detection of the anomalies in the production curve of commercial laying hens is based on a classifier combined with a sliding time window, which allows for the detection of anomalies and was based on the study by Bennett and Campbell (2000), who indicated that classification techniques can be used to detect unusual cases.

Data partitioning was performed using a 5-fold cross-validation technique, which guarantees a reduction of possible overfitting (Refaeilzadeh et al. 2009; Kuhn and Johnson 2013). The optimisation of the model was performed by three consecutive steps:

First step: formation of the input patterns

From the preliminary tests, an initial architecture with one hidden layer with 10 neurons, a value of 0.5 for parameter S and a prediction interval equal to one was chosen. The window size and the feature selection threshold were evaluated simultaneously using a grid search technique. An average of 100 runs in the 5-fold cross-validation is shown in Figure 4. The axes of the grid search are the window size and the feature selection threshold.

The best ACC results were found with a feature selection threshold greater than 40 along with a window size greater than 15 as shown in Figure 4(a). This finding is similar to what occurs in the case of SPC, which is shown in Figure 4 (b). This situation was expected, because the dataset is imbalanced and most targets are negative. Therefore, the influence of these values on the overall ACC of the model is remarkable.

The SEN and PPV results are shown in Figure 4(c,d), respectively. These values express a more marked area of the feature selection threshold between 40 and 70 and the window size between 15 and 20.

Because the selection of relevant features is conditioned by the window size (Frank et al. 2001) and the feature selection threshold, a window size of 18 was chosen along with a feature selection threshold equal to 65. This configuration yields the best results for all performance metrics. The window size and the feature selection threshold, which are far from ideal, display a lower performance, which may be related to insufficient information contained in the selected features or excessive and noisy information for the model. In the model developed by Ramírez-Morales et al. (2016), multiples of 7 for the window size values were required because of the method used by the expert to calculate these features. In the model proposed here, the information contained in the selected relevant features allows for better modelling of the problem without the restriction mentioned before.

The selected relevant features are as follows:

- The number of eggs produced in the first 5 d and the last 4 d of the sliding window.
- The number of dead hens over the past 10 d in the sliding window.
- The number of live hens in the first 14 d of the sliding window.
- All data of cracked eggs in the sliding window.

An illustration of these relevant features is shown in Figure 5, in which the selected features are marked in black.

In a similar work published by Ramírez-Morales et al. (2016), an early warning classifier for the productive

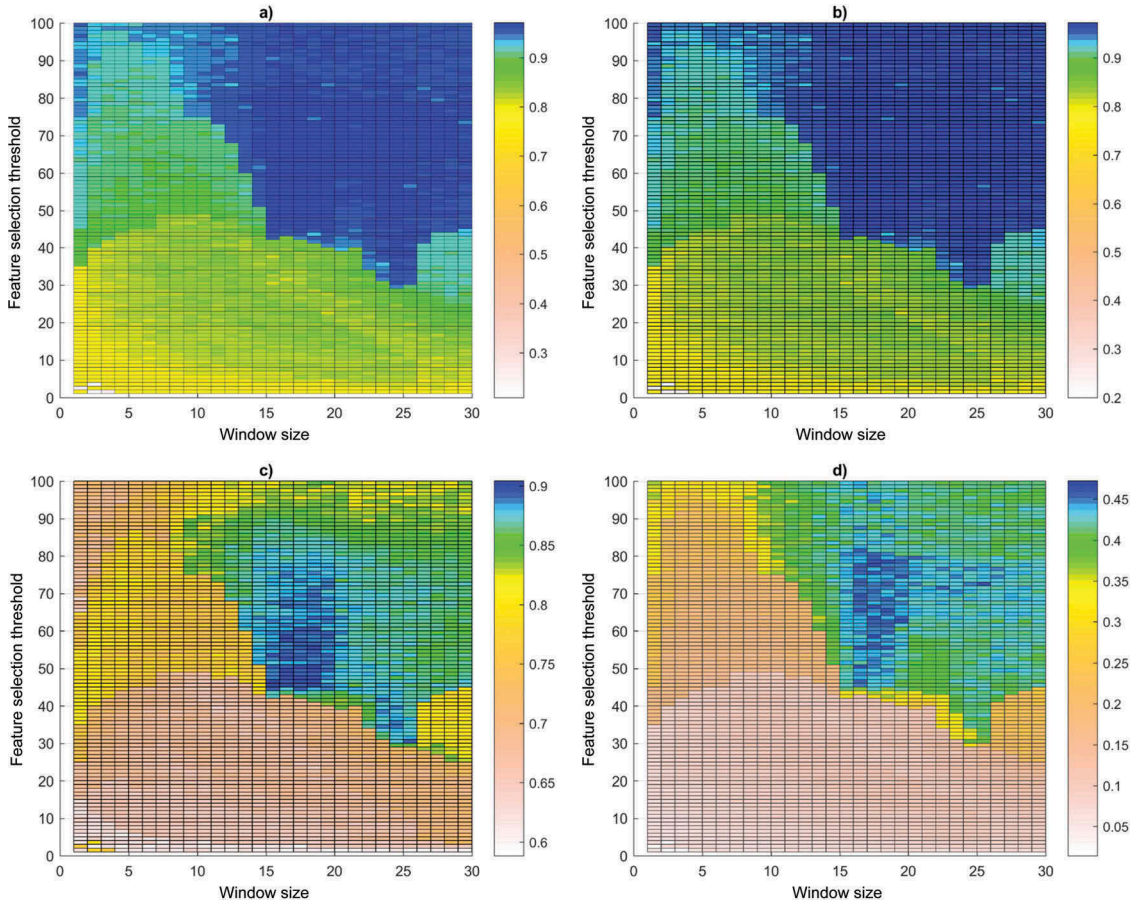


Figure 4. Grid search for window size and feature selection threshold: (a) accuracy, (b) specificity, (c) sensitivity and (d) positive predictive value.

	# of previous days																	
Features	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Number of eggs produced	■	■	■	■	□	□	□	□	□	□	□	□	□	■	■	■	■	■
Dead hens	■	■	■	■	■	■	■	■	■	■	□	□	□	■	■	■	■	■
Number of existng hens	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Cracked eggs	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Figure 5. Selected features with a feature selection threshold of 65 in a window of size equal to 18 d.

problems of layers was proposed. A poultry expert proposed more than 30 features as inputs to the model. These features went through a process of previous selection and 6 were ultimately selected. The features consisted of indicators calculated from the combination of raw data and other parameters calculated within the sliding time window using a manual trial-error process to choose the best ones. In the current study, an automatic process of selection of relevant features was conducted starting with an analysis of raw data, and an expert was not required.

According to Isabelle Guyon and Elisseeff (2003), the FS is necessary in many cases to prevent overfitting, and it is also related to an improvement of performance and a reduction of computational costs (Saeys et al. 2007; Guyon et al. 2008). For the proposed model, the chosen features correspond to the original raw data, which confer less complexity compared to the required features by Ramírez-Morales et al. (2016). Variables, such as the age of the birds, cumulative mortality, environmental temperature and humidity, were not used as inputs to the model because of their lack of importance according to the automatic FS technique.

Second step: selection of the best ANN architecture

At this stage of the study, 7 architectures were evaluated according to Herrera et al. (2004) and Rivero et al. (2011), who stated that it was necessary to test several different architectures to find the architecture that obtains good results. An ANOVA and Tukey's Honest Significance Test were performed, and a value of $P < 0.01$ indicated significant differences (Abdi and Williams 2010).

The window size and the feature selection threshold were fixed in the first step along with a parameter S-value equal to 0.5 and an advance equal to one. A process of testing several predefined architectures was needed to choose the best architecture for the ANN. Among these architectures, the one with the best performance with the test data set was chosen. The evaluated architectures of the ANN are shown in Table 1. It should be noted that none of the two-layer

Table 1. Multiple comparisons of the evaluated architectures.

Network architecture	Accuracy		Specificity		Sensitivity		Positive predictive value	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
[no HL]	0.955[a]	0.062	0.956[a]	0.063	0.900[a]	0.083	0.387[b]	0.159
[25]	0.968[a]	0.021	0.970[a]	0.022	0.880[a]	0.057	0.451[a]	0.174
[25,25]	0.961[a]	0.026	0.962[a]	0.027	0.884[a]	0.062	0.411[ab]	0.178
[25,50]	0.959[a]	0.028	0.961[a]	0.029	0.884[a]	0.058	0.405[ab]	0.175
[50]	0.967[a]	0.020	0.970[a]	0.021	0.879[a]	0.059	0.450[a]	0.175
[50,25]	0.962[a]	0.026	0.964[a]	0.027	0.876[a]	0.070	0.418[ab]	0.180
[50,50]	0.962[a]	0.025	0.964[a]	0.026	0.876[a].	0.061	0.416[ab]	0.175

Means in a row with no common superscript letter differ significantly according to Tukey's Honest Significance Test for a value of $P < 0.01$.

architectures improved upon the results obtained by a single hidden layer architecture consisting of 25 neurons, which showed the best performance metrics.

ANN training has two sources of random variability: the partitioning of patterns and the initialisation of weights. Consequently, returning the best results of multiple runs is not sufficient to select a proper configuration of parameters. It is recommended to calculate the average and the standard deviation of the performance metrics in all repetitions. Hypothesis tests should be performed to check the significance of differences between ANN configurations (Amiri et al. 2014; Singh et al. 2015; Anwar et al. 2016). Statistically significant differences were not observed for architectures with hidden layers of between 25 and 50 neurons. When deciding between two architectures with statistically equal performances, the less complex architecture is preferred because simpler models tend to be less prone to overfitting. Therefore, to continue with the optimisation, a single hidden layer architecture consisting of 25 neurons was chosen.

Third step: optimisation of the cost-sensitive learning parameter

The third step of the methodology was to tune the value of the parameter S between 0.05 and 0.95 in intervals of 0.05. The ANNs used the window size and a feature selection threshold fixed in the first stage along with the architecture set in the second stage and a prediction interval of 1 d.

The influence of the S-value on the performance metrics of the model is plotted in Figure 6. The line in this figure represents the average of 100 evaluations, and the shaded area corresponds to the standard deviation. The low S-values yielded a high SPC because greater importance was given to the negative patterns; however, the SEN was low. As the S-value increased, the positive patterns had more importance while negative patterns showed reduced importance, which increased the SEN and lowered the SPC. The PPV was also reduced as the S-value increased and the number of false alarms also increased.

Cost-sensitive learning enabled the model to be fine-tuned to obtain a better response regarding PPV and SEN. The resulting PPV was 0.8125 and 0.7564 for the 0-d and 1-d prediction intervals, respectively, which showed improvements of more than 23% compared to the model published by Ramírez-Morales et al. (2016).

Imbalanced datasets might lead to overfitting of the training algorithms to the most common class and many mistakes in the least common class, leading to a poor generalisation performance (Huang et al. 2006; Blagus and Lusa 2010). A common solution to the overfitting problem in imbalanced datasets is using cost-sensitive learning ANN. This algorithm uses a parameter to weigh the errors in the most infrequent classes (Pazzani et al. 1994; Elkan 2001). Although cost-sensitive learning has attracted the attention of experts in machine learning and data mining, few studies have applied this concept to ANN training (Zhi-Hua and Xu-Ying 2006; Zahirnia et al. 2015).

Similar to other detection problems with low-frequency events, rare examples contain relevant information, such as an early warning system for industrial equipment malfunctioning. The detection of abnormal values in the production curve presents a similar challenge because production drops are usually not observed and few positive cases are found. Vannucci and Colla (2016) argued that traditional classification methods often fail when faced with imbalanced

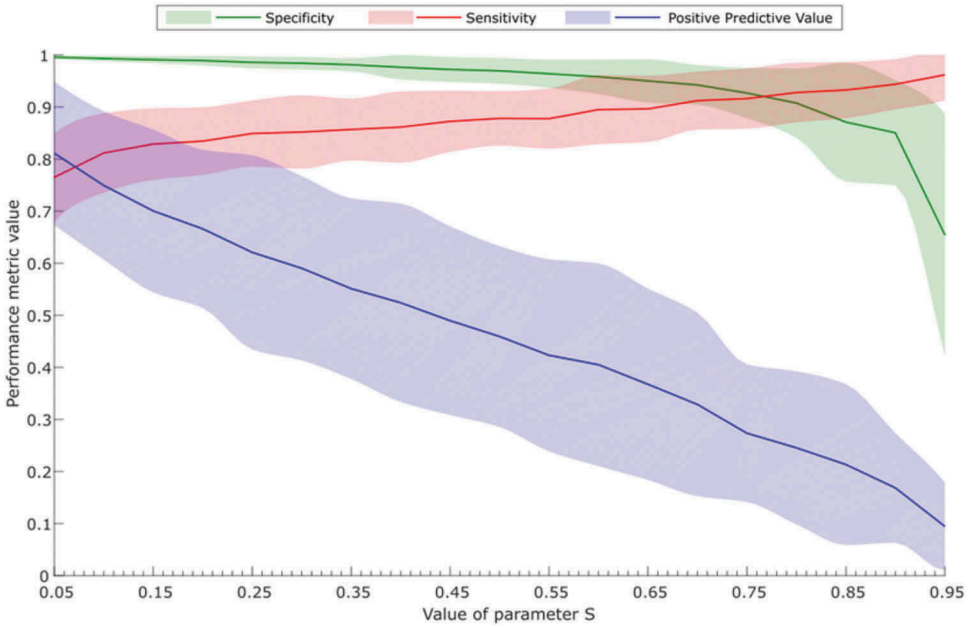


Figure 6. Performance chart of the model for different values of the parameter S.

datasets, which is mainly because the misclassification of rare samples is assumed to penalise the misclassification of frequent occurrences. Consequently, classifiers tend to focus on more frequent events. In the present paper, an adjustable parameter S was defined, from which an array of compensation weights was assigned to the unusual patterns. According to Elkan (2001), this adjustment allowed for the fine tuning of the model to obtain a better response regarding PPV and SEN.

A good system must maintain a proper balance between SPC and SEN along with a reasonable value of false alarms. For that reason, an S-value equal to 0.1 was chosen because at this point, SEN was maximised without excessively affecting PPV.

Performance analysis

To evaluate how well the model performed at different prediction intervals, data were reorganised by applying a day shift based on a time delay to the expected outputs (Martinerie et al. 1998; Kapoor and Bedi 2013). Therefore, in the 0-day prediction interval, the targets correspond to the same day of the pattern, whereas in the 5-d prediction interval, the targets correspond to 5 d later in the pattern. A range of prediction intervals was tested from 0 to 5.

The distribution of performance metrics results in 100 replicates of the 5-fold cross-validation as shown in Figure 7. The mean and standard deviation values for 6 prediction intervals are found in Table 2. The ACC and SPC values are above 0.9758 and 0.9874, respectively. However, a larger mean and a lower dispersion for values of 0 and 1 are observed as shown in Table 2, Figure 7(a,b). On average, the SEN results fall below 0.8 in the prediction interval values ranging from 2 to 5 as shown in Table 2 and Figure 7(c). The PPV value is kept above 0.7 until the prediction interval of 3 d as shown in Figure 7(d).

Few previous studies have analysed production parameters to warn of possible problems in egg production; thus, performing comparisons is difficult. The closest work is that of Lokhorst and Lamaker (1996), who proposed an expert system for egg production using a massive amount of data. These data were later turned into relevant information for the model, which obtained an SEN of 0.64 and an SPC of 0.72 in the same day of the anomaly. Compared to these

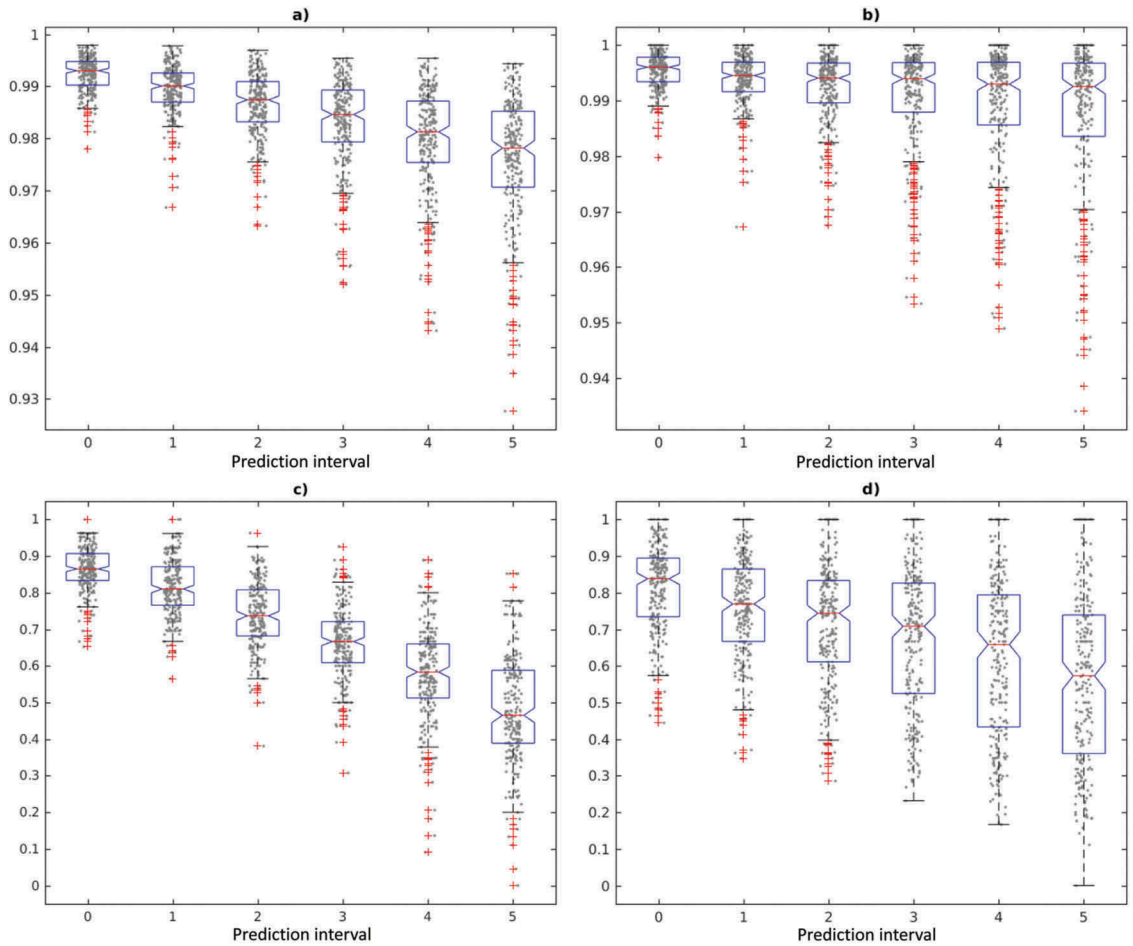


Figure 7. Boxplots of performance metrics for prediction intervals: (a) accuracy, (b) specificity, (c) sensitivity and (d) positive predictive value.

Table 2. Performance metrics for 5 prediction intervals.

	Prediction interval					
Performance	0 d	1 d	2 d	3 d	4 d	5 d
Accuracy	0.9925 ± 0.003	0.9896 ± 0.005	0.9864 ± 0.006	0.983 ± 0.009	0.9797 ± 0.01	0.9758 ± 0.012
Specificity	0.9954 ± 0.003	0.9936 ± 0.005	0.9923 ± 0.006	0.9905 ± 0.009	0.9891 ± 0.011	0.9874 ± 0.014
Sensitivity	0.8639 ± 0.061	0.8169 ± 0.078	0.7362 ± 0.091	0.6643 ± 0.104	0.5795 ± 0.132	0.4821 ± 0.152
Positive predictive value	0.8125 ± 0.119	0.7564 ± 0.14	0.7134 ± 0.166	0.6723 ± 0.19	0.6246 ± 0.22	0.567 ± 0.233

results, the model proposed in this paper obtained an SEN of 0.8639 and an SPC of 0.9954 along with an ACC of 0.9925 and a PPV of 0.8125 for the day prior to the anomaly.

Mertens et al. (2009) also proposed the development of a system based on an intelligent control chart for monitoring the egg production process. This system provided alarms for problems in the production curve for the same day that production dropped. The authors claimed that even when the production curve dropped, this situation remained unnoticed by the layer managers. However, it was not possible to compare this result with that of the present work because the system did not include calculated performance metrics. It should be highlighted that the model presented in the present paper is able to set alarms for good performance metrics up to 3 d before the problem is evident.

Xiao et al. (2011) examined in greater detail the concept of intelligent analysis by developing a software tool that performs statistical analyses of the production curve to identify possible problems. However, these authors did not provide performance metrics to allow for comparisons with our model.

Woudenberg et al. (2014) developed a real-time adaptive egg drop detection method that resulted in a perfect classification of problems for the same day. However, this method was validated in a single flock; thus, measuring its accuracy in a more realistic situation was impossible.

Finally, Ramírez-Morales et al. (2016) presented a model based on support vector machines and features elaborated by an expert. The ACC, SPC, SEN and PPV values were 0.9854, 0.9865, 0.9333 and 0.6135, respectively, and these results correspond to a prediction interval of 1 d before the occurrence of the anomaly. Although the model proposed here reached a lower SEN value of 0.8169, important improvements in the other three performance metrics were obtained (ACC of 0.9896, SPC of 0.9936 and PPV of 0.7564). Therefore, the proposed model moves from expert dependency to expert independency and achieves a greater than 23% improvement in the PPV. This improvement implies a lower number of false alarms and a significant advancement in the practical applicability of continually monitoring poultry production lots.

When the prediction intervals increase, the performance of the model decreases. SEN is the most affected parameter at higher prediction intervals. In the authors' experience, SEN values above 0.8 are acceptable. Therefore, the optimal prediction interval is 1 d.

Conclusions

The results clearly indicate that anomalies in the production of commercial laying hens can be automatically detected at an early date. The best configuration uses a feature selection threshold equal to 65, a sliding windows size of 18, an ANN with one hidden layer and 25 neurons, and a cost-sensitive learning parameter S-value of 0.1. The model proposed herein can be established as a detector of production drops if it is set for the same day because it has an ACC of 0.9925. Also, it can be configured as an early detector for the previous day because of its ACC of 0.9896. At the farm level, a 1-d advance prediction interval could be useful for on-farm decision-making. This tool would improve the preventive capacity in poultry production systems by providing automatically assisted monitoring as a complement to human observation, and such information is especially useful when handling large animal populations.

Acknowledgements

This work is part of the DINTA-UTMACH and RNASA-UDC research groups. We thank Agrolomas Ltd. for providing access to their data. We acknowledge the support of the Center of Supercomputing of Galicia (CESGA), which allowed the execution of the experimental stage of this work. Ivan Ramirez-Morales and Enrique Fernandez-Blanco would like to acknowledge NVidia because of their support through the grants research programme.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially funded by the Universidad Técnica de Machala, the Xunta de Galicia and the European Regional Development Fund (ERDF) as part of the project GRC2014/049.

ORCID

I. Ramírez-Morales http://orcid.org/0000-0002-2245-0016
E. Fernández-Blanco http://orcid.org/0000-0003-3260-8734
D. Rivero http://orcid.org/0000-0001-8245-3094
A. Pazos http://orcid.org/0000-0003-2324-238X

References

Abdi, H., and L. J. Williams. 2010. "Tukey's Honestly Significant Difference (HSD) Test." In *Encyclopedia of Research Design*, 1–5. Thousand Oaks, CA: Sage.

Amiri, A. S., T. A. Niaki, and A. T. Moghadam. 2014. "A Probabilistic Artificial Neural Network-Based Procedure for Variance Change Point Estimation." *Soft Computing* 19 (3): 691–700. doi:10.1007/s00500-014-1293-x.

Antonov, L. V., K. V. Makarov, and A. A. Orlov. 2015. "Development and Experimental Research on Production Data Analysis Algorithm in Livestock Enterprises." *Procedia Engineering* 129: 664–669. doi:10.1016/j.proeng.2015.12.088.

Anwar, H. M., B. V. Ayodele, C. K. Cheng, and M. R. Khan. 2016. ""Artificial Neural Network Modeling of Hydrogen-Rich Syngas Production from Methane Dry Reforming over Novel Ni/CaFe 2 O 4 Catalysts." *International Journal of Hydrogen Energy* 41 (26): 11119–11130.

Bennett, K. P., and C. Campbell. 2000. "Support Vector Machines." *ACM SIGKDD Explorations Newsletter* 2 (2): 1–13. doi:10.1145/380995.380999.

Blagus, R., and L. Lusa. 2010. "Class Prediction for High-Dimensional Class-Imbalanced Data." *BMC Bioinformatics* 11: 523. doi:10.1186/1471-2105-11-523.

Cameron, A. 2012. *Manual of Basic Animal Disease Surveillance*. Kenya: Interafrican Bureau for Animal Resources.

De Vries, A., and J. K. Reneau. 2010. "Application of Statistical Process Control Charts to Monitor Changes in Animal Production Systems." *Journal of Animal Science* 88 (13 Suppl): E11–24. doi:10.2527/jas.2009-2622.

Elkan, C. 2001. "The Foundations of Cost-Sensitive Learning." International joint conference on artificial intelligence, 2001.

Flanders, F., and J. R. Gillespie. 2015. *Modern Livestock & Poultry Production*. Boston, MA: Cengage Learning.

Frank, R. J., N. Davey, and S. P. Hunt. 2001. "Time Series Prediction and Neural Networks." *Journal of Intelligent and Robotic Systems* 31 (1/3): 91–103. doi:10.1023/a:1012074215150.

Frost, A. R., C. P. Schofield, S. A. Beaulah, T. T. Mottram, J. A. Lines, and C. M. Wathes. 1997. "A Review of Livestock Monitoring and the Need for Integrated Systems." *Computers and Electronics in Agriculture* 17 (2): 139–159. doi:10.1016/S0168-1699(96)01301-4.

Gates, M. C., L. K. Holmstrom, K. E. Biggers, and T. R. Beckham. 2015. "Integrating Novel Data Streams to Support Biosurveillance in Commercial Livestock Production Systems in Developed Countries: Challenges and Opportunities." *Front Public Health* 3: 74. doi:10.3389/fpubh.2015.00074.

Guo, L., D. Rivero, and A. Pazos. 2010. "Epileptic Seizure Detection Using Multiwavelet Transform Based Approximate Entropy and Artificial Neural Networks." *Journal of Neuroscience Methods* 193 (1): 156–163. doi:10.1016/j.jneumeth.2010.08.030.

Guyon, I., S. Gunn, M. Nikravesh, and L. A. Zadeh. 2008. *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Heidelberg: Springer..

Guyon, I., and A. Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3: 1157–1182.

Hepworth, P. J., A. V. Nefedov, I. B. Muchnik, and K. L. Morgan. 2012. "Broiler Chickens Can Benefit from Machine Learning: Support Vector Machine Analysis of Observational Epidemiological Data." *Journal of the Royal Society Interface* 9 (73): 1934–1942. doi:10.1098/rsif.2011.0852.

Herrera, F., C. Hervas, J. Otero, and S. Luciano 2004. "Un Estudio Empırico Preliminar Sobre Los Tests Estadısticos Más Habituales En El Aprendizaje Automático." *Tendencias de la Mineria de Datos en Espana, Red Espanola de Mineria de Datos y Aprendizaje (TIC2002-11124-E)*:403–412.

Huang, Y.-M., C.-M. Hung, and H. C. Jiau. 2006. "Evaluation of Neural Networks and Data Mining Methods on a Credit Assessment Task for Class Imbalance Problem." *Nonlinear Analysis: Real World Applications* 7 (4): 720–747. doi:10.1016/j.nonrwa.2005.04.006.

Kalhor, T., A. Rajabipour, A. Akram, and M. Sharifi. 2016. "Modeling of Energy Ratio Index in Broiler Production Units Using Artificial Neural Networks." *Sustainable Energy Technologies and Assessments* 17: 50–55. doi:10.1016/j.seta.2016.09.002.

Kapoor, P., and S. S. Bedi. 2013. "Weather Forecasting Using Sliding Window Algorithm." *International Scholarly Research Network Signal Processing* 2013: 1–5. doi:10.1155/2013/156540.

Kruse, R., C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held. 2013. "Multi-Layer Perceptrons." In *Computational Intelligence*, 47–81. London: Springer.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. New York: Springer.

Lindsay, D., and S. Cox. 2005. *Effective Probability Forecasting for Time Series Data Using Standard Machine Learning Techniques*, 35–44. Berlin: Springer.

Lokhorst, C., and E. J. J. Lamaker. 1996. "An Expert System for Monitoring the Daily Production Process in Aviary Systems for Laying Hens." *Computers and Electronics in Agriculture* 15 (3): 215–231. doi:10.1016/0168-1699(96)00017-8.

Long, A., and S. Wilcox. 2011. *"Optimizing Egg Revenue for Poultry Farmers." Poultry*, 1–10. Science.

Ma, X., Y. Zhang, and Y. Wang. 2015. "Performance Evaluation of Kernel Functions Based on Grid Search for Support Vector Regression." 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), 2015/7.

Martens, D., and B. Baesens. 2010. *"Building Acceptable Classification Models*, 53–74. Boston, MA: Springer.

Martinerie, J., C. Adam, M. Le Van Quyen, M. Baulac, S. Clemenceau, B. Renault, and F. J. Varela. 1998. "Epileptic Seizures Can Be Anticipated by Non-Linear Analysis." *Nature Medicine* 4 (10): 1173–1176. doi:10.1038/2667.

Mertens, K., I. Vaesen, J. Löffel, B. Kemps, B. Kamers, J. Zoons, P. Darius, E. Decuypere, J. De Baerdemaeker, and B. De Ketelaere. 2009. "An Intelligent Control Chart for Monitoring of Autocorrelated Egg Production Process Data Based on a Synergistic Control Strategy." *Computers and Electronics in Agriculture* 69 (1): 100–111. doi:10.1016/j.compag.2009.07.012.

Møller, M. F. 1993. "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning." *Neural Networks* 6 (4): 525–533. doi:10.1016/s0893-6080(05)80056-5.

Mucherino, A., P. J. Papajorgji, and P. M. Pardalos. 2009. *Data Mining in Agriculture. Vol. 34, Springer Optimization and Its Applications*. Edited by Panos M. Pardalos. New York, NY: Springer.

Narinc, D., F. Uckardes, and E. Aslan. 2014. "Egg Production Curve Analyses in Poultry Science." *Worlds Poultry Sciences Journal* 70 (04): 817–828. doi:10.1017/S0043933914000877.

Pazzani, M., C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. 1994. "Reducing Misclassification Costs." Proceedings of the Eleventh International Conference on Machine Learning, 1994.

Pica-Ciamarra, U., D. Baker, N. Morgan, and A. Zezza. 2014. *Investing in the Livestock Sector: Why Good Numbers Matter, A Sourcebook for Decision Makers on How to Improve Livestock Data*. Washington, DC: The World Bank and FAO.

Ramírez-Morales, I., D. Rivero-Cebrián, E. Fernández-Blanco, and A. Pazos-Sierra. 2016. "Early Warning in Egg Production Curves from Commercial Hens: A SVM Approach." *Computers and Electronics in Agriculture* 121: 169–179. doi:10.1016/j.compag.2015.12.009.

Refaeilzadeh, P., L. Tang, and H. Liu. 2009. "Cross-Validation." In *Encyclopedia of Database Systems*, edited by L. Liu and M. Tamer Özsu. 532-538. New York: Springer US.

Rivero, D., E. Fernandez-Blanco, J. Dorado, and A. Pazos. 2011. "Using Recurrent ANNs for the Detection of Epileptic Seizures in EEG Signals." 2011 IEEE Congress of Evolutionary Computation (CEC), 2011/6.

Saeed, K., and S. Václav 2014. *Computer Information Systems and Industrial Management: 13th IFIP TC 8 International Conference, CISIM 2014, Ho Chi Minh City, Vietnam, November 5-7, 2014*, Proceedings: Springer.

Saeys, Y., I. Inza, and P. Larranaga. 2007. "A Review of Feature Selection Techniques in Bioinformatics." *Bioinformatics* 23 (19): 2507–2517. doi:10.1093/bioinformatics/btm344.

Samborska, I. A., V. Alexandrov, L. Sieczko, B. Kornatowska, V. Goltsev, M. D. Cetner, and H. M. Kalaji. 2014. "Artificial Neural Networks and Their Application in Biological and Agricultural Research." *NanoPhotoBioSciences* 2: 14–30.

Schaefer, A. L., N. Cook, S. V. Tessaro, D. Deregt, G. Desroches, P. L. Dubeski, A. K. W. Tong, and D. L. Godson. 2004. "Early Detection and Prediction of Infection Using Infrared Thermography." *Canadian Journal of Animal Science* 84 (1): 73–80. doi:10.4141/a02-104.

Singh, P. K., R. Sarkar, and M. Nasipuri. 2015. "Statistical Validation of Multiple Classifiers over Multiple Datasets in the Field of Pattern Recognition." *International Journal of Applied Pattern Recognition* 2 (1): 1. doi:10.1504/ijapr.2015.068929.

Sun, Y., A. K. C. Wong, and M. S. Kamel. 2009. "Classification of Imbalanced Data: A Review." *International Journal of Pattern Recognition and Artificial Intelligence* 23 (04): 687–719. doi:10.1142/s0218001409007326.

Vannucci, M., and V. Colla. 2016. "Smart Under-Sampling for the Detection of Rare Patterns in Unbalanced Datasets." In *Intelligent Decision Technologies 2016*, edited by I. Czarnowski, A. M. Caballero, R. J. Howlett, and L. C. Jain, 395–404, Basel: Springer International Publishing.

Venkatesan, M., A. Thangavelu, and P. Prabhavathy. 2013. *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*. Vol. 202, *Advances in Intelligent Systems and Computing*. India: Springer India.

Wheeler, E. F., K. D. Casey, J. S. Zajaczkowski, P. A. Topper, R. S. Gates, H. Xin, Y. Liang, and A. Tanaka. 2003. "Ammonia Emissions from US Poultry Houses: Part III–broiler Houses." Air Pollution from Agricultural Operations-III, Raleigh, NC, October 12–15.

Woudenberg, S. P., D. Linda Gaag, A. Feelders, and A. R. Elbers. 2014. "Real-Time Adaptive Problem Detection in Poultry." In *Ecai 2014,* 1217–1218. Amsterdam: IOS Press.

Xiao, J., H. Wang, L. Shi, L. Mingzhe, and M. Haikun. 2011. "The Development of Decision Support System for Production of Layer." Computer and Computing Technologies in Agriculture V, Beijing, October 29-31. Berlin: Springer.

Zahirnia, K., M. Teimouri, R. Rahmani, and A. Salaq. 2015. "Diagnosis of Type 2 Diabetes Using Cost-Sensitive Learning." 2015 5th International Conference on Computer and Knowledge Engineering (ICCKE), 2015/10.

Zhi-Hua, Z., and L. Xu-Ying. 2006. "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem." *IEEE Transactions on Knowledge and Data Engineering* 18 (1): 63–77. doi:10.1109/tkde.2006.17.