Universidad de Valladolid

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TRABAJO DE FIN DE MÁSTER

MÁSTER EN INGENIERÍA DE TELECOMUNICACIÓN

Técnicas de visión por ordenador para la estimación automática y no intrusiva del grado de acidez en cítricos

Autor:

Fernando García Gómez

Tutor:

Juan Ignacio Arribas Sánchez

Valladolid, Septiembre de 2019

TÍTULO:	**"Técnicas de visión por ordenador para la estimación automática y no intrusiva del grado de acidez en cítricos"**
AUTOR:	Fernando García Gómez
TUTOR:	Juan Ignacio Arribas Sánchez
DEPARTAMENTO:	Teoría de la Señal y Comunicaciones e Ingeniería Telemática

TRIBUNAL

PRESIDENTE:	Belén Carro Martínez
VOCAL:	César Gutiérrez Vaquero
SECRETARIO:	Juan Pablo Casaseca de la Higuera

FECHA:	Septiembre de 2019
CALIFICACIÓN:	

RESUMEN DEL TFM

El estudio y estimación de las propiedades de los productos generados en la industria alimentaria de forma no intrusiva es una de las aplicaciones más utilizadas para evaluar la calidad del producto ofrecido. Dentro del conjunto de productos derivados de esta industria, cobran gran importancia la gran variedad de frutas que existen en el planeta, y en especial los cítricos, con la naranja como su máximo exponente, debido a su gran aporte de vitamina C.

Una de las propiedades más importantes de la naranja es el grado de acidez (pH), que es indicativo del nivel de madurez de la fruta y también sirve de ayuda para diferenciar entre las distintas variedades que existen en el planeta.

A partir de la segmentación y el análisis de un conjunto de imágenes capturadas de diferentes variedades de naranjas, es posible la extracción de características que definan cada imagen. Utilizando una serie de mecanismos de aprendizaje automático se tendrá la capacidad de estimar el grado de acidez de cada naranja disponible y de realizar una comparativa entre este pH estimado y el pH real medido de tal forma que se seleccione el mecanismo que mejor rendimiento ofrezca a la hora de estimar correctamente el grado de acidez de cada naranja analizada.

PALABRAS CLAVE

Cítricos, Naranja, pH, *Bam*, *Blood*, *Thomson*, *3-variety*, Análisis de Componentes Principales (PCA), R, MSE, RMSE, MAE, SSE, Máquina de Vectores Soporte (SVM), Perceptrón Multicapa (MLP), Algoritmo Competitivo Imperialista (ICA).

ABSTRACT

The study and estimation of products properties in food industry in a non-intrusive way is one of the best applications in order to evaluate the product quality. One of the most important products of this industry are the great variety of fruits that exist in the planet, especially citrus, with the orange as its best variety, due to its contribution of C vitamin.

One of the most important properties of orange is the acidity degree (pH), which is indicative of fruit level maturity and also helps to discriminate between varieties on the planet.

Starting from segmentation and the analysis of the image set captured from different orange varieties, it is possible to extract features to define each image. Using a set of automatic learning machines, it will be possible to estimate the acidity degree of each available orange and to make a comparison between estimated and real measured pH in order to choose the best performance machine to estimate the acidity degree of each analyzed orange.

KEYWORDS

Critics, Orange, pH, *Bam*, *Blood*, *Thomson*, *3-variety*, Principal Component Analysis (PCA), R, MSE, RMSE, MAE, SSE, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Imperialist Competitive Algorithm (ICA).

A mis padres, por su apoyo.
A Juan Ignacio, por su ayuda a pesar de las dificultades.
A María, por su paciencia y optimismo.
A Carmen, porque juntos saldremos adelante.

El 90 % del éxito se basa simplemente en insistir.
-Woody Allen-

Lista de Siglas y Acrónimos

ANFIS: Adaptive Neuro-Fuzzy Inference System.
ANN: Artificial Neural Network. Red Neuronal Artificial.
GLCM: Gray Level Co-ocurrence Matrix. Matriz de Co-ocurrencia de Nivel de Gris.
ICA: Imperialist Competitive Algorithm. Algortirmo Imperialista Competitivo.
MAE: Mean Absolute Error. Error Medio Absoluto.
MLP: MultiLayer Perceptron. Perceptrón Multicapa.
MSE: Mean Squared Error. Error Cuadrático Medio.
PCA: Principal Component Analysis. Análisis de Componentes Principales.
pH: Pontential of hydrogen. Potencial de hidrógeno.
PSO: Particle Swarm Optimization. Optimización por Enjambre de Particulas.
R: Coeficiente de correlación lineal.
$\mathbf{R}^2$: Coeficiente de determinación.
RMSE: Root Mean Squared Error. Raíz del Error Cuadrático Medio.
SSE: Sum of Squared Errors. Suma de Errores Cuadráticos.
SVD: Singular Value Descomposition. Descomposición en Valores Singulares.
SVM: Support Vector Machine. Máquina de Vectores Soporte.
Tr.: Contenido de trazas.

Índice general

Índice de figuras

Índice de tablas

1

Introducción

Las frutas son un conjunto de alimentos que engloban una parte muy importante de la alimentación del ser humano ya que son las encargadas de proveer al organismo de una serie de nutrientes necesarios como son vitaminas, sales minerales y antioxidantes. Es posible clasificar la fruta en función de su variedad y diferentes propósitos de uso atendiendo a un gran número de propiedades, de las cuales una de las más importantes es el grado de acidez (pH), lo que hace una cuestión muy interesante el conocimiento del mismo a la hora de evaluar la calidad del producto.

En este capítulo se expondrá la motivación del proyecto propuesto así como los diferentes objetivos perseguidos y los recursos empleados para la consecución de los mismos. Por último, se presentara la estructura a seguir del documento.

1.1. Motivación

La estimación de las propiedades de diferentes productos de forma no intrusiva es una de las principales aplicaciones de la industria alimentaria a la hora de reconocer la calidad del producto ofrecido. En los últimos años se han realizado estudios sobre diversos campos de la industria alimentaria como la clasifición de peces en función del color y la textura [1], predicción de los niveles de azucar y dureza en melones [2] o en el análisis de daños debido a plagas en la planta de colza [3]. Dentro del conjunto de productos derivados de la industria alimentaria cobran gran importancia la gran variedad de frutas que existen en el planeta. Se han realizado estudios para realizar la estimación del volumen y la masa en papayas a partir del análisis de imágenes [4] y de su procesado utilizando filtros ExG (Excess Green), obteniendo resultados en la estimación y clasificación de un 90 % de acierto. Por otro lado, se han investigado también diferentes métodos de segmentación para la clasificación de manzanas utilizando máquinas de vectores soporte (*Support Vector Machine*, SVM) y el método Otsu [5], sobre una base de datos de 300 imágenes de manzana roja y de un conjunto de características extraídas sobre las mismas,

obteniendo unos porcentajes de error en la clasificación menores del 2 %.

De las variedades de frutas existentes cabe destacar la importancia de los cítricos, que son uno de los tipos de fruta más cultivada en el mundo con la naranja como mayor exponente de esta variedad.

A la hora de clasificar las diferentes variedades de cítricos se tienen en cuenta determinadas propiedades que tienen que ver tanto con la fisionomía como con las propiedades químicas del producto. En relación con las naranjas, base del presente proyecto, numerosos estudios has sido realizados a la hora de estimar la calidad y el uso de los distintos tipos y variedades de esta fruta, en particular se han estudiado la masa [6] a partir de la segmentación de las imagenes de 300 naranjas de la variedad *Thomson* y del uso del método ANFIS, proporcionando altas tasas de acierto en la estimación; y el grosor de la piel [7] utilizando una red neuronal híbrida PSO-GA. A partir de técnicas de aprendizaje automático, muy en auge en la industria alimentaria, se ha conseguido estimar diferentes propiedades de las mencionadas anteriormente.

Por lo tanto, a partir de estas técnicas de aprendizaje automático y un conjunto de muestras de determinadas variedades de naranjas, la motivación del presente proyecto será la estimación del grado de acidez (pH) para tres variedades de naranjas diferentes y comprobar si los resultados obtenidos ofrecen una precisión en la estimación comparable a los estudios realizados recientemente mediante técnicas de aprendizaje automático para este tipo de fruta [8].

1.2. Objetivos

Los principales objetivos que se persiguen en el presente proyecto son los siguientes:

- Estudio de los cítricos, en especial las variedades de naranjas utilizadas, y compresión de la propiedad relativa al grado de acidez de las mismas (pH).
- Elaboración de un conjunto de datos manipulable a partir de la base de datos de imágenes relativa a las tres variedades de naranjas disponibles (*Bam*, *Blood* y *Thomson*) basado en un conjunto de características de distinta índole útiles a la hora de describir cada muestra.
- Desarrollo y utilización de un *software* en MATLAB que lleve a cabo, por un lado, la extracción de diferentes parámetros característicos de las imágenes de cada naranja y, por otro lado, la reducción de la dimensionalidad de todo el conjunto.
- Estimación del grado de acidez (pH) de cada muestra a partir de una serie de algoritmos de aprendizaje automático y su comparación con el grado de acidez real de cada una de las muestras.

- Comparación de las diferentes técnicas de aprendizaje automático no intrusivas utilizadas a partir los resultados y gráficos obtenidos.
- Obtener las conclusiones pertinentes a partir de los resultados obtenidos en el presente proyecto.

1.3. Recursos

Los medios que se emplearán para llevar a cabo los objetivos del proyecto son los siguientes:

- Ordenador portátil conectado a Internet con *Windows 10* como sistema operativo.
- Entorno de programación MATLAB R2017b y uso de la aplicación *Neural Network*.
- Paquete *Microsoft Office*.
- Sistema de composición de textos *LaTex TexMaker*.
- Base de datos de naranjas proporcionada por Sajad Sazbi de tres variedades distintas (*Bam*, *Blood* y *Thomson*), con 100 muestras en formato de imagen *jpg* para cada una de las tres variedades, sumando en su conjunto un total de 300 imágenes.

1.4. Estructura

La estructura diseñada para el documento será la siguiente:

- **Capítulo 2. Estudio de los cítricos: la naranja.** Descripción y explicación de los cítricos, y en especial de la naranja prestando atención a los tipos de naranjas utilizadas en la base de datos del proyecto. Por otro lado se explicara el concepto de grado de acidez.
- **Capítulo 3. Métodos.** Extracción de un conjunto de parámetros de diferente índole capaces de caracterizar cada una de las imágenes de la base de datos de naranjas disponible. Una vez obtenidas las características se llevara a cabo una reducción de la dimensionalidad del banco de datos obtenido sin producir pérdida de información relevante a fin de aplicar diferentes mecanismos de aprendizaje automático para estimar el grado de acidez de las naranjas que conforman la base de datos propuesta.

- **Capítulo 4. Resultados.** Extracción de valores y gráficos para comparar los valores estimados y reales medidos de pH de cada una de las naranjas de la base de datos disponible. Además se realizará una comparación entre distintos mecanismos de aprendizaje automático a fin de seleccionar aquellos que menor error en la estimación ofrezcan.

- **Capítulo 5. Conclusiones y lineas futuras.** Extracción de las conclusiones más relevantes a partir de los resultados obtenidos y de las lineas futuras posibles para mejorar el presente proyecto.

2

Estudio de los cítricos: la naranja

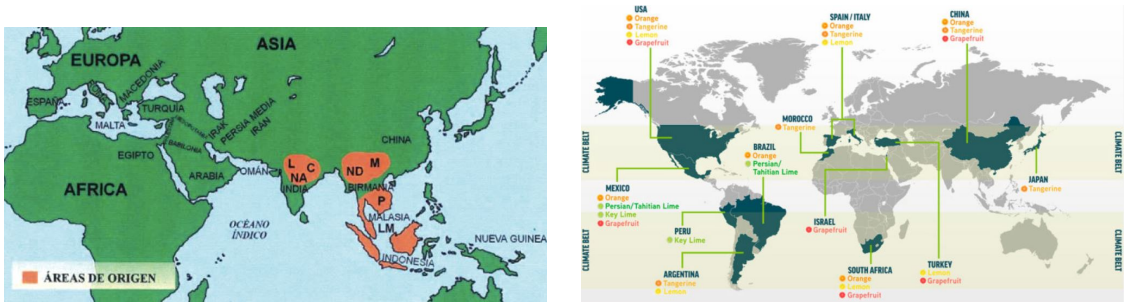
Como se ha explicado anteriormente, los cítricos se pueden clasificar en función de una gran variedad de propiedades con el fin de su asignación hacia diferentes ámbitos de la industria alimentaria. En el presente capítulo se presentara el cítrico, sus propiedades y sus variedades, de las cuales se prestará especial atención a la naranja y a las tres clases que forman la base de datos disponible: *Bam*, *Blood* y *Thomson*. Por otro lado, se describirá el concepto de grado de acidez, que será la propiedad principal a la hora de clasificar y evaluar cada tipo de naranja propuesta. Por último, se analizara la base de datos de imágenes disponible con el objetivo de preparar los datos para su posterior procesado y extracción de las características necesarias para la definición completa de cada imagen.

2.1. El cítrico en la industria alimentaria

La citricultura es una rama de la fruticultura que estudia los cultivos y características de un grupo de plantas denominadas cítricos. Son una de las especies arbóreas más cultivadas en todo el mundo y su gran variedad de especies se desarrolla principalmente en casi todas las regiones del mundo situadas en una latitud de 40 grados [9].

El origen de su cultivo se remonta desde hace más de 4000 años en la zona geográfica sudeste asiático cuyos climas tropicales y subtropicales eran propicios para su explotación y crecimiento. Desde allí se han ido extendiendo al resto de regiones del planeta donde se cultivan actualmente [10]. Los primeros exploradores occidentales al continente asiático fueron los encargados de extender esta variedad frutal por el resto de continentes. En torno al ano 300 a.C llegaron a Europa mientras que hasta el año 1500 d.C no fueron llevados hasta el continente americano. Respecto a España, el naranjo y el limonero fueron introducidos por los árabes entre los siglos XI y XII; la mandarina común fue introducida en 1845, mientras que el pomelo es la última de las variedades cítricas más importantes que se incorporo a la península, en 1910. En la actualidad, los principales

países productores de cítricos son China, Brasil, Estados Unidos y México.



(a) (b)

Figura 2.1: a) Zona geográfica de origen de las primeras plantaciones de cítricos [11] y b) Distribución geográfica de las zonas de mayor cultivo y producción de cítricos [cit].

Atendiendo a las características y propiedades de los cítricos se pueden distinguir las siguientes [9]:

- Los árboles cítricos se caracterizan por ser árboles o arbustos de hoja perenne, florecen en primavera y los frutos se recogen a partir del otoño hasta el invierno.
- Su cultivo requiere climas cálidos y una humedad abundante.
- Los frutos son grandes y generalmente de piel gruesa, y su sabor es ácido.
- Se caracterizan por ser fuente de vitaminas (vitamina C, ácido fólico y vitamina A), minerales (potasio) y flavonoides. De aquí vienen sus propiedades antioxidantes.

2.1.1. Variedades

Existen infinidad de cítricos en el mundo, cada uno de ellos con sus propias especies. Las variedades más cultivadas y consumidas en la actualidad son cuatro: naranja, limón, mandarina y pomelo [12]:

- **Naranja**: La naranja es un fruto de piel gruesa y sabor dulce cuyo nombre proviene por el característico color del mismo y cuyos tres principales productores son Brasil, Estados Unidos y México. Se caracteriza por sus propiedades antioxidantes debidas a la gran cantidad de vitamina C que contiene.
- **Limón**: Es una fruta comestible de color amarillo perteneciente a la familia de las rutáceas de sabor ácido que crece en los limoneros y cuyo principal productor es México seguido por la India. Sus principales propiedades son antioxidantes y anti-bacterianas [13].

- **Mandarina**: es un cítrico similar a la naranja aunque de menor tamaño, sabor menos ácido al tener mayor cantidad de azucar y con una piel más fina. Su origen se remonta a las zonas tropicales de China, siendo en la actualidad los mayores productores China, España y México. Sus principales propiedades son antioxidantes.
- **Pomelo**: Es un fruto carnoso de corteza gruesa de color amarillento y sabor agrio que crece en el árbol rutáceo cultivado en lugares tropicales y subtropicales como Bolivia, y que alberga gran variedad de beneficios para la salud gracias a sus propiedades antioxidantes, como la pérdida de peso y reducción del colesterol [14].

Figura 2.2: De izquierda a derecha: naranja [nar], mandarina [man], limón [lim] y pomelo [pom].

2.1.2. La naranja

Es el fruto obtenido del árbol frutal *citrus sinesis*, uno de los frutales más cultivados del mundo y cuyo origen se sitúa en el sur de China y el nordeste de la India. El fruto recolectado se caracteriza por una pulpa jugosa y tierna de color amarillo, naranja o rojizo, con presencia de semillas poliembriónicas en mayor o menor medida. El zumo que proporcionan es abundante y de alta calidad debido al equilibrio entre azucares y acidez [12], [15]. El principal uso de la naranja dulce es el consumo, ya sea en fresco o en zumo preparado. Las principales propiedades de la naranja es que es un alimento con un alto contenido en vitamina C (ver Tabla 2.1) y con efectos antioxidantes que favorecen la cicatrización y refuerzan el sistema inmunológico del organismo.

El gran número de variedades existentes de naranjas dulces se clasifican en cuatro grupos: grupo Navel, grupo Blancas, grupo Sangre y grupo Sucreñas [12].

Grupo Navel

Las naranjas del grupo Navel se diferencian del resto por tener un pequeño fruto rudimentario en la zona estilar que recuerda a un ombligo (*navel*) y carecer de semillas. Es el principal tipo de naranja empleado para el consumo en fresco debido a su excelente calidad y a su sabor dulce, mientras que no son recomendables para la elaboración de

Valores	**Por cada 100g de porción comestible**
Energía (Kcal)	42
Proteínas (g)	0.8
Lípidos totales (g)	Tr.
Hidratos de carbono (g)	8.6
Fibra (g)	2
Agua (g)	88.6
Calcio (mg)	36
Hierro (mg)	0.3
Yodo (μg)	2
Magnesio (mg)	12
Zinc (mg)	0.18
Sodio (mg)	3
Potasio (mg)	200
Fósforo (mg)	28
Selenio (μg)	1
Tiamina (mg)	0.1
Riboflavina (mg)	0.03
Equivalentes niacina (mg)	0.3
Vitamina B6 (mg)	0.06
Folatos (μg)	37
Vitamina B12 (μg)	0
Vitamina C (mg)	50
Vitamina A: Eq. Retinol (μg)	40
Vitamina D (μg)	0
Vitamina E (mg)	0.2

Tabla 2.1: Composición nutricional de la naranja (*citrus sinensis*)

zumos. Las variedades más destacables son las siguientes: *Navelina*, *Washington navel*, *Lanelate* y *Thomson*.

Grupo Blancas

Engloba todos aquellos tipos de naranjas que no poseen ombligo (navel), pigmentación roja o una acidez escasa. Son utilizadas en su gran mayoría para la fabricación de zumos. Las variedades más destacables son: *Valencia*, *Pera* y *Salustiana*.

Grupo Sangre

Son similares a las blancas, distinguiéndose por la pigmentación rojiza presente tanto en la corteza como en la pulpa y el zumo. Sus variedades más destacables son: *Tarocco* y *Sanguinelli*.

Grupo Sucreñas

Se distinguen de las del grupo Blancas por un contenido en ácido extremadamente bajo. Variedades más destacables: *Sucreña* o *Imperial*.

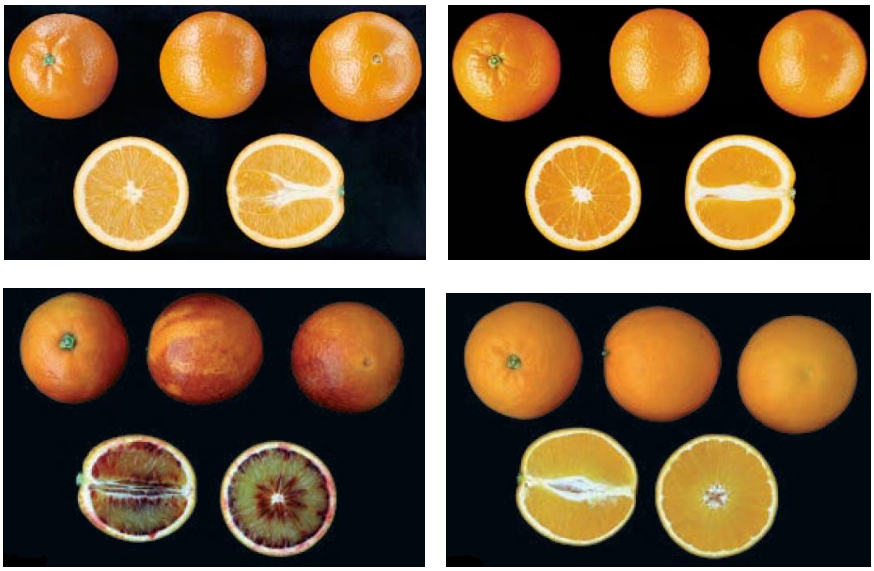


Figura 2.3: De izquierda a derecha y de arriba hacia abajo: navel (navelina), blanca (salustiana), sangre (sanguinelli) y sucreña (sucreña) [12].

2.2. Grado de acidez: el pH

El pH es el coeficiente que indica la medida de acidez o basicidad de una disolución. Es el logaritmo negativo de la concentración de iones de hidrógeno (en mol/litro) en un determinado medio [16]:

$$pH = -\log_{10}[H^+] \tag{2.1}$$

donde H^+ representa la concentración de iones de hidrógeno en el medio.

Debido a que el pH sólo es una manera de expresar la concentración del ion hidrógeno, las disoluciones ácidas y básicas se identifican por sus valores pH como se muestra a continuación [17]:

- **Disoluciones ácidas**: $[H^+] > 1 \cdot 10^{-7} \quad mol/litro$, $pH < 7$
- **Disoluciones básicas**: $[H^+] < 1 \cdot 10^{-7} \quad mol/litro$, $pH > 7$
- **Disoluciones neutras**: $[H^+] = 1 \cdot 10^{-7} \quad mol/litro$, $pH = 7$

Teniendo esto en cuenta, se implementa la escala de valores del pH, con un rango desde el valor 0 (el más ácido) hasta el 14 (el más básico) y con un etiquetado por colores en intervalos de 1 indicando el grado de acidez o basicidad de la disolución en cuestión.

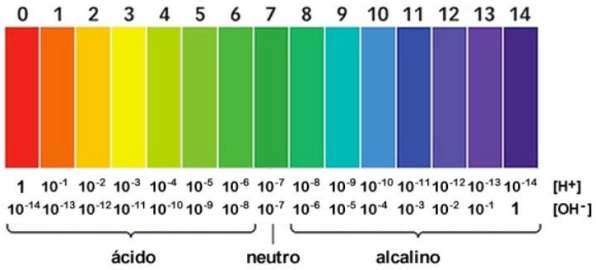


Figura 2.4: Escala de valores del pH [ph].

2.2.1. Medición del pH

El método más utilizado para medir el pH se basa en la utilización de un potenciómetro o pH-metro que mide la diferencia de potencial de los electrodos: un electrodo de referencia externo y otro electrodo de vidrio que es sensible al ion de hidrógeno [18].

El electrodo de vidrio utilizado para la medida del pH está formado por un filamento de vidrio compuesto por tres componentes principales: un 72 % de dióxido de silicio (SiO_2), 22 % de óxido de sodio (Na_2O) y un 6 % de óxido de calcio (CaO). Contiene en su interior una solución ácida de ácido clorhídrico de normalidad 0.1 ($0{,}1N\ HCl$) combinada con una solución saturada de cloruro de plata ($AgCl$), y un electrodo interno de referencia de cloruro de plata. El electrodo de vidrio se pone en contacto con la mezcla propuesta.

El segundo electrodo, el de referencia externo formado por cloruro de plata también se encuentra inmerso en la mezcla propuesta y contiene en su extremo una solución de cloruro de potasio (KCl).

La diferencia de potencial E entre ambos terminales, el electrodo de vidrio y el electrodo externo de referencia, se define como:

$$E = E_{in} - \Delta E_m - \Delta E_{lj} - E_{ext} \tag{2.2}$$

donde E_{in} y E_{ext} son los potenciales de cada uno de los electrodos del diseño, ΔE_m simboliza la caída de potencial entre los extremos de la membrana de vidrio, y ΔE_{lj} la caída de potencial en la unión electrolítica del electrodo externo de referencia. Considerando E_{in}, E_{ext} y ΔE_{lj} como valores constantes y definiendo la caída de potencial δE_m mediante la ecuación de Nernst [19], [20]:

$$E = E_{const} - \Delta E_m \tag{2.3}$$

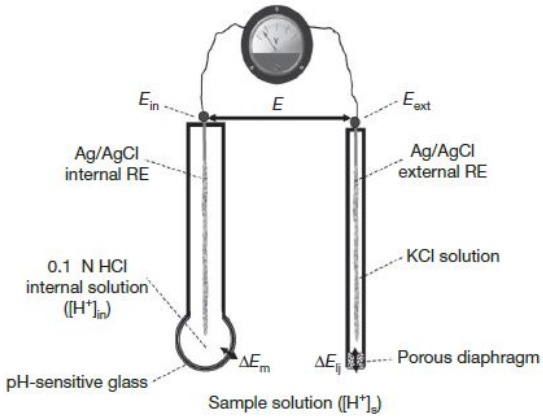


Figura 2.5: Representación esquemática de un medidor de pH con el electrodo de vidrio separado del electrodo externo de referencia [19].

$$\Delta E_m = 2{,}303 \cdot \frac{RT}{F} \cdot log_{10} \frac{[H^+]_{in}}{[H^+]_s} \tag{2.4}$$

donde R es la constante de los gases ideales en Julios, T la temperatura en Kelvin, F la constante de Faraday, $[H^+]_{in}$ la concentración de iones de hidrógeno de la solución interna del electrodo de vidrio, que se considerará constante, y $[H^+]_s$ la concentración de iones de hidrógeno de la solución propuesta. Por lo tanto, con estas premisas y sabiendo que $pH_s = -log_{10}[H^+]_s$, la diferencia de potencial se define como:

$$E = E_{const} - 2{,}303 \cdot \frac{RT}{F} \cdot pH_s \tag{2.5}$$

indicando que el pH es una función lineal de la diferencia de potencial medido.

Para la realización de la medida del pH real de cada una de las muestras de naranjas disponibles en el presente proyecto, se ha utilizado un pH-metro (pH/ORP/medidor de temperatura) fabricado en Taiwan.

2.2.2. El pH en la industria alimentaria

El conocimiento del valor del pH en los alimentos tiene una importancia relevante desde varias perspectivas. Es un factor de suma importancia en la absorción de agua y, por otro lado, afecta significativamente las propiedades químicas y físicas de ciertos componentes del alimento como son las proteínas, azúcares y aminoácidos [19].

Por lo general, la mayoría de los alimentos tienen un valor de pH por debajo de 7, y el valor del grado de acidez de aquellos alimentos que provienen de plantas es menor que el valor de los alimentos de origen animal. Con estas consideraciones los alimentos se pueden clasificar en alto contenido en ácido (pH <3.7), ácidos (pH = 3.7-4.6), contenido en ácido medio (pH = 4.6-5.3) y bajo contenido en ácido (pH >5.3).

Atendiendo a la medición del valor del grado de acidez en cítricos, este grupo de alimentos se pueden considerar ácidos o con un alto contenido en ácido, siendo el limón el cítrico con el pH medio más bajo y la mandarina con el valor más elevado:

Tipo de cítrico	**pH medio**
Naranja (genérico)	3.6
Naranja variedad *Bam*	3.77
Naranja variedad *Blood*	3.80
Naranja variedad *Thomson*	4.68
Mandarina	4.0
Limón	2.4
Pomelo	3.3

Tabla 2.2: Grado de acidez (pH) medio para los cuatro cítricos principales y las variedades de naranja *Bam*, *Blood* y *Thomson*.

2.3. Utilización de la base de datos

Una vez definidas los diferentes grupos de naranjas y teniendo en cuenta que la base de datos disponible está formada por tres variedades de naranjas: *Bam* (grupo Blancas), *Blood* (grupo Sangre) y *Thomson* (grupo Navel), la primera etapa consiste en la preparación de las imágenes para su posterior procesado y extracción de las características pertinentes. La base de datos está formada por 300 imágenes en formato *jpg* con un tamaño medio de 250 x 250 píxeles capturadas en un entorno con las siguientes características [21]:

Figura 2.6: Entorno de captura de imágenes de cada una de las muestras de naranjas disponibles [21].

- Cámara digital BOSCH fabricada en Alemania y situada a una distancia de 10 centímetros de la muestra a capturar.
- Estructura con tres tipos de lámparas utilizados (LED, fluorescente y tungsteno) a una intensidad de 50.33 luxes para evitar cualquier tipo de sombra en la muestra capturada.
- Superficie de color negro para facilitar la segmentación de la muestra capturada.

Es necesario separar la naranja del fondo (Figura 2.7) [22]: a partir del espacio de color *RGB* (Red, Green, Blue) se realiza un reconocimiento de la imagen respecto al binomio fruta/fondo píxel a píxel en función del siguiente umbral:

$$\begin{aligned} &if\ (R(x,y)+G(x,y)+B(x,y))/3 < 40\ or\ R(x,y) < 100 \\ &then \quad (x,y) \text{ is background} \end{aligned} \tag{2.6}$$

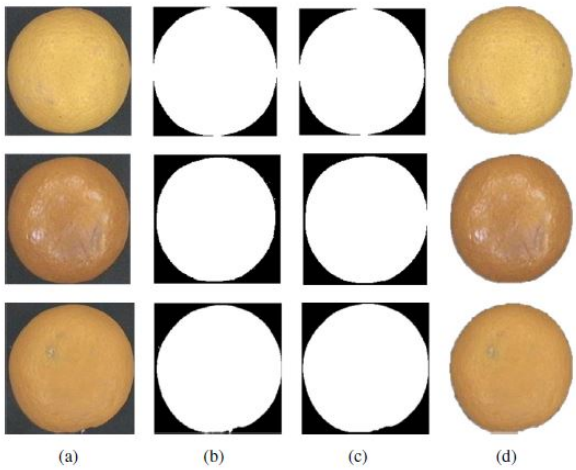


Figura 2.7: Etapas del proceso de segmentación para tres muestras de naranjas de la base de datos: a) imagen original. b) imagen binaria tras la clasificación de los píxeles. c) imagen binaria mejorada. d) imagen segmentada resultante [22].

Una vez que las imágenes han sido segmentadas se puede comenzar el estudio, procesado y extracción de características a realizar en el presente proyecto. El banco de datos estará formado por el siguiente conjunto de imágenes segmentadas (ver Apéndice A):

- Conjunto de 100 imágenes correspondientes a 100 muestras de la variedad de naranja *Bam* perteneciente al grupo Blancas, junto con los pH reales medidos para cada una de las muestras.

- Conjunto de 100 imágenes correspondientes a 100 muestras de la variedad de naranja *Blood* perteneciente al grupo Sangre, junto con los pH reales medidos para cada una de las muestras.
- Conjunto de 100 imágenes correspondientes a 100 muestras de la variedad de naranja *Thomson* perteneciente al grupo Navel, junto con los pH reales medidos para cada una de las muestras.

3

Métodos

A la hora de analizar una base de datos es muy importante la selección de una serie de características que sean capaces de definir cada uno de los elementos de la misma de forma unívoca, en este caso, cada una de las imágenes y valores de pH que conforman el banco de datos.

En este capítulo se abordará la extracción de un conjunto de características de interés a la hora de describir cada imagen segmentada de la base de datos propuesta, para después realizar una reducción de la dimensionalidad del conjunto de datos a partir de la técnica conocida como Análisis de Componentes Principales (*Principal Component Analysis*, PCA). Por último se explicarán diferentes mecanismos de aprendizaje automático utilizados para realizar una regresión lineal que permita estimar el grado de acidez de las naranjas que conforman la base de datos propuesta.

3.1. Extracción de características

El proceso de extracción de características es una etapa muy importante a fin de caracterizar un evento en concreto. En el caso que concierne a este proyecto la extracción se centrará en el análisis de las diferentes imágenes de naranjas de cada variedad que conforman la base de datos. Dichas características extraídas representarán de la forma más eficiente posible cada una de las muestras disponibles.

Como se ha comentado anteriormente las características extraídas para cada imagen serán elegidas de tal forma que su representación sea lo más fiable posible en términos de similitud. Para conseguirlo, se elegirán de acuerdo con diferentes criterios, siendo el más importante la clase a la que pertenecen cada una de ellas [22], así, se dividirán en los siguientes grupos: color, textura, momentos invariantes, histograma y forma.

3.1.1. Color

Las características de color son una potencial fuente de información a la hora de analizar las imágenes de la base de datos disponible ya que ofrecen grandes ventajas como su simplicidad, eficiencia y robustez. Su uso está extendido en numerosas aplicaciones tales como monitorización de cultivos, control de calidad en frutas y verduras, y monitorización por satélite [23]. El objetivo será extraer todas aquellas características más discriminantes que ayuden a extraer la mayor cantidad de información posible de la base de datos propuesta [22]. Dentro de las características de color se pueden distinguir dos grupos fundamentales: espacios de color e índices de vegetación. El primer grupo se encargará del análisis de las tres componentes de cada uno de los espacios de color propuestos a fin de conocer su media y su desviación típica; en cuanto al segundo grupo de características estará formado por 14 valores relacionados con los indices de vegetación propuestos por diferentes autores [22], [24].

Espacios de color

Un espacio de color es una representación matemática de un conjunto de colores que conforman una imagen o vídeo. Está formado habitualmente por tres componentes distintas que describen la imagen en función de una serie de características conocidas como son la luminancia, la saturación, la tonalidad, el nivel cromático, etc [25].

En el presente apartado se especificarán los 16 espacios de color utilizados a la hora de obtener las características deseadas de las diferentes imágenes que conforman la base de datos disponible.

Espacio de color	**Características**						
RGB, CMYK, YPbPr, YCbCr, JPEG-YCbCr, YDbDr, YUV, YIQ, HSV, HSL, HSI, XYZ, Lab, Luv, LCH, CAT02-LMS	First component mean	Second component mean	Third component mean	Three components mean together	Standard deviation of the first component	Standard deviation of the second component	Standard deviation of the third component

Tabla 3.1: Características de los diferentes espacios de color propuestos.

RGB El espacio de color RGB (Red, Green, Blue) es comúnmente utilizado en la representación gráfica de computadoras y dispositivos electrónicos, así como para procesamiento y almacenamiento de imágenes [25]. Las tres componentes del espacio

de color (Rojo, Verde y Azul) se consideran las componentes primarias del color y su mezcla da como resultado el resto de colores deseados del espectro.

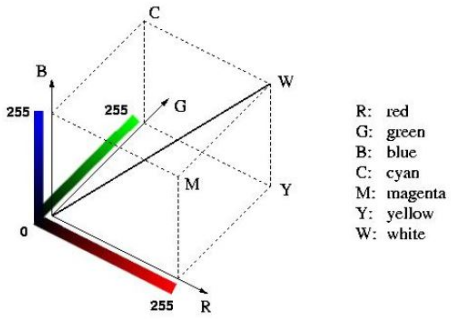


Figura 3.1: Representación de los espacios de color RGB y CMYK [25].

CMYK El espacio de color CMYK (Cyan, Magenta, Yellow, Key) se utiliza en impresión de documentos e imágenes a color. Está formado por tres componentes principales: **C** (Cyan), **M** (magenta) e **Y** (Amarillo), y una cuarta componente **K** referente al color negro [26]. Es posible realizar la conversión del espacio de color RGB al CMYK a partir de las siguientes ecuaciones [27]:

$$C = 1 - R \tag{3.1}$$

$$M = 1 - G \tag{3.2}$$

$$Y = 1 - B \tag{3.3}$$

YPbPr Es un espacio de color utilizado en el video electrónico y una versión analógica del espacio de color YCbCr. Está formado por tres componentes: **Y** representa la luminancia, **Pb** la diferencia entre la tercera componente del RGB (Blue) y la luminancia, y **Pr** la diferencia entre la primera componente del espacio RGB (Red) y la luminancia **Y**. Las siguientes ecuaciones son utilizadas para realizar la transformación del espacio de color YPbPr partiendo del espacio RGB:

$$Y = 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \tag{3.4}$$

$$Pb = B - Y \tag{3.5}$$

$$Pr = R - Y \tag{3.6}$$

YCbCr El formato YCbCr es muy utilizado en compresión de imágenes, sobre todo en los estándares JPEG y MPEG [23]. El espacio de color está formado por la componente

de luminancia **Y** y dos componentes de crominancia que se corresponden con las componentes R y B del espacio de color RGB. La transformación al espacio de color YCbCr desde el espacio de color RGB se realiza a partir de las siguientes ecuaciones [28]:

$$Y = 16 + 0{,}279 \cdot R + 0{,}504 \cdot G + 0{,}098 \cdot B \tag{3.7}$$
$$Cb = 128 - 0{,}148 \cdot R - 0{,}291 \cdot G + 0{,}439 \cdot B \tag{3.8}$$
$$Cr = 128 + 0{,}439 \cdot R - 0{,}368 \cdot G - 0{,}071 \cdot B \tag{3.9}$$

JPEG-YCbCr El espacio de color JPEG-YCbCr es similar al YCbCr explicado anteriormente, con la salvedad de que el rango de entrada utilizado son 8 bits. Este formato es utilizado para la representación de imágenes JPEG [28]:

$$JPEG - Y = 0 + 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \tag{3.10}$$
$$JPEG - Cb = 128 - 0{,}169 \cdot R - 0{,}331 \cdot G + 0{,}5 \cdot B \tag{3.11}$$
$$JPEG - Cr = 128 + 0{,}5 \cdot R - 0{,}419 \cdot G - 0{,}081 \cdot B \tag{3.12}$$

YDbDr Está formado por tres componentes: la luminancia **Y**, en conjunto con dos componentes de crominancia **Db** y **Dr** [29]:

$$Y = 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \tag{3.13}$$
$$Db = -0{,}450 \cdot R - 0{,}883 \cdot G + 1{,}333 \cdot B \tag{3.14}$$
$$Dr = -1{,}333 \cdot R + 1{,}116 \cdot G + 0{,}217 \cdot B \tag{3.15}$$

YUV El espacio de color YUV es utilizado por los formatos de video PAL (Phase Alernation Line), NTSC (National Television System Committee) y SECAM (Sequentiel Coleur Avec Mémoire). La luminancia **Y** se encarga de la representación de los tonos blancos y negros mientras que la informacion a color corre a cargo de las componentes **U** y **V** [23]:

$$Y = 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \tag{3.16}$$
$$U = 0{,}493 \cdot (B - Y) \tag{3.17}$$
$$V = 0{,}877 \cdot (R - Y) \tag{3.18}$$

YIQ Espacio de color derivado del YUV y utilizado en el formato de video a color NTSC, donde sus tres componentes representan la luminancia (**Y**), la fase (**I**) y la cuadratura (**Q**) [30]:

$$Y = 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \tag{3.19}$$

$$I = 0{,}596 \cdot R - 0{,}274 \cdot G - 0{,}322 \cdot B \tag{3.20}$$

$$Q = 0{,}211 \cdot R - 0{,}522 \cdot G + 0{,}311 \cdot B \tag{3.21}$$

HSV El espacio de color HSV se caracteriza por la utilización de magnitudes fácilmente interpretables por el ser humano como son la tonalidad **H** (Hue) que indica los diferentes colores del espectro, la saturación **S** (Saturation), idicativa de la pureza del color, y el brillo **V** (Value), que indica el grado de claridad u oscuridad del color [31]. Las tres componentes, obtenidas a partir del espacio de color RGB y basadas en coordenadas cilíndricas, se definen a través de las siguientes ecuaciones:

$$H = \begin{cases} arccos\frac{(R-G)+(R-B)}{2\cdot\sqrt{(R-G)^2+(R-B)\cdot(G-B)}}, & B \leq G \\ 2\pi - arccos\frac{(R-G)+(R-B)}{2\cdot\sqrt{(R-G)^2+(R-B)\cdot(G-B)}}, & B > G \end{cases} \tag{3.22}$$

$$S = \frac{max(R,G,B) - min(R,G,B)}{max(R,G,B)} \tag{3.23}$$

$$V = max(R,G,B) \tag{3.24}$$

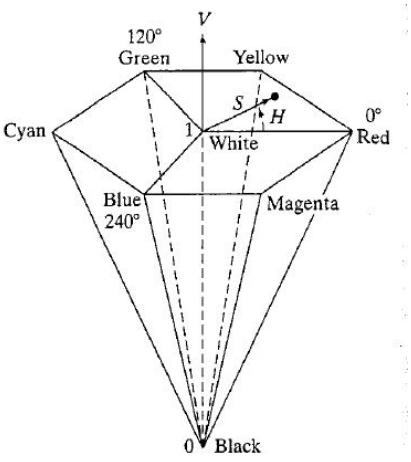


Figura 3.2: Representación del espacio de color HSV [32].

HSL Espacio de color muy similar al HSV cambiando la tercera componente **V** (Brillo) por **L** (Luminosidad) [23].

$$H = \begin{cases} arccos\frac{(R-G)+(R-B)}{2\cdot\sqrt{(R-G)^2+(R-B)\cdot(G-B)}}, & B \leq G \\ 2\pi - arccos\frac{(R-G)+(R-B)}{2\cdot\sqrt{(R-G)^2+(R-B)\cdot(G-B)}}, & B > G \end{cases} \tag{3.25}$$

$$S = \frac{max(R,G,B) - min(R,G,B)}{min(max(R,G,B) + min(R,G,B), 2 - max(R,G,B) - min(R,G,B))} \tag{3.26}$$

$$L = \frac{max(R,G,B) + min(R,G,B)}{2} \tag{3.27}$$

HSI El espacio de color HSI comparte la primera y la segunda componente con el HSV, cambiando la tercera componente **V** (Brillo) por **I** (Intensidad). Este espacio de color es utilizado para funciones de procesado de imágenes como la convolución [32].

$$H = \begin{cases} arccos\frac{(R-G)+(R-B)}{2\cdot\sqrt{(R-G)^2+(R-B)\cdot(G-B)}}, & B \leq G \\ 2\pi - arccos\frac{(R-G)+(R-B)}{2\cdot\sqrt{(R-G)^2+(R-B)\cdot(G-B)}}, & B > G \end{cases} \tag{3.28}$$

$$S = 1 - \frac{3}{(R+G+B)} \cdot [min(R,G,B)] \tag{3.29}$$

$$I = \frac{1}{3} \cdot (R+G+B) \tag{3.30}$$

XYZ Es un estándar internacional desarrollado por la CIE (*Commission Internationale de l'Eclairage*) formado por tres componentes **X**, **Y** y **Z** donde la segunda componente (Y) se encarga de la definición de luminancia mientras que las otras dos componentes proporcionan la información del color [33]:

$$X = 0{,}412453 \cdot R + 0{,}35758 \cdot G + 0{,}180423 \cdot B \tag{3.31}$$

$$Y = 0{,}212671 \cdot R + 0{,}71516 \cdot G + 0{,}072169 \cdot B \tag{3.32}$$

$$Z = 0{,}019334 \cdot R + 0{,}119193 \cdot G + 0{,}950227 \cdot B \tag{3.33}$$

Lab y Luv Los espacios de color CIE-lab y CIE-Luv son ampliamente utilizados debido a que relacionan los valores numéricos de color consistentemente con la percepción visual humana. Fueron modelados en base a la teoría de que dos colores no pueden ser rojo y verde al mismo tiempo, o azul y amarillos al mismo tiempo (colores oponentes), así la componente **L** indica la luminosidad, la segunda componente **a** (**u**) indica las coordenadas rojo/verde (+a indica rojo y -a indica verde), y la componente **b** (**v**) se refiere a las coordenadas amarillo/azul (+b inidica amarillo y -b indica azul) [23].

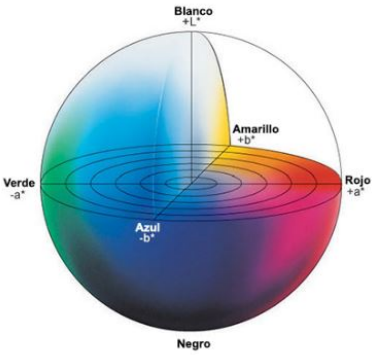


Figura 3.3: Representación del espacio de color Lab [lab].

LCH Similar a los espacios de color CIE-Lab y CIE-Luv con la diferencia de que la segunda y tercera componente (**C** y **H** respectivamente) representan el nivel de saturación y la tonalidad.

CAT02-LMS El espacio de color CAT02-LMS está representado por las tres celdas cónicas del ojo humano (células fotorreceptoras de las retinas), descritas por sus picos de responsividad en longitudes de onda pequeñas, medianas y grandes.

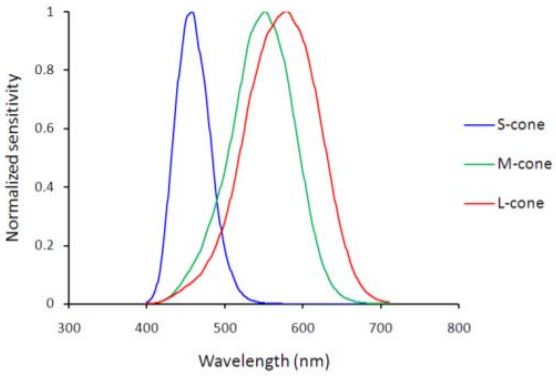


Figura 3.4: Representación del espacio de color CAT02-LMS [34].

Índices de vegetación

Los índices de vegetación son una de las medidas más populares estudiadas en la detección a través de satélites de informaciones sobre la superficie terrestre. Utilizan una combinación de transformaciones espectrales de las diferentes bandas del espectro electromagnético para caracterizar y detectar la actividad de la flora terrestre [35].

A través de la contribución de diferentes autores y artículos publicados [22], [24] se han definido 14 índices de vegetación distintos en función del espacio de color RGB, pero extrapolables al resto de espacios de color propuestos:

Índice de vegetación (VI)	**Expresión**
Normalized first component of RGB color space	$R_n = R/(R+G+B)$
Normalized second component of RGB color space	$G_n = G/(R+G+B)$
Normalized third component of RGB color space	$B_n = B/(R+G+B)$
Gray channel	$gray = 0{,}2898 \cdot R_n + 0{,}5870 \cdot G_n + 0{,}1140 \cdot B_n$
Additional green [36]	$EXG = 2 \cdot G_n - R_n - B_n$
Additional red [37]	$EXG = 1{,}4 \cdot R_n - G_n$
Color index for extracted vegetation cover [38]	$CIVE = 0{,}441 \cdot R_n - 0{,}811 \cdot G_n + 0{,}385 \cdot B_n + 18{,}78$
Subtraction between additional green and additional red [39]	$EXGR = EXG - EXR$
Normalized difference index [40]	$NDI = (G_n - B_n)/(G_n + B_n)$
Green index minus blue [36]	$GB = (G_n - B_n)$
Red-blue contrast [41]	$RBI = (R_n - B_n)/(R_n + B_n)$
Green-Red index [41]	$ERI = (R_n - G_n) \cdot (R_n - B_n)$
Additional green index [41]	$EGI = (G_n - R_n) \cdot (G_n - B_n)$
Additional blue index [41]	$EBI = (B_n - G_n) \cdot (B_n - R_n)$

Tabla 3.2: Índices de vegetación para el espacio de color RGB con sus expresiones matemáticas definitorias.

3.1.2. Textura

La textura se define como la sensación que produce al tacto el roce con una determinada materia. En función de los materiales y de la sensación experimentada el objeto se puede clasificar como liso o rugoso entre otras propiedades.

Para definir y clasificar la textura en imágenes se utiliza la matriz de co-ocurrencias de nivel de gris o GLCM (Gray Level Co-occurrence Matrix), que es una matriz cuadrada donde el número de filas y columnas es equivalente al número de niveles de gris de la imagen y que representa la distribución de las intensidades e información sobre píxeles vecinos en una imagen determinada. Dada una imagen **I**, de tamaño $N \times X$, la matriz GLCM se representa como [42]:

$$P(i,j) = \sum_{x=1}^{N} \sum_{y=1}^{N} \begin{cases} 1, & \text{if } I(x,y) = i \quad \&\& \quad I(x+\Delta_x, y+\Delta_y) = j \\ 0, & resto \end{cases} \tag{3.34}$$

donde Δ_x y Δ_y representan el offset, la dirección espacial en la cual la matriz de co-ocurrencias es calculada [43]. Las direcciones más comunes son cuatro: 0, 45, 90 y 135 grados (Figura 3.5), representadas como [0 1], [-1 1], [-1 0], [-1 1] respectivamente.

A partir de la matriz GLCM se han definido un conjunto de características que ayudan a caracterizar la textura de una imagen. En concreto, se han descrito 20 características distintas extraidas para cada imagen de la base de datos propuesta que se explicarán a continuación [42], [44], [45] teniendo en cuenta las siguientes consideraciones:

Se define $p(i,j)$ como la entrada (i,j)-ésima de la matriz GLMC normalizada:

$$p(i,j) = \frac{P(i,j)}{\sum_{i,j} P(i,j)} \tag{3.35}$$

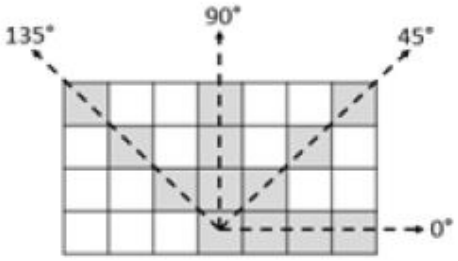


Figura 3.5: Representación del valor del offset en la matriz de co-ocurrencias [44].

donde $P(i.j)$ representa la entrada (i, j)-ésima de la matriz de co-ocurrencia GLCM.

Para cada conjunto de filas y columnas se definen dos probabilidades marginales p_x y p_y como:

$$p_x(i) = \sum_{i=1}^{N_g} p(i,j) \tag{3.36}$$

$$p_y(j) = \sum_{j=1}^{N_g} p(i,j) \tag{3.37}$$

A partir ellas se obtienen la entropías HX y HY; y HXY, $HXY1$ y $HXY2$ se definen como:

$$HXY = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \cdot \log[p(i,j)] \tag{3.38}$$

$$HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \cdot \log[p_x(i) \cdot p_y(j)] \tag{3.39}$$

$$HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i) \cdot p_y(j) \cdot \log[p_x(i) \cdot p_y(j)] \tag{3.40}$$

Por otro lado, se definen μ_x, μ_y y σ_x, σ_y como la media y la desviación típica de las filas y columnas de la matriz GLCM respectivamente. El valor N_g representa el número total de niveles de gris de la imagen.

Por último, p_{x+y} y p_{x-y} se definen como la suma y la diferencia entre los píxeles de co-ocurrencia de la matriz GLCM:

$$p_{x+y} = \sum_{\substack{i=1 \\ i+j=k}}^{N_g}\sum_{j=1}^{N_g} p(i,j), \quad k = 2, 3, ..., 2N_g \tag{3.41}$$

$$p_{x-y} = \sum_{\substack{i=1 \\ |i-j|=k}}^{N_g} \sum_{j=1}^{N_g} p(i,j), \quad k = 0, 1, 2, ..., N_g - 1 \tag{3.42}$$

Probabilidad máxima (Maximum probability) Valor máximo de $p(i,j)$ para todo el espacio de i filas y j columnas que forman la matriz de co-ocurrencia GLCM.

$$\text{Maximum probability} = \max(p(i,j)) \tag{3.43}$$

Suma de cuadrados (Sum of squares) Dispersión de la distribución de nivel de gris de una imagen.

$$\text{Sum of squares} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2 \cdot p(i,j) \tag{3.44}$$

Media suma (Sum average) Medida indicativa de la media de la suma de píxeles co-ocurrentes de nivel de gris de la imagen.

$$\text{Sum average} = \sum_{i=2}^{2N_g} i \cdot p_{x+y}(i) \tag{3.45}$$

Varianza suma (Sum variance) Dispersión de la suma de píxeles co-ocurrentes de nivel de gris de la imagen.

$$\text{Sum variance} = \sum_{i=2}^{2N_g} \left(i - \left[\sum_{i=2}^{2N_g} i \cdot p_{x+y}(i) \right] \right)^2 \tag{3.46}$$

Entropía suma (Sum entropy) Medida del desorden referido a la suma de píxeles co-ocurrentes de nivel de gris de la imagen.

$$\text{Sum entropy} = \sum_{i=2}^{2N_g} p_{x+y}(i) \cdot \log[p_{x+y}(i)] \tag{3.47}$$

Varianza diferencia (Difference variance) Medida de la dispersión de la diferencia entre píxeles co-ocurrentes de nivel de gris de la imagen.

$$\text{Difference variance} = \sum_{i=2}^{2N_g} \left(i - \left[\sum_{i=2}^{2N_g} i \cdot p_{x-y}(i) \right] \right)^2 \tag{3.48}$$

Entropía diferencia (Difference entropy) Medida del desorden de la diferencia entre píxeles co-ocurrentes de nivel de gris de la imagen.

$$\text{Difference entropy} = -\sum_{i=2}^{2N_g} p_{x-y}(i) \cdot \log[p_{x-y}(i)] \tag{3.49}$$

Medida de información de la correlación1 (Information measure of correlation1)

$$\text{Information measure of correlation1} = \frac{HXY - HXY1}{\max(\text{HX,HY})} \tag{3.50}$$

Medida de información de la correlación2 (Information measure of correlation2)

$$\text{Information measure of correlation2} = \sqrt{1 - \exp[-2 \cdot (HXY2 - HXY)]} \tag{3.51}$$

Diferencia inversa en homogeneidad (Inverse difference in homogeneity) Medida de la homogeneidad de la distribución del nivel de gris en la imagen. Es inversamente proporcional al contraste: si el contraste tiene un valor reducido, la homogeneidad es elevada.

$$\text{Inverse difference in homogeneity} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p(i,j)}{1 + (i-j)^2} \tag{3.52}$$

Diferencia inversa normalizada (Inverse difference normalized) Medida de la homogeneidad local del nivel de gris en la matriz de co-ocurrencias.

$$\text{Inverse difference normalized} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p(i,j)}{1 + \frac{|i-j|}{N_g}} \tag{3.53}$$

Diferencia de momentos inversa normalizada (Inverse difference moment normalized) Medida equivalente a la diferencia inversa en homogeneidad, pero con la consideración de la normalización del valor de $p(i,j)$.

$$\text{Inverse difference moment normalized} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p(i,j)}{1 + \left[\frac{(i-j)}{N_g}\right]^2} \tag{3.54}$$

Momento diagonal (Diagonal moment)

$$\text{Diagonal moment} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \sqrt{|i-j| \cdot \frac{1}{2} \cdot p(i,j)} \tag{3.55}$$

Segundo momento diagonal (Second diagonal moment)

$$\text{Second diagonal moment} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} |i-j| \cdot \frac{1}{2} \cdot p(i,j) \tag{3.56}$$

Varianza (Variance) Dispersión en términos de valores de filas y columnas de la matriz de co-ocurrencia de nivel de gris.

$$\sigma_x^2 = \sum_{i=1}^{N_g} p_x(i) \cdot (i-\mu_x)^2 \qquad \sigma_y^2 = \sum_{i=1}^{N_g} p_y(i) \cdot (i-\mu_y)^2 \tag{3.57}$$

Media (Mean) Valor característico del conjunto de probabilidades de filas y columnas de la matriz GLCM.

$$\mu_x = \sum_{i=1}^{N_g} i \cdot p_x(i) \qquad \mu_y = \sum_{i=1}^{N_g} i \cdot p_y(i) \tag{3.58}$$

Desviación estándar (Standard deviation) Se define como la raíz cuadrada de la varianza.

Coeficiente de variación (Coefficient of variation) Relacion entre la media y la desviación típica de la matriz de co-ocurrencias GLCM.

$$\text{Coefficient of variation} = \frac{\sigma}{\mu} \tag{3.59}$$

Contraste (Contrast) Representa la variación del nivel de gris en una imagen determinada. Un alto valor de contraste indicará la presencia de bordes o texturas arrugadas en la imagen.

$$\text{Contrast} = \sum_{n=0^{N_g-1}} (i-j)^2 \left\{ \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \right\} \tag{3.60}$$

Correlación (Correlation) Medida de la dependencia lineal de los niveles de gris con cada uno de los píxeles de co-ocurrencia vecinos.

$$\text{Correlation} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{i \cdot j \cdot p(i,j) - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y} \tag{3.61}$$

Caracteristica #	Nombre característica	Caracteristica #	Nombre característica
1	Maximum probability	11	Inverse difference normalized (INN)
2	Sum of squeres	12	Inverse difference moment normalized
3	Sum average	13	Diagonal moment
4	Sum variance	14	Second diagonal moment
5	Sum entropy	15	Variance
6	Difference variance	16	Mean
7	Difference entropy	17	Standard deviation
8	Information measure of correlation1	18	Coefficient of deviation
9	Information measure of correlation2	19	Contrast
10	Inverse difference in homogeneity	20	Correlation

Tabla 3.3: Características referentes a la textura extraídas de la matriz de co-ocurrencias de nivel de gris.

3.1.3. Momentos invariantes

El momento bidimensional de orden $(p+q)$ para una imagen digital $f(x,y)$ se define como:

$$m_{pq} = \sum_x \sum_y x^p \cdot y^q f(x,y), \quad p,q = 0,1,2,... \tag{3.62}$$

donde x e y representan el conjunto de las coordenadas espaciales de la imagen.

Por otro lado, el momento central correspondiente a la imagen $f(x,y)$ es el siguiente:

$$\mu_{pq} = \sum_x \sum_y (x-\overline{x})^p \cdot (y-\overline{y})^q f(x,y) \tag{3.63}$$

donde

$$\overline{x} = \frac{m_{10}}{m_{00}}, \qquad \overline{y} = \frac{m_{01}}{m_{00}} \tag{3.64}$$

A partir del momento central se obtiene el momento central normalizado de orden $(p+q)$:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{3.65}$$

para $p,q = 0,1,2,...$ y con $\gamma = \frac{p+q}{2} + 1$.

Una vez definidos estos conceptos se pueden obtener un conjunto de siete momentos invariantes bidimensionales que no sufren modificaciones frente al cambio de escala, la traslación y la rotación [32]:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{3.66}$$

$$\phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \tag{3.67}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{3.68}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{3.69}$$

$$\begin{aligned}\phi_5 = {} & [(\eta_{30} - 3\eta_{12}) \cdot (\eta_{30} + \eta_{12}) \cdot (\eta_{30} - \eta_{12})^2 \\ & - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03}) \cdot (\eta_{21} + \eta_{03}) \cdot \\ & [3 \cdot (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{3.70}$$

$$\begin{aligned}\phi_6 = {} & [(\eta_{20} - \eta_{02}) \cdot [(\eta_{30} + \eta_{12})^2 - (\eta_{21} - \eta_{03})^2] \\ & + 4\eta_{11} \cdot (\eta_{30} + \eta_{12}) \cdot (\eta_{21} - \eta_{03})\end{aligned} \tag{3.71}$$

$$\begin{aligned}\phi_7 = {} & [(3\eta_{21} - \eta_{03}) \cdot (\eta_{30} + \eta_{12}) \cdot (\eta_{30} - \eta_{12})^2 \\ & - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{30}) \cdot (\eta_{21} + \eta_{03}) \cdot \\ & [3 \cdot (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{3.72}$$

Caracteristica #	**Nombre característica**
1	ϕ_1
2	ϕ_2
3	ϕ_3
4	ϕ_4
5	ϕ_5
6	ϕ_6
7	ϕ_7
8	$\phi_1 - \phi_7$
9	$\phi_2 - \phi_6$
10	$\phi_3 - \phi_5$

Tabla 3.4: Características referentes a los momentos invariantes extraídas de la imagen.

3.1.4. Histograma

El conjunto de características de intensidad del histograma se utilizan para describir una región concreta en términos del estudio de las texturas que contiene [32].

Las características referentes a la intensidad del histograma son seis que se explican a continuación:

Media (Mean) Medida del valor característico de la intensidad.

$$m = \sum_{i=0}^{L-1} z_i \cdot p(z_i) \tag{3.73}$$

donde z_i es una variable aleatoria indicativa de la intensidad, $p(z)$ el histograma de los niveles de intensidad en una región determinada y L el número de niveles de intensidad posibles.

Caracteristica #	Nombre característica
1	Mean
2	Standard deviation
3	Smoothness
4	Third moment
5	Uniformity
6	Entropy

Tabla 3.5: Características referentes a la textura en términos de la intensidad del histograma extraídas de la imagen.

Desviación estándar (Standard deviation) Medida del contraste medio de la imagen.

$$\sigma = \sqrt{\mu_2(z)} \tag{3.74}$$

donde $\mu_2 = \sum_{i=0}^{L-1}(z_i - m)^2 \cdot p(z_i)$ representa el segundo momento.

Homogeneidad (Smoothness) Caracterización de la suavidad relativa de la intensidad de una región concreta. Su valor es cero para aquellas regiones con intensidad constante y uno en caso de que se disponga de diferentes valores del nivel de intensidad.

$$R = 1 - \frac{1}{(1 + \sigma^2)} \tag{3.75}$$

Tercer momento (Third moment) Medida indicativa de la asimetría del histograma. Su valor es cero para histogramas simétricos, positivo si la asimetría tiende hacia el lado derecho y negativo si tiende hacia el lado izquierdo.

$$\mu_3 = \sum_{i=0}^{L-1}(z_i - m)^3 \cdot p(z_i) \tag{3.76}$$

Uniformidad (Uniformity) Medida del grado de uniformidad del histograma. Su valor es máximo cuando todos los niveles de grises son iguales (Uniformidad máxima).

$$U = \sum_{i=0}^{L-1} p^2(z_i) \tag{3.77}$$

Entropía (Entropy) Medida del grado de desorden del histograma.

$$e = -\sum_{i=0}^{L-1} p(z_i) \cdot \log_2(p(z_i)) \tag{3.78}$$

3.1.5. Forma

La forma se define como la descripción geométrica de la parte del espacio ocupado por un determinado objeto y que viene dada por el contorno del mismo [32].

La caracteristicas referentes a la forma son aquellas medidas físicas que permiten caracterizar la apariencia de un objeto contenido en una imagen, y su principal objetivo la descripción del mismo con el menor número de características posible [46]. En el presente proyecto se estudiarán 20 características referentes a la forma y extraídas a partir de cada una de las imágenes segmentadas en formato binario que conforman la base de datos propuesta.

Caracteristica #	**Nombre característica**	**Caracteristica #**	**Nombre característica**
1	Length	11	ConvexArea
2	Width	12	FilledArea
3	Area	13	EquivDiameter
4	Log. of the ratio length to width	14	EulerNumber
5	Ratio of perimeter to broadness	15	Solidity
6	Ratio of width to length	16	Elongation
7	Ratio Area to length	17	Compactness
8	Eccentricity	18	Extent
9	Perimeter	19	Aspect Ratio
10	Orientation	20	Ratio of length to perimeter

Tabla 3.6: Características referentes a la forma extraídas de la imagen.

Longitud (Length) Medida que equivale a la longitud (en píxeles) del eje mayor de la elipse que tiene el segundo momento igual al área de la región analizada.

Anchura (Width) Medida equivalente a la longitud (en píxeles) del eje menor de la elipse que tiene su segundo momento igual al área de la región analizada.

Área (Area) Número de píxeles que conforman una región específica.

Logaritmo de la relación longitud/anchura (Log. of the ratio of the length to width) Logaritmo de la relación entre la longitud y la anchura del objeto analizado.

$$\text{Log. of the ratio of the length to width} = \log\left(\frac{\text{Length}}{\text{Width}}\right) \tag{3.79}$$

Relación perímetro/amplitud (Ratio of perimeter to broadness) Relación entre el perímetro de la región analizada y su amplitud.

$$\text{Ratio of perimeter to broadness} = \frac{\text{Perimeter}}{2 \cdot (\text{Length} + \text{Width})} \tag{3.80}$$

Relación anchura/longitud (Ratio of width to length) Relación entre la anchura y la longitud del objeto estudiado.

$$\text{Ratio width to length} = \frac{\text{Width}}{\text{Length}} \tag{3.81}$$

Relación área/longitud (Ratio Area to length) Relación entre el área del objeto analizado y su longitud.

$$\text{Ratio area to length} = \frac{\text{Area}}{\text{Length}} \tag{3.82}$$

Excentricidad (Eccentricity) Se define como la relación entre el foco de la elipse y su eje mayor. Su valor se sitúa entre 0 y 1.

Perímetro (Perimeter) Número de píxeles que forman el borde de la región analizada.

Orientación (Orientation) Ángulo (en grados) formado entre el eje mayor de la elipse y el eje de abscisas (eje x).

Área convexa (ConvexArea) Es el área que rodea e incluye al objeto analizado.

Área rellena (FilledArea) Número de píxeles que forman el cuadro delimitador (Bounding Box, región rectangular del mínimo tamaño que pueda contener la región a analizar) en formato binario.

Diámetro equivalente (EquivDiameter) Valor del diámetro del círculo con el mismo área que la región estudiada.

Número de Euler (EulerNumber) Valor escalar igual al número de objetos contenidos en una determinada región menos el número de huecos de esos objetos.

Solidez (Solidity) Mide la proporción de píxeles que existen en el área que rodea al objeto y que además pertenecen al objeto.

$$\text{Solidity} = \frac{\text{Area}}{\text{ConvexArea}} \tag{3.83}$$

Elongación (Elongation) Relación entre la diferencia de la longitud y la anchura del objeto, y la suma de ambas magnitudes.

$$\text{Elongation} = \frac{\text{Length} - \text{Width}}{\text{Length} + \text{Width}} \tag{3.84}$$

Compactibilidad (Compactness) Relación entre el área del objeto con el área del círculo con el mismo perímetro.

$$\text{Compactness} = \frac{\text{Perimeter}^2}{4\pi \cdot \text{Area}} \tag{3.85}$$

Alcance (Extent) Proporción de píxeles del cuadro delimitador que también pertenecen al objeto analizado.

$$\text{Extent} = \frac{\text{Area}}{\text{Bounding-Box Area}} \tag{3.86}$$

Proporción de aspecto (Aspect Ratio) Relación entre la longitud y la anchura del objeto estudiado.

$$\text{Aspect Ratio} = \frac{\text{Length}}{\text{Width}} \tag{3.87}$$

Relación longitud/perímetro (Ratio of length to perimeter) Relación entre la longitud de la región analizada y su perímetro.

$$\text{Ratio of length to perimeter} = \frac{\text{Length}}{\text{Perimeter}} \tag{3.88}$$

3.2. Reducción de la dimensionalidad: Análisis de componentes principales

El análisis de componentes principales (PCA) es un método utilizado para conseguir la reducción de la dimensionalidad de grandes bancos de datos a partir de la transformación del mismo en otro más reducido que contenga la misma cantidad de información. El principal objetivo consiste en extraer la información más relevante del banco de datos original y definirla como un nuevo conjunto de variables ortogonales llamadas componentes principales [47].

La primera componente principal obtenida es la que posee mayor cantidad de información y por lo tanto mayor varianza (Figura 3.6). A partir de aquí, el resto de

componentes principales contendra menor cantidad de información y menor valor de varianza según una ordenación descendente.

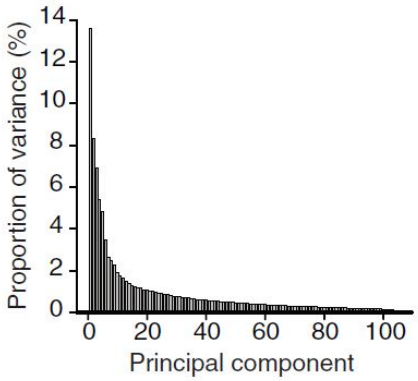


Figura 3.6: Porcentaje total de la varianza para las diferentes componentes principales de un banco de datos [48].

3.2.1. Resolución de PCA utilizando descomposición en autovectores

La primera solución para la obtención de las componentes principales de un banco de datos se basa en la descomposición en autovectores. Un autovector es aquel vector que cuando sufre una transformación lineal determinada no cambia su dirección. Relacionado con el concepto de autovector está la definición del autovalor, que representa la constante que reduce o incrementa el autovalor a lo largo de su dirección en una transformación lineal.

La matriz que representa el banco de datos se define como X, de tamaño $m \times n$, donde m representa el número de características medidas y n el número de muestras disponibles, sabiendo que:

$$N \cdot V = \lambda \cdot V \tag{3.89}$$

donde λ representa el conjunto de autovalores y V los autovectores del espacio obtenidos a partir de la matriz de varianzas-covarianzas N definida como [8]:

$$N = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,p} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{q,1} & C_{q,2} & \cdots & C_{q,p} \end{bmatrix} \tag{3.90}$$

donde $C_{q,p}$ es la covarianza de la matriz X:

$$Cov(X_q, X_p) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_{q,i} - \overline{X_q}) \cdot (X_{p,i} - \overline{X_p}) \tag{3.91}$$

Por lo tanto, la obtención de las componentes principales de la matriz X se realiza mediante el desarrollo de las siguientes etapas:

1. Banco de datos X de tamaño $m \times n$, con m el número de características disponibles y n como el número de muestras.

2. Estandarización de los datos de entrada X para conseguir una contribución equitativa entre el conjunto total de valores:

$$X_{standardize} = \frac{X - \bar{X}}{std(X)} \tag{3.92}$$

3. Cálculo de la matriz de varianzas-covarianzas N.

4. Cálculo de los autovalores λ y autovectores V a partir de la matriz de varianzas-covarianzas.

5. Ordenación descendente de los autovalores λ y autovectores V.

6. Selección de los n primeros autovectores (componentes principales) para reducir la dimensionalidad del banco de datos original X.

7. Reconstrucción del nuevo espacio de dimensionalidad reducida a partir de las componentes principales seleccionadas:

$$X_{PCA} = V^T \cdot X_{standardize} \tag{3.93}$$

3.2.2. Resolución de PCA utilizando descomposición en valores singulares SVD

La descomposición en valores singulares representa la factorización de una matriz X de dimensiones $m \times n$, donde m representa el número de características y n el numero de muestras, a partir de la siguiente expresión:

$$X_{m \times n} = U_{m \times m} \cdot \Sigma_{m \times n} \cdot V^T_{n \times n} \tag{3.94}$$

donde U representa el conjunto de vectores singulares izquierdos obtenidos a partir del producto $A \cdot A^T$. Los autovalores obtenidos de este producto forman las columnas de la matriz U, cumpliéndose además que $U^T \cdot U = I_{n \times n}$, con I la matriz identidad.

Por otro lado, la matriz V representa el conjunto de vectores singulares derechos obtenidos a partir de $A^T \cdot A$. Se cumple la relación $V^T \cdot V = I_{m \times m}$, con I definida como la matriz identidad.

La matriz Σ, del mismo tamaño que la matriz original X, se caracteriza por su forma diagonal y contiene el conjunto de valores singulares ordenados de forma descendente ($\sigma_1 \geq \sigma_1 \geq \cdots \geq \sigma_r$) y obtenidos mediante la raíz cuadrada de los autovalores de $A \cdot A^T$ y $A^T \cdot A$ [49].

$$\underbrace{\begin{pmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{pmatrix}}_{\mathbf{X_{m \times n}}} = \underbrace{\begin{pmatrix} u_{1,1} & \dots & u_{m,1} \\ \vdots & \ddots & \vdots \\ u_{1,m} & \dots & u_{m,m} \end{pmatrix}}_{\mathbf{U_{m \times m}}} \cdot \underbrace{\begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \sigma_r & \vdots \\ 0 & \dots & 0 \end{pmatrix}}_{\mathbf{\Sigma_{m \times n}}} \cdot \underbrace{\begin{pmatrix} v_{1,1} & \dots & v_{1,n} \\ \vdots & \ddots & \vdots \\ u_{n,1} & \dots & v_{n,n} \end{pmatrix}}_{\mathbf{V^T_{m \times m}}} \tag{3.95}$$

Así el objetivo principal del calculo SVD consiste en encontrar el conjunto de autovalores y autovectores de $A \cdot A^T$ y $A^T \cdot A$.

Las etapas realizadas para obtener las componentes principales de X serán las siguientes:

1. Banco de datos X de tamaño $m \times n$, con m el número de características disponibles y n como el número de muestras.

2. Estandarización de los datos de entrada X para conseguir una contribución equitativa entre el conjunto total de valores:

$$X_{standardize} = \frac{X - \bar{X}}{std(X)} \tag{3.96}$$

3. Cálculo de la matriz Y [49]:

$$Y = \frac{1}{\sqrt{n}} \cdot X^T \tag{3.97}$$

4. Cálculo de los vectores singulares U y V, y de los valores singulares Σ ordenados de forma descendente a partir de la matriz Y.

5. Selección de los n primeros vectores singulares de V (componentes principales) a fin de reducir la dimensionalidad del banco de datos original X.

6. Reconstrucción del nuevo espacio de dimensionalidad reducida a partir de las componentes principales seleccionadas:

$$X_{PCA} = V^T \cdot X_{standardize} \tag{3.98}$$

3.3. Métodos de regresión

A la hora de estimar el grado de acidez de las naranjas que conforman la base de datos propuesta se hará uso de un cojunto de tecnicas de aprendizaje automático no intrusivas que haran uso de la regresión lineal simple como método de estimación.

Los métodos de regresión propuestos para el presente estudio serán tres: Máquinas de vectores Soporte (*Support Vector Machine*, SVM) , Perceptrón Multicapa (*MultiLayer Perceptron*, MLP) y una red neuronal híbrida basada en el algortimo ICA (*Imperialist Competitive Algorithm*).

3.3.1. Máquinas de vectores soporte

Las Máquinas de Vectores Soporte (*Support Vector Machine*, SVM) son unos sistemas de aprendizaje que pertenecen a la rama de clasificadores lineales generalizados, aplicables tanto en problemas de clasificación como en regresión. Se basan en una red estática de núcleos que operan sobre vectores de características que han sido transformados a un espacio de dimensión mayor a la del espacio de características original [50].

A la hora de la implementación de una regresión lineal de los datos de entrada es necesario tener en cuenta las siguentes características:

Se define la función del hiperplano de regresión lineal como:

$$f(x, w) = w^T \cdot x + b \tag{3.99}$$

siendo la entrada x un conjunto de vectores n-dimensionales $x \in \Re^n$, y la salida f un conjunto de valores continuos $f \in \Re$.

A diferencia de un problema de clasificación, donde la medida utilizada para evaluar el rendimiento y el correcto funcionamiento del clasificador es la maximización del margen, ahora interesa también la medida del error de aproximación definido a partir de diferentes funciones de error lineales. Dos de las funciones de error más comunes son la función cuadrática ($(y - f)^2$) y el error absoluto ($|y - f|$), pero la utilizada para realizar la regresión SVM se conoce como función de pérdidas de Vapnik [51]:

$$e(x, y, f) = max(0, |y - f(x, w)| - \varepsilon) \tag{3.100}$$

donde las pérdidas serán igual a cero en el caso de la diferencia entre en valor medido y con el valor estimado $f(x, w)$ sea menor que el umbral propuesto ε.

A la hora de estudiar la regresión debemos tener en cuenta dos variables diferenciadas. Por un lado, la variable w que indicará el margen maximo disponible del regresor SVM, y por otro lado, el error de penalización R^{ε}_{emp} definido por la función de Vapnik [52]:

$$R_{emp}^{\varepsilon} = C \cdot \sum_{i=1}^{l} |y_i - f(x_i, w)|_{\varepsilon} \tag{3.101}$$

donde $C = \frac{1}{L}$, y_i representa el valor medido y $f(x_i, w)$ el estimado.

A partir de aquí se podrá construir el hiperplano de regresión lineal $f(x, w)$ minimizando la función R:

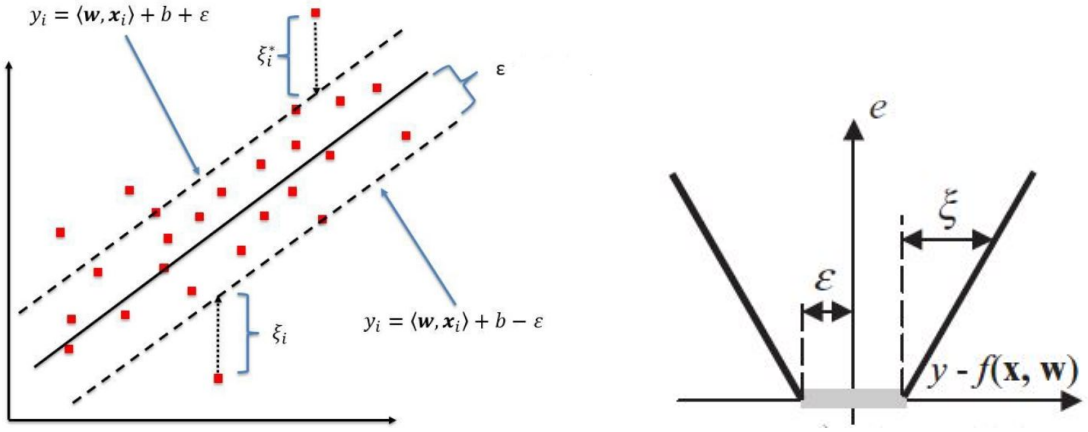
$$\text{Minimizar} \quad R = \frac{1}{2} \cdot ||w||^2 + C \cdot \sum_{i=1}^{l} |y_i - f(x_i, w)|_{\varepsilon} \tag{3.102}$$

y teniendo en cuenta las siguientes restricciones:

$$|y - f(x, w)| - \varepsilon \geq \xi^* \tag{3.103}$$

$$|y - f(x, w)| - \varepsilon \leq \xi \tag{3.104}$$

con $\xi, \xi^* \geq 0$, las variables de holgura, indicativas de la eficiencia de la regresión lineal realizada.



(a) (b)

Figura 3.7: a) Ejemplo de regresión lineal SVM [53] y b) Función de error de Vapnik [51].

3.3.2. Perceptrón multicapa

El perceptrón multicapa (*MultiLayer Perceptron*, MLP) es una red neuronal estática de aprendizaje supervisado formado por varias capas de perceptrones denominados neuronas o nodos: una capa de entrada, una serie de capas ocultas y una capa de salida. Estas capas poseen una conectividad total, esto quiere decir que la entrada a una neurona está conectada con la salida de todas las neuronas de la capa anterior [54].

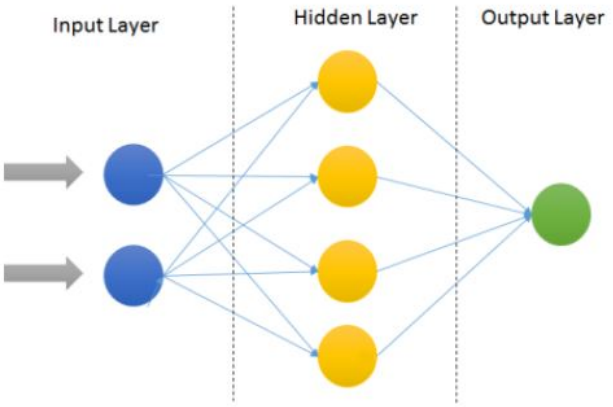


Figura 3.8: Ejemplo de una red MLP con una capa oculta de cuatro neuronas [55].

El algoritmo de aprendizaje más característico de MLP se denomina *backpropagation* o propagación hacia atrás [54], que consiste en la presentación de patrones, donde un patrón de entrada se corresponde con un patron de salida deseada. Tiene dos fases diferenciadas:

- Fase de propagación hacia adelante: se inicia cuando se presenta un patrón en la capa de entrada de la red.
- Fase de propagación hacia atrás o corrección: se inicia una vez finalizada la primera etapa. Se modifican los valores de los pesos, desde la capa de salida hacia la capa de entrada, a fin de reducir el error entre la salida deseada para una determinada entrada y la respuesta que ofreció la red a dicha entrada.

3.3.3. Red neuronal híbrida ANN-ICA

El algoritmo competitivo imperialista (*Imperialist Competitive Algorithm*, ICA) es un algoritmo de optimización basado en la diferenciación de dos poblaciones: imperios y colonias [56].

Este algoritmo se utiliza junto a las redes neuronales MLP para optimizar el entrenamiento de la red y mejorar los resultados obtenidos tanto en clasificación como en regresión.

En la figura 3.11 se explica el funcionamiento y las etapas seguidas del algoritmo ICA. Como la mayoría de los algoritmos de optimización se comienza con una población inicial. Las poblaciones más poderosas serán consideradas como imperios y el resto como colonias, donde estas serán repartidas entre los diferentes imperios en función del poder que tenga cada uno de ellos, siendo el poder inversamente proporcional al coste.

Una vez repartidas las colonias entre los imperios disponibles, estas tendrán la capacidad de desplazarse hacia su correspondiente imperio, acercándose o alejándose en

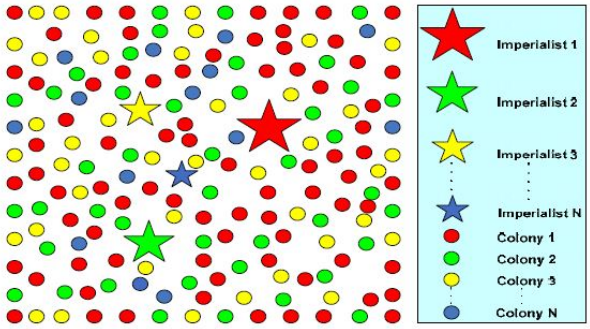


Figura 3.9: Generación de imperios y colonias: a mayor poder de un imperio, mayor tamaño de la estrella que lo define [56].

función del coste. Puede darse el caso de que una colonia posea menor coste que el imperio que la contiene; en este caso se producirá un intercambio de posiciones: la colonia pasará a ser el imperio y viceversa.

Por otro lado, los imperios tienen la capacidad de conquistar colonias en función del poder de los mismos (Figura 3.10). El poder total de un imperio se define como la suma del poder del imperio más un porcentaje de la media del poder de todas sus colonias. Las colonias que tienen más probabilidades de cambiar de imperio serán aquellas que posean menor poder (o mayor coste) y cambiarán su posición a aquel imperio con más probabilidad de poseerla (el de mayor poder). Si un imperio es muy débil o si se queda sin colonias, este se eliminará y sus colonias serán repartidas entre el resto de imperios, y el algoritmo continuará hasta que sólo se mantenga un único imperio y se alcance la convergencia.

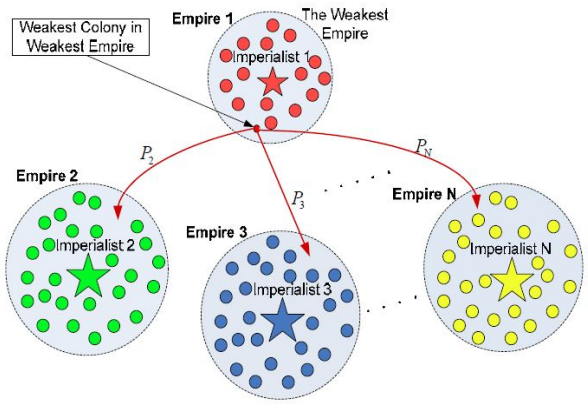


Figura 3.10: Conquista de colonias [56].

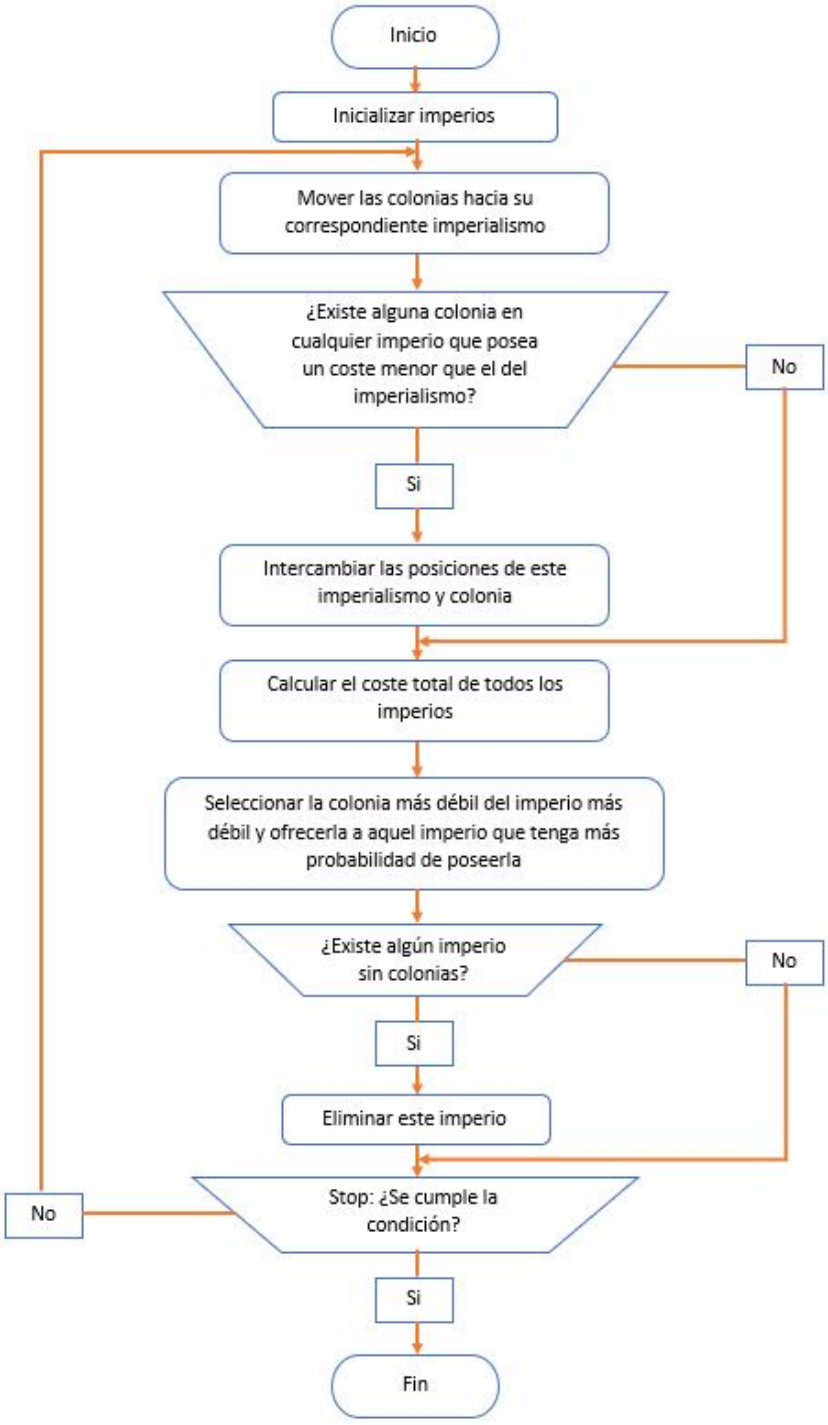


Figura 3.11: Diagrama de flujo del algoritmo ICA adaptado de [56].

3.4. Implementación en el software MATLAB

Una vez definido el conjunto de características más representativas de cada uno de los grupos propuestos se procederá a la extracción de características de cada imagen que conforman la base de datos disponible, para posteriormente reducir la dimensionalidad del mismo a partir de la implementación del algoritmo PCA mediante la descomposición en autovectores con el objetivo de mejorar el rendimiento y disminuir la carga computacional de los mecanismos de aprendizaje automático propuestos en el presente estudio.

Las diferentes etapas a seguir en el proyecto para la obtención de los resultados deseados serán las siguientes (Figura 3.12): análisis de la base de datos y extracción de características (pre-reducción), implementación del algoritmo PCA (post-reducción), implementación de los mecanismos de regresión lineal y obtención de resultados

3.4.1. Etapa pre-reducción

A partir de las características definidas en la sección 3.1 cada imagen de la base de datos se representará a través del siguiente conjunto de características:

- Un total de 336 características referentes al color: por un lado 7 características significativas (Tabla 3.1) para cada uno de los 16 espacios de color disponibles (*7 características/espacio × 16 espacios de color = 112 características*) y, por otro lado, 14 características referentes a los índices de vegetación (Tabla 3.2) para cada uno de los 16 espacios de color disponibles (*14 características/espacio × 16 espacios de color = 224 características*).

- Un total de 80 características referentes a la textura obtenidas a partir de las 20 características significativas definidas (Tabla 3.3) para cuatro niveles de offset diferentes: 0, 45, 90 y 135 grados (*20 características/offset × 4 offset = 80 características*).

- Un total de 10 características referentes a los momentos invariantes (Tabla 3.4).

- Un total de 6 características referentes al histograma (Tabla 3.5).

- Un total de 20 características referentes a la forma (Tabla 3.6).

Por lo tanto la base de datos de características del proyecto estará formada por 452 características extraídas para cada imagen, generando las siguientes matrices:

- Una matriz de 452 × 100 representativa del conjunto de naranjas de la clase *Bam* (100 muestras totales).

- Una matriz de 452 × 100 representativa del conjunto de naranjas de la clase *Blood* (100 muestras totales).
- Una matriz de 452 × 100 representativa del conjunto de naranjas de la clase *Thomson* (100 muestras totales).
- Una matriz de 452 × 300 representativa de la combinación de las tres clases de naranjas disponibles *Bam*, *Blood* y *Thomson* (300 muestras totales, 100 por variedad).

3.4.2. Etapa post-reducción

La etapa de reducción de la dimensionalidad es un proceso muy importante a la hora mejorar el rendimiento y reducir la carga computacional de los mecanismos de regresión, perdiendo la menor cantidad de información posible. Como se ha explicado en la sección 3.2, la primera componente principal es la que posee mayor cantidad de información al tener la mayor varianza. Las siguentes componentes principales irán reduciendo su valor de la varianza en orden descendente, por lo que la selección de las primeras componentes principales es suficiente para recabar la mayor cantidad de información disponible. En el presente proyecto se han decidido seleccionar las cuatro primeras componentes principales debido a que la cantidad de información contenida en ellas es muy alta y las pérdidas reducidas, además el reducido tamaño de las matrices obtenidas permitirá reducir la carga computacional y el tiempo de ejecución de los mecanismos de regresión propuestos; y, por último, los resultados obtenidos podrán ser comparables a los obtenidos en [8] al utilizar las mismas condiciones en la reducción.

Una vez ejecutado el algortimo PCA mediante el método de descomposición en autovectores, se obtienen las siguientes matrices de entrada a los diferentes mecanismos de aprendizaje automático propuestos.

- Una matriz de 100 muestras × 4 componentes principales representativa del conjunto de naranajas de la clase *Bam*.
- Una matriz de 100 muestras × 4 componentes principales representativa del conjunto de naranajas de la clase *Blood*.
- Una matriz de 100 muestras × 4 componentes principales representativa del conjunto de naranajas de la clase *Thomson*.
- Una matriz de 300 muestras × 4 componentes principales representativa de la combinación de las tres clases de naranjas disponibles *Bam*, *Blood* y *Thomson*.

3.4.3. Regresión

A partir de las matrices de datos obtenidas al aplicar el algoritmo PCA, se aplican los métodos de regresión propuestos con las siguientes consideraciones:

- Del conjunto total de muestras se utilizarán un 80 % de ellas como valores de entrenamiento y un 20 % como valores de test.
- Debido a que el número de valores disponibles para el test es muy reducido se realizarán un número de iteraciones $n_{cic} = 1000$ para cada uno de los experimentos propuestos (*Bam*, *Blood*, *Thomson* y combinación de las tres variedades) a fin de aumentar el número de valores. Esto implicará un total de 20000 valores de test para cada uno de los tres tipos y 60000 valores de test para el conjunto combinación de los tres tipos de naranja. Por lo tanto, la media aproximada de valores de test para cada muestra de naranja disponible será de 200 valores.
- Los resultados obtenidos más importantes para cada uno de los métodos de regresión utilizados son la salida real, la salida estimada, el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE), el coeficiente de correlación lineal (R), el coeficiente de determinación (R^2), la suma de errores cuadrados (SSE) y el error medio absoluto (MAE); siempre teniendo en cuenta los valores de test.

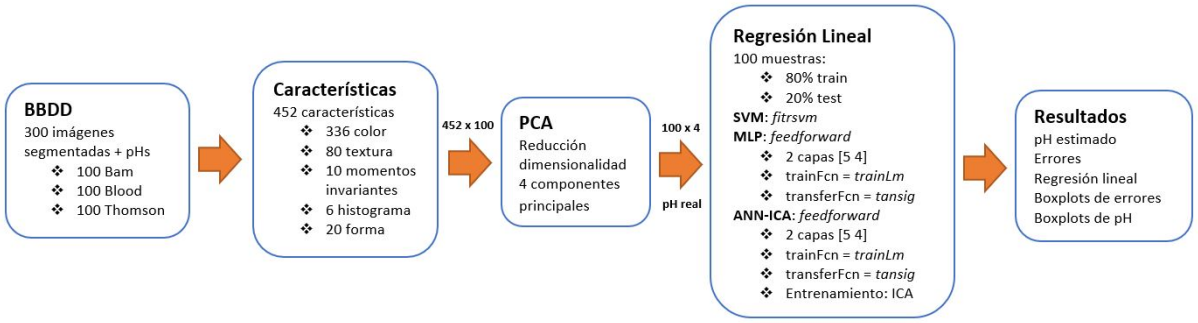


Figura 3.12: Diagrama de las etapas a seguir en el proyecto.

4

Resultados

La implementación de algoritmos de aprendizaje automático para realizar predicciones sobre unos datos de entrada arroja unos resultados que se deben interpretar de la forma correcta con el objetivo de realizar un estudio comparativo de los mismos y decidir cuáles de ellos son los más eficientes para una tarea en concreto.

En este capítulo se abordará la interpretación de los resultados obtenidos a fin de decidir si la estimación del pH de cada una de las muestras de las diferentes variedades de naranjas propuestas es correcto. Para ello se compararán diferentes mecanismos de estimación para evaluar su eficiencia mediante la interpretación de coeficientes de error y de gráficos realizados a partir de los resultados obtenidos de cada una de ellos. Por último, se seleccionará el mejor algoritmo de aprendizaje automático a la hora de estimar el pH de las variedades de naranjas disponibles.

4.1. Técnicas de análisis de resultados

El análisis de resultados se basa en la interpretación y la extracción de conclusiones de un conjunto de datos representados en gráficos o en tablas de valores. La representación de los resultados tiene especial importancia ya que los lectores que los visualizan deben de ser capaces que interpretarlos correctamente y de forma sencilla sin albergar ningún tipo de duda sobre los mismos.

En el presente proyecto se utilizarán diferentes métodos de presentación de resultados para los datos obtenidos: por un lado se utilizarán gráficos (*Regresión lineal, boxplot, diagramas de dispersión*) y tablas de coeficientes de error para representar la estimación correcta o incorrecta del pH para cada una de las variedades de naranjas que forman la base de datos. Por otro lado, se compararán los diferentes mecanismos de aprendizaje automático a partir de boxplots representativos de cada uno de los coeficientes de error propuestos a fin de concluir cuál es el que mejor rendimiento ofrece para la tarea de la estimación del pH.

4.1.1. Regresión lineal simple

Se conoce como regresión lineal simple (Figura 4.1) al cálculo de aquella ecuación correspondiente a la línea que mejor representa la relación entre la respuesta y la variable que la explica, partiendo de la base que la entrada es la variable objetivo y la salida la respuesta estimada esperada.

Al ser una regresión simple sólo se dispone de una única variable predictora. En el caso de que se disponga de más de una variable predictora se tendría que recurrir al uso de una regresión múltiple [57]. En el caso de el presente proyecto, la variable de entrada objetivo está definida por los valores de pH reales medidos para cada una de las variedades de naranjas que conforman la base de datos, mientras que la variable de respuesta la definen los pH estimados para cada caso.

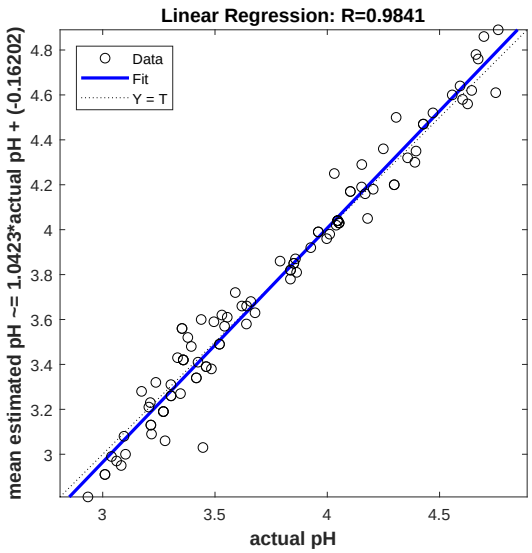


Figura 4.1: Representación gráfica de una regresión lineal simple.

La ecuación que define una regresión lineal simple es una linea recta:

$$y = b_0 + b_1 \cdot x \tag{4.1}$$

donde y es la variable respuesta (salida), x es la variable predictora (entrada), b_0 es la intersección (determina el valor de y cuando x es cero) y b_1 es la pendiente.

La distancia entre los puntos estimados y la línea de regresión se denominan residuos y representan el error cometido de la salida respecto a la entrada. Esto implica que los puntos que estén más cercanos a la línea supondrán un menor error cometido y mejor será el ajuste entre la ecuación de regresión y el dato. En la regresión lineal además se utilizan los mínimos cuadrados (Figura 4.2), que determinan la línea que minimiza la suma de las

distancias verticales cuadradas, desde los puntos hasta la recta (valor del error).

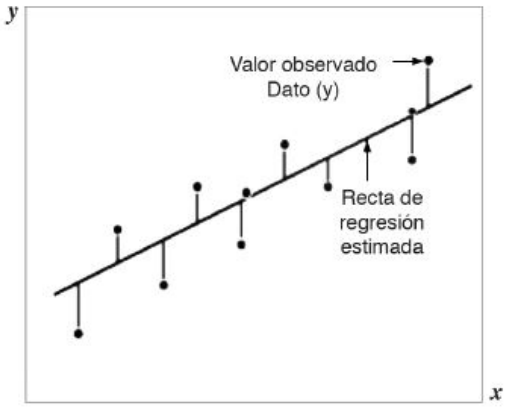


Figura 4.2: Representación de los mínimos cuadrados y del error en la regresión lineal [57].

Coeficientes de error

Los coeficientes de error son medidas empleadas para evaluar el rendimiento de una máquina de aprendizaje automático en términos de la capacidad de predicción de un modelo en concreto. Se definen cinco coeficientes distintos para evaluar este rendimiento [8], [58]:

- **R^2. Coeficiente de determinación**: Es el estadístico que representa la proporción de variación explicada por la regresión y una medida relativa al grado de asociación lineal entre los valores reales y los estimados al igual que el coeficiente de correlación lineal R. Si el valor de R^2 es cero o cercano a cero, el modelo de regresión no tiene un buen ajuste, mientras que si el valor es 1 o próximo a 1 el ajuste es correcto ya que la dependencia lineal del valor estimado respecto al real es elevada. Se define como:

$$R^2 = 1 - \left\{ \frac{\sum_{k=1}^{n}(X_k - X_0)^2}{\sum_{k=1}^{n}(X_k - X_m)^2} \right\} \tag{4.2}$$

  donde X_k representa el valor actual real, $X_m = \frac{1}{n}\sum_{k=1}^{n} X_k$ la media del conjunto de valores actuales reales, X_0 el valor estimado obtenido y n el número de muestras disponibles.

- **MSE. Error cuadrático medio**: medida de la media de los cuadrados de los errores, lo que equivale a la media de los cuadrados de la diferencia entre el valor real menos el valor estimado. Interesan valores cercanos a cero:

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(X_k - Xo)^2 \tag{4.3}$$

- **SSE. Suma de errores cuadrados**: medida relativa al sumatorio de los cuadrados de la diferencia entre el valor actual y la media de los valores estimados:

$$SSE = \sum_{k=1}^{n}(X_k - \bar{X}_0)^2 \tag{4.4}$$

  donde $\bar{X}_0$ simboliza la media de los valores estimados obtenidos.

- **MAE. Error medio absoluto**: es una medida indicativa del error equivalente a la distancia entre la línea de regresión y los valores estimados. Cuanto menor sea esta distancia, menor será el error cometido (MAE $\rightarrow$ 0):

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|X_k - X_0| \tag{4.5}$$

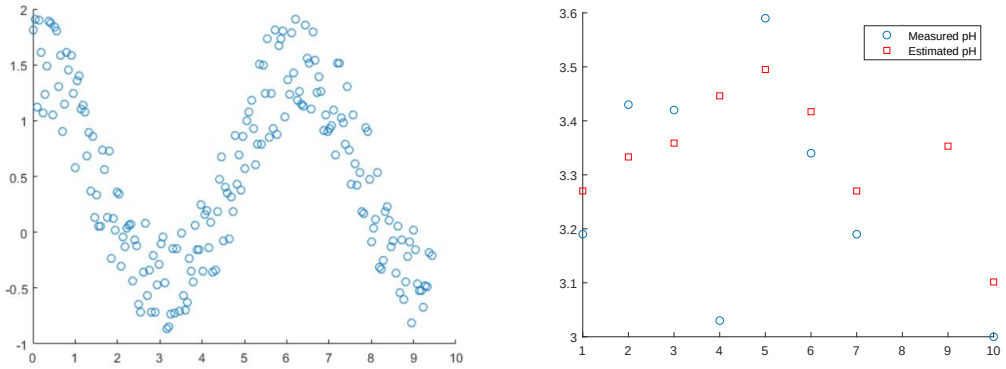
- **RMSE. Raíz del error cuadrático medio**: es una de las medidas más utilizadas a la hora de evaluar el rendimiento de un un modelo de predicción de valores. Se calcula a partir de la raíz cuadrada del error cuadrático medio. Un valor cercano a cero será indicativo de que el error cometido es bajo:

$$RSME = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - X_0)^2} \tag{4.6}$$

4.1.2. Diagrama de dispersión

Un diagrama de dispersión es un gráfico de puntos que permite estudiar la relación entre dos variables. Dadas las dos variables X e Y, se dice que existe una correlación positiva entre ambas si cada vez que aumenta el valor de X aumenta proporcionalmente el valor Y, mientras que existirá un correlación negativa si cada vez que aumenta el valor de X disminuye proporcionalmente el valor de Y.

En el presente proyecto el diagrama de dispersión se va a utilizar de una forma especial debido a que el objetivo no será comparar las variables reales y estimadas de la manera habitual, sino representar el valor de cada una de los pH, reales y estimados, para cada una de las 100 muestras disponibles (Figura 4.3).



(a) (b)

Figura 4.3: a) Ejemplo de un diagrama de dispersión tradicional [disp] y b) ejemplo de un diagrama de dispersión implementado en el presente proyecto.

4.1.3. Boxplot

Es un gráfico basado en cuartiles en el que se presenta una distribución de una o más series de datos cuantitativos. Este tipo de representación utiliza una sola escala, que es la correspondiente a la variable de los datos que se presentan [59].

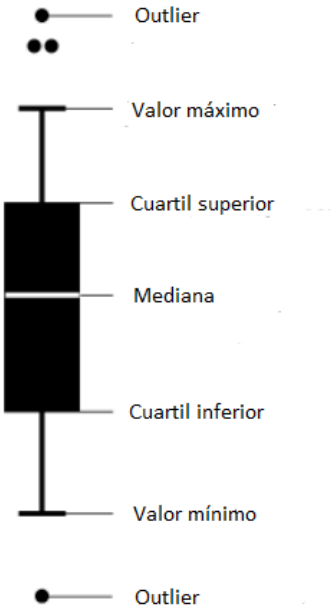


Figura 4.4: Representación gráfica de un boxplot.

Los elementos que constituyen el boxplot (Figura 4.4) son los siguientes:

- Cuartil inferior: representa el valor mayor que el 25 % de los valores de la distribución.

- Cuartil superior: representa el valor que sobrepasa al 75 % de los valores de la distribución.
- Caja: rectángulo que presenta el recorrido intercuartílico (RIC) de la distribución, es decir, el conjunto de valores que van desde el cuartil inferior hasta el cuartil superior.
- Mediana: linea horizontal dibujada dentro de la caja que corresponde al valor de esta medida.
- Bigotes *(Whiskers)*: lineas verticales que salen de los costados de la caja y que sirven como referencia para ubicar todas aquellas observaciones que están fuera del 50 % central de la distribución.
- Outlier: representación de aquellos valores que se encuentran fuera del conjunto RIC y los bigotes de la distribución. Se considera un outlier el valor que supera en un factor de 3/2 el valor de los cuartiles.
- Valores máximo y mínimo: valores extremos de la distribución excluyendo los outliers.

En función de la distribución de los valores del boxplot podemos distinguir entre tres tipos de gráficos: si los datos siguen una distribución normal la mediana se situará en el centro de la caja y cada bigote tendrá una longitud de 1.5 veces la caja. Si la distribución es asimétrica podemos distinguir entre distribuciones con sesgo negativo (mediana más cercana al cuartil superior y bigote inferior más largo que el superior) y con sesgo positivo (mediana mas cercana al cuartil inferior y bigote superior más largo que el inferior).

4.2. Comparación de los métodos de regresión utilizados: boxplots de errores y de pH

El primer bloque de resultados obtenidos se basa en la comparación de los tres métodos de regresión utilizados propuestos en la sección 3.3: SVM, MLP y ANN-ICA. El objetivo principal es seleccionar el mejor mecanismo de aprendizaje automático a la hora de estimar el grado de acidez de las diferentes muestras de naranjas que conforman la base de datos propuesta.

Para conseguir el objetivo propuesto se emplearán gráficos representativos de boxplots a fin de analizar dos variables principales: por un lado los coeficientes de error derivados de la regresión lineal para cada uno de los métodos propuestos: coeficiente de determinación (R^2), coeficiente de correlación lineal (R), error cuadrático medio (MSE), raíz del error

cuadrático medio (RMSE), error medio absoluto (MAE) y suma de errores cuadrados (SSE).Y, por otro lado, la estimación del grado de acidez de cada muestra disponible y su relación con el pH real medido para cada una de ellas, con el objetivo de analizar el error cometido a partir de la siguiente relación:

$$\text{error pH} = \text{pH real} - \text{pH estimado} \tag{4.7}$$

4.2.1. Boxplots de errores

En la sección 4.1.1 se explicaron los seis coeficientes de error que se analizarán como boxplots con el objetivo de seleccionar el mecanismo de aprendizaje automático más eficiente. La forma de proceder consistirá en mostrar los boxplots representativos de cada uno de los errores para los métodos de regresión propuestos (SVM, MLP y ANN-ICA) dividiéndolos en tres gráficos distintos: por un lado, boxplots de R y R^2; por otro lado, boxplots de MSE, MAE y RMSE; y para finalizar, boxplots de SSE. El estudio de los boxplots de errores se realizará para las tres variedades de naranjas propuestas en la base de datos (*Bam*, *Blood* y *Thomson*) y para un nuevo conjunto de datos formado a partir de la combinación de las tres variedades (*3-Variety*). Para un mayor detalle de los resultados obtenidos se dispondrá del apéndice B.2.

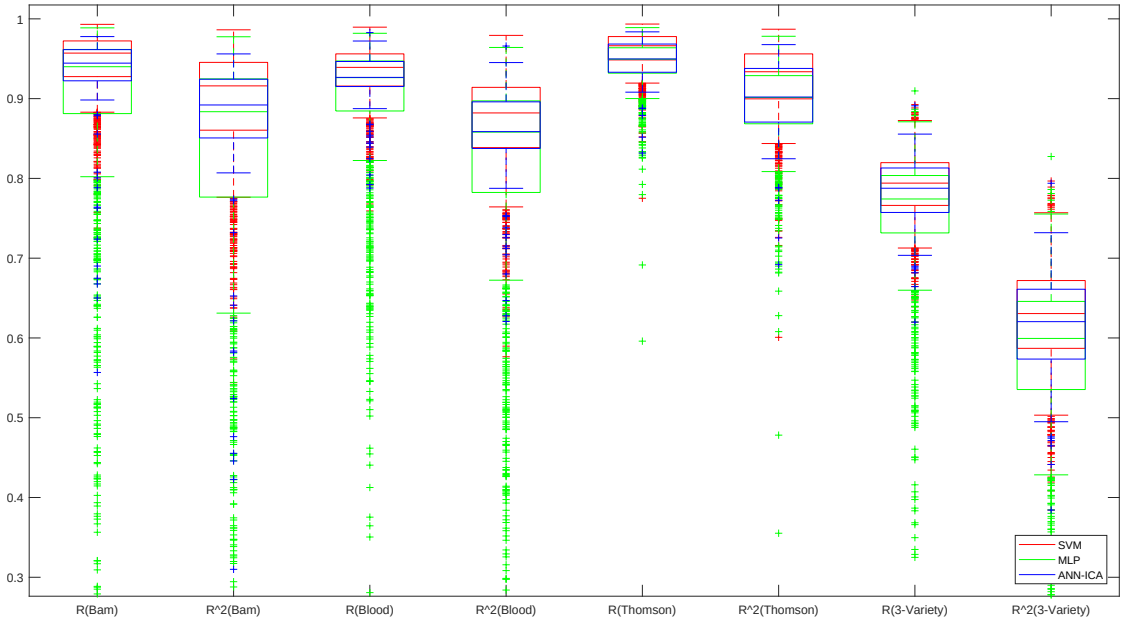


Figura 4.5: Boxplots de los coeficientes de error R y R^2 para los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul) y un número de iteraciones $n_{cic} = 1000$.

En la figura 4.5 se presentan los coeficientes de error R y R^2 para los mecanismos de regresión SVM, MLP y ANN-ICA. Los coeficientes de determinación y de correlación lineal indican una buena estimación en la regresión si su valor es cercano a 1, mientras

que un valor cercano a cero indica que el modelo de regresión no dispone de un buen ajuste, por lo tanto, aquellos boxplots cuya media sea cercana a 1 serán los indicativos de una buena estimación mediante el mecanismo de regresión no intrusivo representado.

Se puede observar que para los métodos de regresión SVM (rojo) y ANN-ICA (azul) los valores del coeficiente de correlación lineal se sitúan en el rango de valores de 0.9 a 1 para las variedades de naranjas *Bam*, *Blood* y *Thomson*, mientras que los valores de R y R^2 empeoran para el estudio de las tres variedades combinadas. Sin embargo para MLP (verde) los valores de los coeficientes R y R^2 empeoran en las cuatro situaciones propuestas, produciéndose además gran cantidad de valores muy alejados de la media del error representado (outliers).

Para los coeficientes de error MSE, RMSE y MAE, encargados de la evaluación del rendimiento de la regresión (Figura 4.6), interesan valores lo mas cercanos al cero, ya que esto implicará que el error en la estimación del grado de acidez es bajo en los tres métodos de regresión propuestos. Analizando el siguiente gráfico de boxplots se observa que los valores de los tres coeficientes de error son reducidos para las tres variedades de naranjas disponibles, sin embargo el error aumenta para el conjunto de las tres variedades combinadas. En concreto, los métodos de regresión SVM (rojo) y ANN-ICA (azul) ofrecen resultados similares para los tres coeficientes con muy pocos valores que se salgan del recorrido intercuartílico, mientras que el método MLP (verde) presenta gran cantidad de outliers que aumentan los valores del error cometido.

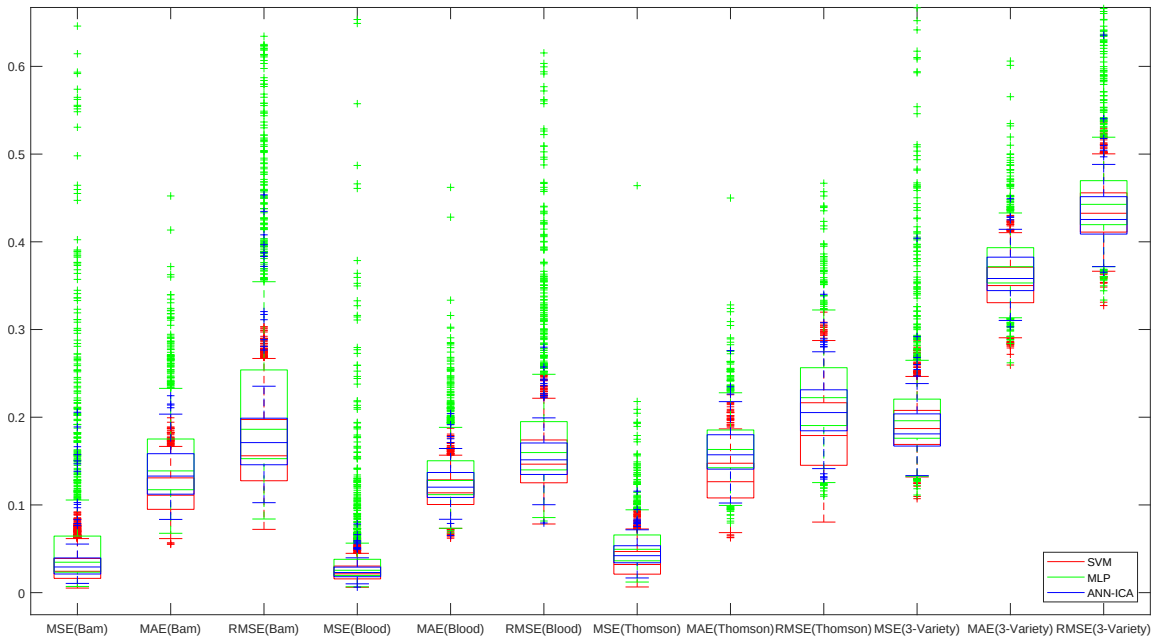


Figura 4.6: Boxplots de los coeficientes de error MSE, MAE y RMSE para los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul) y un número de iteraciones $n_{cic} = 1000$.

Por último, el coeficiente de error SSE (Figura 4.7) representa una medida de la discrepancia entre el modelo real y el estimado. Si el valor del error SSE es cero, no existirá ningún tipo de error entre ambos modelos y estos coincidirán. Analizando el gráfico se puede deducir, como en los resultados vistos anteriormente, que la combinación de variedades de naranjas ofrece peores resultados que su análisis individual para los tres métodos de regresión propuestos y, por otro lado, que el mecanismo MLP (verde) introduce mayor cantidad de valores con un error alejado de la media del total de iteraciones realizadas (mayor cantidad y valor de los outliers representados).

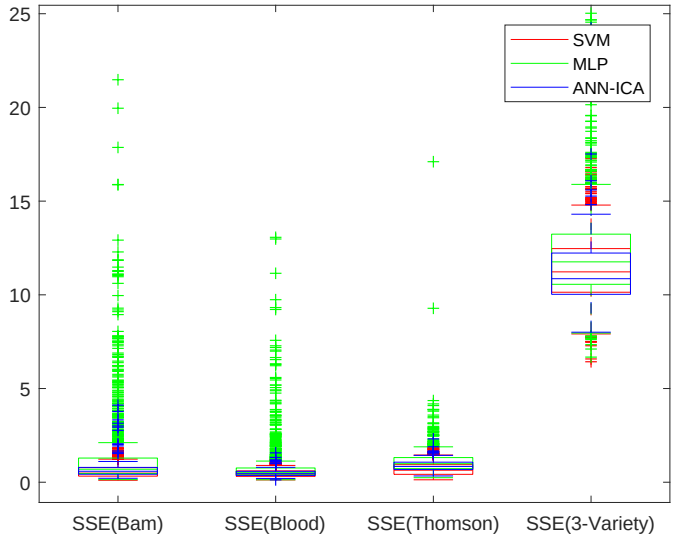


Figura 4.7: Boxplots del coeficiente de error SSE para los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul) y un número de iteraciones $n_{cic} = 1000$.

A partir del análisis los gráficos representativos de los coeficientes de error se puede concluir que, por un lado, los experimentos realizados para las tres variedades de naranjas de la base de datos, *Bam*, *Blood* y *Thomson*, de forma individual ofrecen mejores resultados que el experimento realizado con el conjunto combinación de las variedades. El motivo principal se debe a que diferentes muestras de distintas variedades de naranja pueden tener el mismo valor de grado de acidez (pH) pero características totalmente distintas, esto hace que a la hora de estimar se produzcan un número mayor de errores.

Por otro lado, en lo que se refiere a los mecanismos de regresión no intrusivos utilizados, SVM y ANN-ICA ofrecen mejores resultados que MLP. Si se analizan los valores de media y desviación típica de cada uno de los errores (Tablas 4.1, 4.2, 4.3 y 4.4) se observa que, siendo los valores medios de los errores similares, los valores de desviación típica para la red neuronal MLP son más altos que para los métodos SVM y ANN-ICA, lo cual indica que la dispersión de los datos respecto a la media es más elevada y de ahí la gran cantidad de outliers que se representan en los boxplots.

Tablas de errores

Bam		**R**	**R^2**	**MSE**	**RMSE**	**MAE**	**SSE**
SVM	mean	0.9428	0.8889	0.0309	0.1672	0.1139	0.6171
	standard deviation	0.0407	0.0744	0.0200	0.0538	0.0265	0.4004
MLP	mean	0.9306	0.8660	0.0763	0.2366	0.1520	1.5252
	standard deviation	0.1431	0.2027	0.1232	0.1425	0.0515	2.4640
ANN-ICA	mean	0.9419	0.8872	0.0417	0.1907	0.1374	0.8332
	standard deviation	0.0790	0.1295	0.0387	0.0731	0.0314	0.7740

Tabla 4.1: Media y desviación típica de los coeficientes de error para los métodos de regresión SVM, MLP y ANN-ICA en función de la variedad de naranja *Bam*.

Blood		**R**	**R^2**	**MSE**	**RMSE**	**MAE**	**SSE**
SVM	mean	0.9283	0.8617	0.0239	0.1510	0.1154	0.4787
	standard deviation	0.0335	0.0612	0.0109	0.0339	0.0208	0.2179
MLP	mean	0.9413	0.8860	0.0466	0.1912	0.1373	0.9321
	standard deviation	0.1075	0.1533	0.0826	0.1003	0.0409	1.6520
ANN-ICA	mean	0.9424	0.8881	0.0263	0.1583	0.1237	0.5261
	standard deviation	0.0380	0.0677	0.0126	0.0353	0.0242	0.2526

Tabla 4.2: Media y desviación típica de los coeficientes de error para los métodos de regresión SVM, MLP y ANN-ICA en función de la variedad de naranja *Blood*.

Thomson		**R**	**R^2**	**MSE**	**RMSE**	**MAE**	**SSE**
SVM	mean	0.9582	0.9181	0.0359	0.1835	0.1289	0.7189
	standard deviation	0.0245	0.0460	0.0184	0.0479	0.0290	0.3689
MLP	mean	0.9550	0.9120	0.0557	0.2281	0.1665	1.1140
	standard deviation	0.0314	0.0564	0.0399	0.0607	0.0377	0.7972
ANN-ICA	mean	0.9559	0.9137	0.0458	0.2096	0.1603	0.9154
	standard deviation	0.0271	0.0503	0.0188	0.0430	0.0310	0.3754

Tabla 4.3: Media y desviación típica de los coeficientes de error para los métodos de regresión SVM, MLP y ANN-ICA en función de la variedad de naranja *Thomson*.

3-Variety		**R**	**R^2**	**MSE**	**RMSE**	**MAE**	**SSE**
SVM	mean	0.7906	0.6267	0.1899	0.4345	0.3506	11.3930
	standard deviation	0.0411	0.0641	0.0291	0.0333	0.0298	1.7482
MLP	mean	0.7514	0.5749	0.2237	0.4641	0.3752	13.4238
	standard deviation	0.1011	0.1128	0.1174	0.0916	0.0374	7.0444
ANN-ICA	mean	0.7810	0.6122	0.1888	0.4327	0.3625	11.3276
	standard deviation	0.0463	0.0708	0.0368	0.0392	0.0280	2.2053

Tabla 4.4: Media y desviación típica de los coeficientes de error para los métodos de regresión SVM, MLP y ANN-ICA en función del conjunto combinación de variedades (*3-Variety*).

4.2.2. Boxplots de pH

Una vez analizados los coeficientes de error propuestos para cada uno de los métodos de regresión estudiados, la siguiente etapa consistirá en analizar el error cometido en la estimación del grado de acidez para cada una de las variedades de naranjas de la base de datos propuesta. En esta situación se tendrán en cuenta dos variables relacionadas a través de la ecuación (4.7): el pH real medido y el pH estimado a partir de los tres métodos de regresión propuestos (Apéndice B.1).

El objetivo principal será analizar los boxplots derivados de estas dos variables y comparar el rendimiento en términos de estimación del grado de acidez de los mecanismos de regresión no intrusivos SVM, MLP y ANN-ICA. Para ello, se tienen en cuenta las siguientes premisas: el número de iteraciones (experimentos) n_{cic} realizadas han sido un total de 1000, por lo que teniendo en cuenta que la base de datos está formada por 300 muestras, 100 relativas a cada variedad de naranja disponible (*Bam*, *Blood* y *Thomson*) y que el conjunto de datos de test equivale al 20 % de los datos totales (de cada 100 muestras, 20 de test), cada boxplot de cada muestra estará formado por una media aproximada de 200 valores:

$$\frac{1000 \text{ iteraciones} \cdot 20 \text{ valores/iteración}}{100 \text{ muestras totales}} = 200 \text{ valores/muestra} \tag{4.8}$$

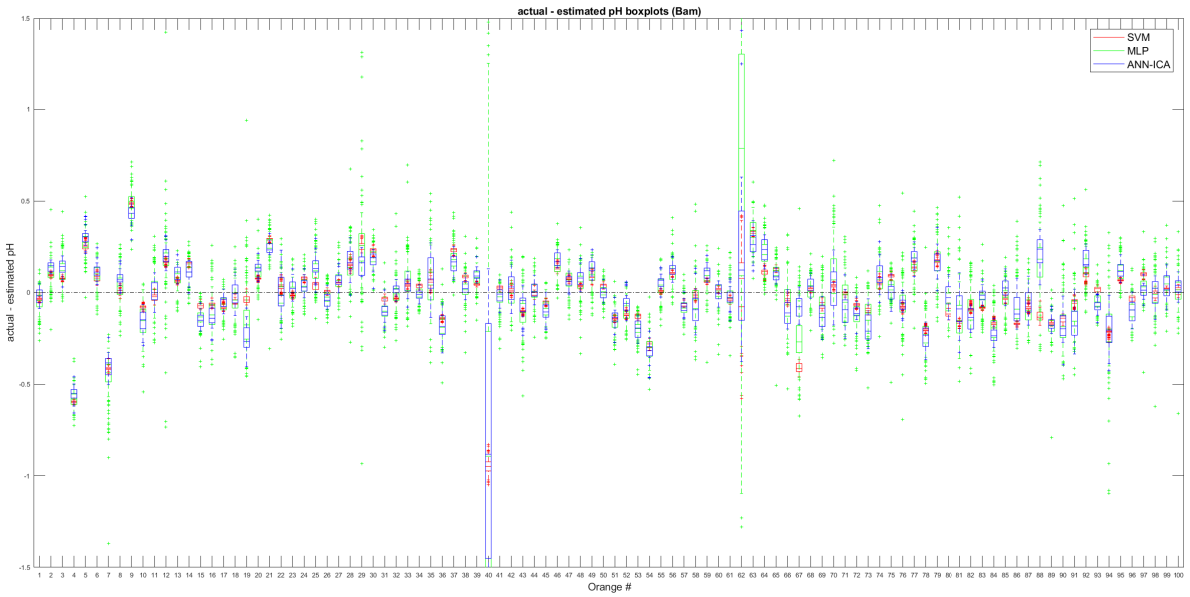


Figura 4.8: Boxplots relativos a las 100 muestras de la variedad de naranja *Bam* indicativos de la diferencia entre el pH real y el pH estimado para los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul). (n_{cic} = 1000, 200 valores de media por muestra).

Por lo tanto, cada gráfico estará formado por 100 boxplots, uno por cada muestra de naranja disponible, generado a partir de 200 valores de grado de acidez, indicando el error cometido entre el valor real y la estimación. Debido a la relación impuesta por la ecuación 4.7, aquellas muestras cuyos boxplots sitúen su media cerca de cero serán las que contengan un error menor en la estimación, mientras que aquellos boxplots que tengan su media alejada del cero implicarán errores más elevados.

Para la variedad de naranja *Bam* (Figura 4.8) se observa que la mayoría de las muestras disponibles posicionan sus boxplots dentro del rango de error [-0.5 0.5], excepto dos muestras, la muestra 40 y la muestra 62, que implican un error más elevado. En concreto, para los métodos de regresión SVM (rojo) y ANN-ICA (azul) se observa que el tamaño de los boxplots para la mayoría de las muestras es reducido y cercanos al valor de error cero, y apenas se aprecia dispersión en los valores del error (número de outliers reducido) lo que implica una buena estimación de los valores del grado de acidez; en cambio para la red neuronal MLP (verde), aunque el tamaño de los boxplots es reducido y se sitúan cercanos al valor cero, la dispersión de los valores de error es elevada ya que se aprecian gran cantidad de outliers para cada una de las muestras, lo que da pie a que la estimación del valor del pH no es adecuada.

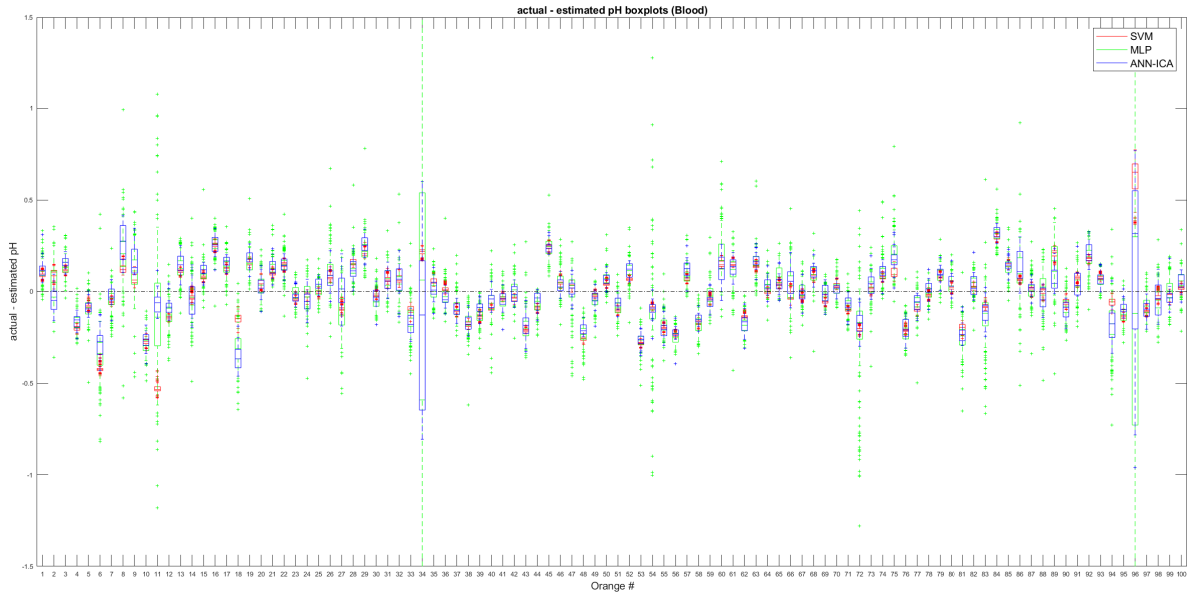


Figura 4.9: Boxplots relativos a las 100 muestras de la variedad de naranja *Blood* indicativos de la diferencia entre el pH real y el pH estimado para los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul). (n_{cic} = 1000, 200 valores de media por muestra).

En la figura 4.9 se representan el conjunto de boxplots para cada muestra de la variedad de naranja *Blood*. Al igual que el gráfico de la variedad de naranja *Bam*, los

boxplots son reducidos en tamaño y se situán en un rango de valores de [-0.5 0.5] excepto las muestras 34 y 96 que introducen altas cantidades de error en la estimación debido a defectos en las imágenes recabadas en la base de datos como brillos, o una segmentación errónea. El comportamiento de los tres métodos de regresión es similar ya que la mayoría de los boxplots se encuentran cercanos al valor cero de error, pero al igual que en el caso anterior la red neuronal MLP (verde) provoca un mayor rango de dispersión en los valores estimados del pH, representados en el alto número de outliers por muestra.

Por último, para la variedad de naranja *Thomson* (Figura 4.10), los resultados obtenidos son muy similares a los representados en las figuras 4.8 y 4.9. La muestra 4 produce una cantidad de errores más elevada, lo que hace que este fuera del rango [-0.5 0.5] que incluye a casi todo el conjunto de muestras disponibles, y la red MLP (verde) ofrece los mismos problemas que los resultados obtenidos para las dos variedades de naranjas estudiadas anteriormente.

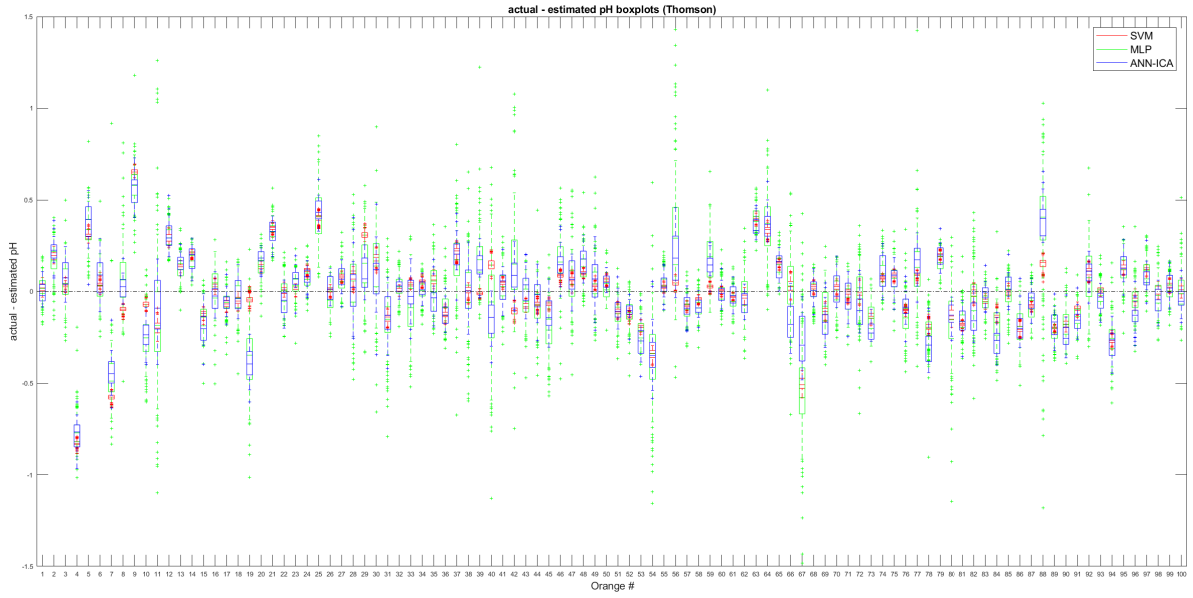


Figura 4.10: Boxplots relativos a las 100 muestras de la variedad de naranja *Thomson* indicativos de la diferencia entre el pH real y el pH estimado para los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul). (n_{cic} = 1000, 200 valores de media por muestra).

Por lo tanto, una vez analizados los errores respecto a la estimación a partir de los boxplots referentes a las tres variedades de naranjas de la base de datos, se puede concluir que el comportamiento de los métodos de regresión no intrusivos SVM y ANN-ICA ofrece resultados muy similares con valores de error reducidos en comparación a la red neuronal MLP debido a la dispersión en el conjunto de los 200 valores por muestra estudiados. Así, combinando las ideas concluidas en la sección 4.2.1 referente a los coeficientes de

error y los resultados obtenidos en la presente sección, los mecanismos de regresión no intrusivos más adecuados para la estimación del grado de acidez para las tres variedades de naranjas disponibles son SVM y ANN-ICA. En la siguiente sección se analizaran los tres mecanismos de regresión para cada variedad de naranja disponible en función de los valores estimados medios de pH obtenidos para cada muestra de la base de datos.

4.3. Resultados para cada variedad de naranja disponible: regresión lineal y valores de pH estimados

El segundo bloque de resultados obtenidos se basa en el análisis de cada una de las tres variedades de naranjas que conforman la base de datos propuesta (*Bam*, *Blood* y *Thomson*) enunciada en la sección 2.3 y en el apéndice A del presente proyecto.

A partir de los resultados obtenidos en la sección 4.2 se ha concluido que los mejores mecanismos de regresión no intrusivos para realizar una estimación del grado de acidez para las variedades de naranjas propuestas son SVM y ANN-ICA, siendo la red neuronal MLP la que peor resultados ha ofrecido. En esta sección se comprobara si se sigue cumpliendo esta tendencia.

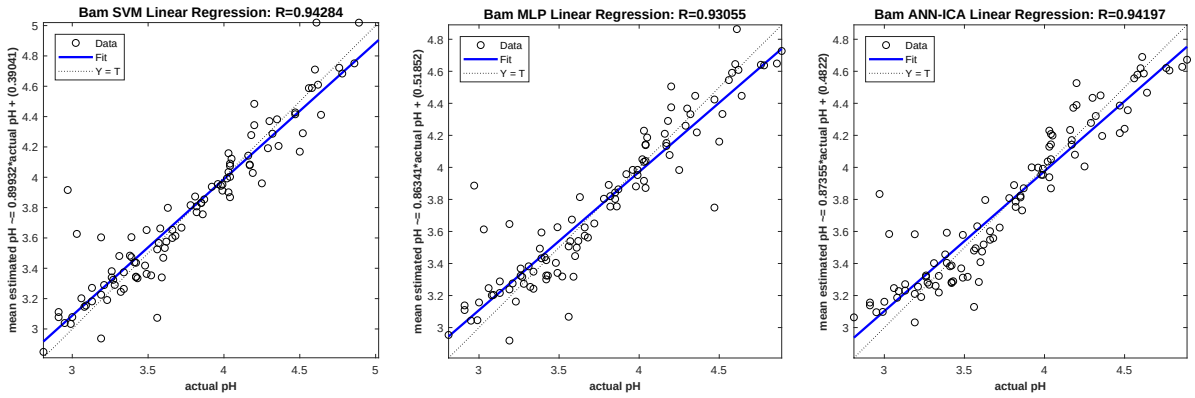
4.3.1. Regresión lineal

La estimación del grado de acidez de cada una de las muestras de naranjas de la base de datos propuesta es el principal resultado a obtener en el presente proyecto. Para comprobar la fiabilidad del mismo es necesario realizar una comparación entre en pH real medido y el pH estimado mediante los mecanismos de regresión no intrusivos propuestos.

La regresión lineal (sección 4.1.1) permite realizar una comparación fiable entre el pH medido y el pH estimado a partir de la ecuación de la recta de regresión y del valor del coeficiente de correlación lineal R. A partir de los métodos de regresión SVM y MLP se calculará el grado de acidez estimado medio para cada muestra de las variedades de naranjas propuestas en la base de datos, sabiendo que se han realizado 1000 iteraciones para cada método propuesto y que cada muestra dispone de una media aproximada de 200 valores estimados de pH. El gráfico representativo de la regresión mostrará la relación entre los valores medidos y estimados.

Así, el objetivo de la regresión lineal consistirá en comparar la tendencia de la ecuación de la recta de regresión para las distintas variedades de naranjas (*Bam*, *Blood* y *Thomson*) mediante los métodos de regresión SVM, MLP y ANN-ICA.

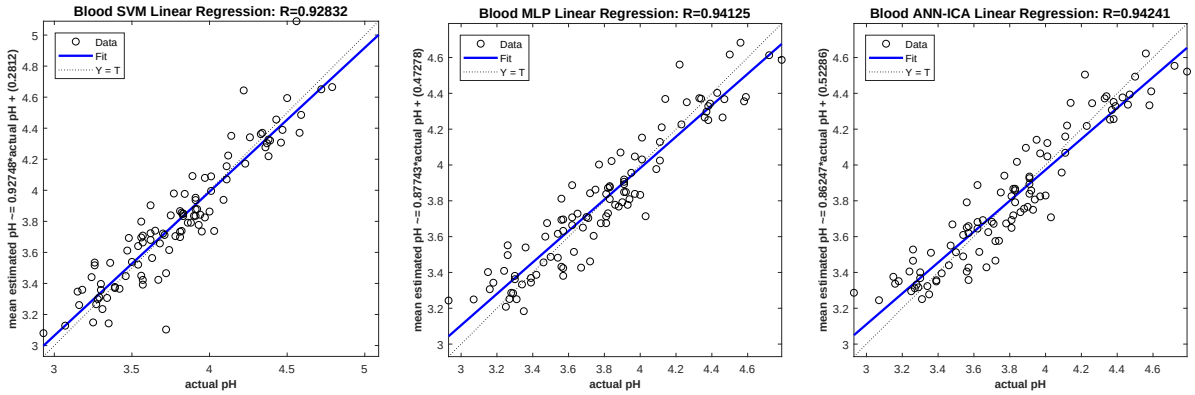
Para la variedad de naranja *Bam* (Figura 4.11) se observa que la tendencia en la estimación del grado de acidez obtenida se mantiene muy cercana a la situación ideal. El coeficiente de correlación lineal R posee un valor aproximado de 0.94 para los métodos



(a) (b) (c)

Figura 4.11: Representación de la regresión lineal para la variedad de naranja *Bam* (n_{cic} = 1000, 200 valores de media por muestra) para a) SVM y b) MLP c) ANN-ICA.

de regresión SVM y ANN-ICA, y un valor de R = 0,93055 para la red neuronal MLP, por lo que la estimación del conjunto de valores del grado de acidez será más fiable para los métodos de regresión SVM y ANN-ICA.



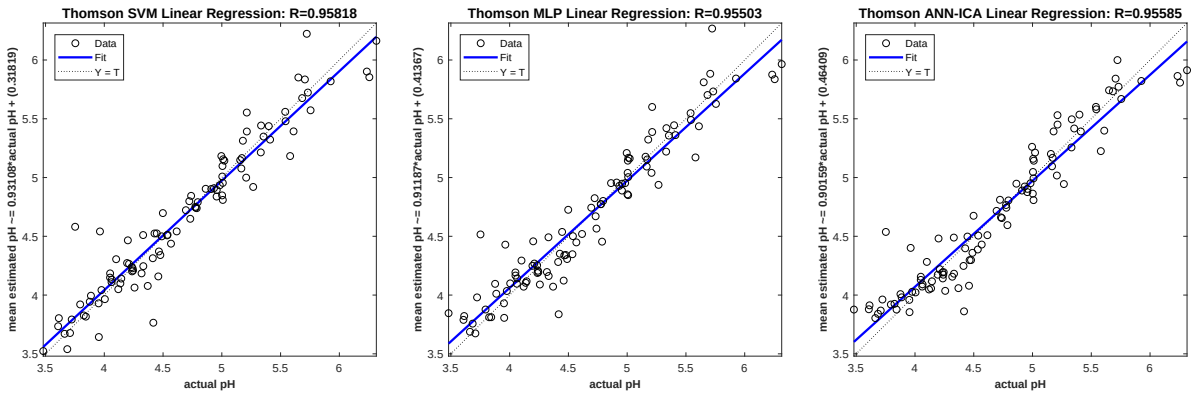
(a) (b) (c)

Figura 4.12: Representación de la regresión lineal para la variedad de naranja *Blood* (n_{cic} = 1000, 200 valores de media por muestra) para a) SVM b) MLP y c) ANN-ICA.

En la figura 4.12 se muestran las representaciones de las gráficas de regresión lineal para la variedad de naranja *Blood*. El coeficiente de correlación lineal para el método de regresión SVM (R = 0,92832) es menor que el obtenido para los métodos MLP (R = 0,94125) y ANN-ICA (R = 0,94241), por lo que la estimación del conjunto de valores de pH de cada muestra será más fiable utilizando estos dos últimos mecanismos.

Por último, para la variedad de naranja *Thomson* (Figura 4.13), el coeficiente de correlación lineal R obtenido en la regresión lineal para los tres mecanismos de regresión propuestos es similar y posee el valor más elevado para las tres variedades de naranjas

propuestas (R = 0,95).



(a) (b) (c)

Figura 4.13: Representación de la regresión lineal para la variedad de naranja *Thomson* (n_{cic} = 1000, 200 valores de media por muestra) para a) SVM b) MLP y c) ANN-ICA.

Por lo tanto, a la vista de los resultados obtenidos, tanto en las representaciones de la regresión lineal como en los valores de pH estimados frente a los reales, se puede concluir que la estimación mediante los métodos de regresión propuestos es muy similar para la variedad de naranja *Thomson*, mientras que para las variedades *Bam* y *Blood* se observan diferencias: los métodos SVM y ANN-ICA ofrece mejores resultados en términos del coeficiente de correlación lineal R que el método de regresión MLP para la variedad de naranja *Bam*, mientras que para la variedad *Blood* los peores resultados se obtienen para SVM. Por otro lado, en lo que se refiere a las variedades de naranjas, la variedad *Thomson* es la que ofrece mejores resultados en la estimación para el conjunto de las 100 muestras disponibles ya que el coeficiente de determinación obtenido para ambos mecanismos de regresión se sitúa en torno a 0.95; la variedad *Bam* también ofrece buenos resultados en la estimación con un valor de R de aproximadamente 0.94 para los métodos SVM y ANN-ICA, y un valor de 0.93 para MLP, mientras que para la variedad de naranja *Blood* también se encuentran discrepancias entre criterios de regresión ya que SVM ofrece una R de 0.92, en cambio, MLP y ANN-ICA incrementan este valor a 0.94. Esto implica que para las variedades *Bam* y *Blood* no se puede discernir cuál de ellas ofrece mejores resultados en la estimación ya que depende del método de regresión utilizado.

4.3.2. pH real vs pH estimado

Los resultados obtenidos a partir de la regresión lineal muestran una tendencia de los valores estimados frente a los reales a partir de la ecuación de la recta estimada a través de estos valores, pero no se permite observar una comparación real entre los valores estimados del pH mediante los métodos de regresión propuestos y el valor medido real.

En esta sección se representarán los valores estimados medios del grado de acidez para los métodos de regresión SVM, MLP y ANN-ICA frente al valor real del pH para observar la diferencia real entre los valores de cada una de las muestras de las tres variedades de naranjas disponibles *Bam*, *Blood* y *Thomson* (Figuras 4.14, 4.15 y 4.16). Además, para una mejor interpretación de los resultados obtenidos, se realizará un desglose de cada una de las gráficas obtenidas en el apéndice B.3.

La notación utilizada en las siguientes representaciones será la siguiente:

- Valor de pH medido real: negro
- Valor medio estimado de pH mediante el método de regresión SVM: rojo.
- . Valor medio estimado de pH mediante el método de regresión MLP: verde
- Valor medio estimado de pH mediante el método de regresión ANN-ICA: azul.

Un valor estimado medio de pH situado lo más cercano posible al valor medido real de pH de una muestra en concreto implicará que la estimación realizada es muy eficiente.

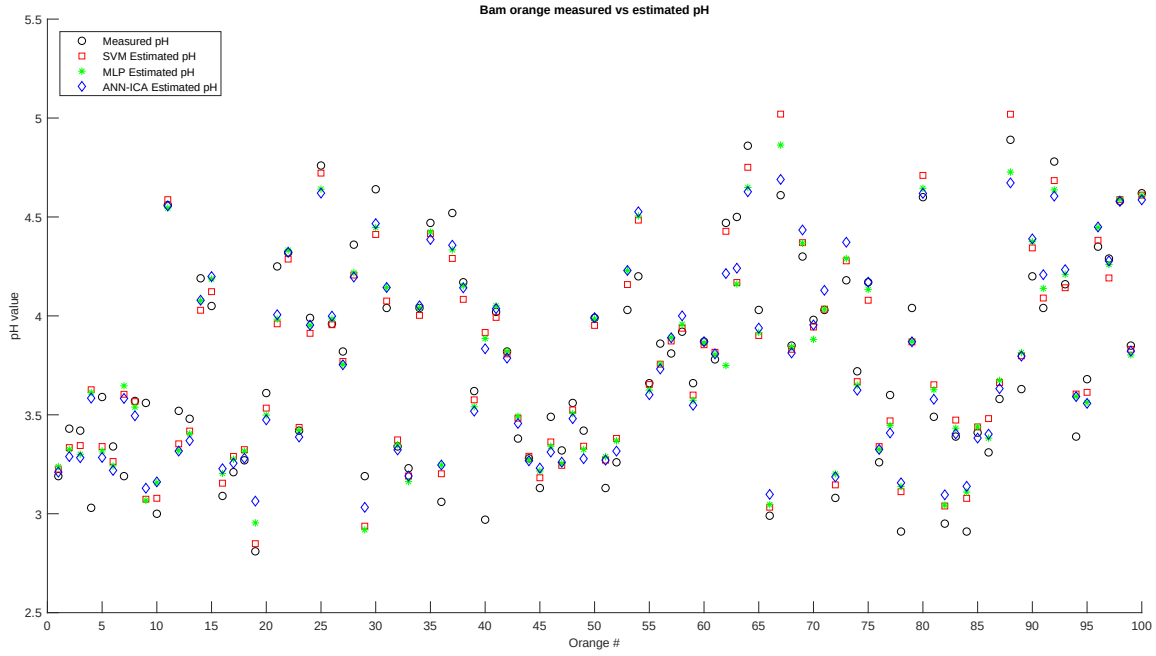


Figura 4.14: Comparación del valor de pH real (negro) con los valores estimados medios de pH obtenidos a partir de los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul) para la variedad de naranja *Bam*. ($n_{cic} = 1000$, 200 valores de media por muestra).

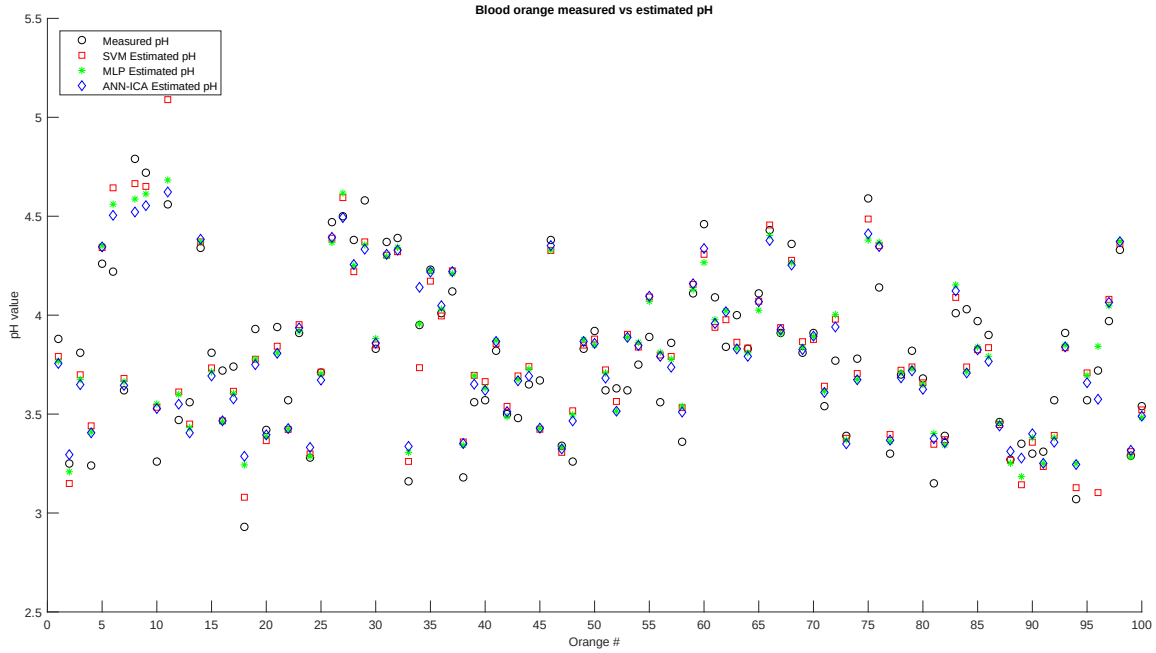


Figura 4.15: Comparación del valor de pH real (negro) con los valores estimados medios de pH obtenidos a partir de los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul) para la variedad de naranja *Blood*. ($n_{cic} = 1000$, 200 valores de media por muestra).

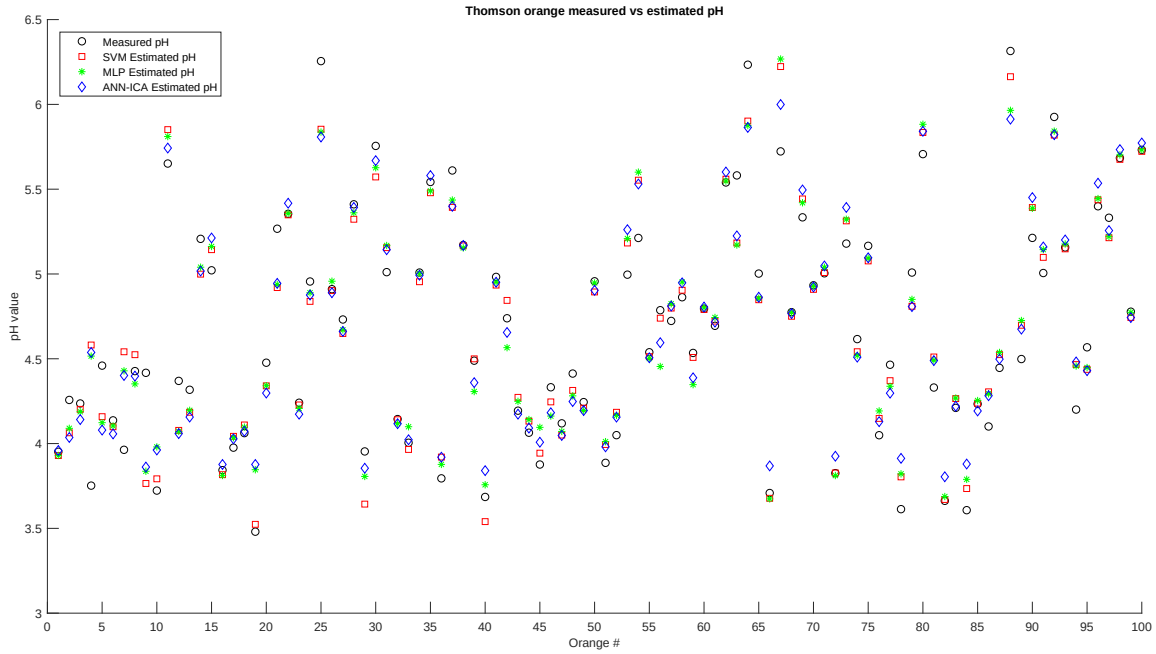


Figura 4.16: Comparación del valor de pH real (negro) con los valores estimados medios de pH obtenidos a partir de los métodos de regresión SVM (rojo), MLP (verde) y ANN-ICA (azul) para la variedad de naranja *Thomson*. ($n_{cic} = 1000$, 200 valores de media por muestra).

Tablas de valores de pH

Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH	Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH	Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH
1	3,19	3,23 ± 0,007	3,24 ± 0,057	3,21 ± 0,043	35	4,47	4,42 ± 0,020	4,42 ± 0,131	4,39 ± 0,107	69	4,3	4,37 ± 0,017	4,37 ± 0,087	4,43 ± 0,066
2	3,43	3,33 ± 0,006	3,32 ± 0,058	3,29 ± 0,026	36	3,06	3,20 ± 0,008	3,25 ± 0,068	3,25 ± 0,070	70	3,98	3,94 ± 0,010	3,88 ± 0,152	3,95 ± 0,111
3	3,42	3,34 ± 0,006	3,30 ± 0,071	3,28 ± 0,044	37	4,52	4,29 ± 0,015	4,33 ± 0,064	4,36 ± 0,051	71	4,03	4,03 ± 0,009	4,03 ± 0,086	4,13 ± 0,085
4	3,03	3,63 ± 0,006	3,61 ± 0,048	3,58 ± 0,042	38	4,17	4,08 ± 0,011	4,15 ± 0,062	4,14 ± 0,037	72	3,08	3,15 ± 0,009	3,20 ± 0,072	3,19 ± 0,045
5	3,59	3,34 ± 0,014	3,32 ± 0,056	3,28 ± 0,052	39	3,62	3,58 ± 0,007	3,54 ± 0,052	3,52 ± 0,035	73	4,18	4,28 ± 0,013	4,29 ± 0,079	4,37 ± 0,080
6	3,34	3,26 ± 0,015	3,24 ± 0,058	3,22 ± 0,047	40	2,97	3,92 ± 0,041	3,89 ± 1,203	3,83 ± 0,662	74	3,72	3,67 ± 0,014	3,65 ± 0,082	3,62 ± 0,085
7	3,19	3,60 ± 0,010	3,65 ± 0,116	3,58 ± 0,063	41	4,02	3,99 ± 0,011	4,05 ± 0,054	4,04 ± 0,047	75	4,17	4,08 ± 0,011	4,13 ± 0,058	4,17 ± 0,053
8	3,57	3,57 ± 0,011	3,54 ± 0,077	3,50 ± 0,047	42	3,82	3,81 ± 0,013	3,82 ± 0,087	3,79 ± 0,080	76	3,26	3,34 ± 0,009	3,32 ± 0,111	3,33 ± 0,047
9	3,56	3,07 ± 0,009	3,07 ± 0,060	3,13 ± 0,051	43	3,38	3,48 ± 0,008	3,49 ± 0,104	3,46 ± 0,073	77	3,6	3,47 ± 0,013	3,45 ± 0,071	3,41 ± 0,048
10	3	3,08 ± 0,008	3,16 ± 0,091	3,16 ± 0,064	44	3,28	3,29 ± 0,014	3,27 ± 0,061	3,27 ± 0,046	78	2,91	3,11 ± 0,012	3,14 ± 0,087	3,16 ± 0,063
11	4,56	4,59 ± 0,018	4,55 ± 0,090	4,56 ± 0,059	45	3,13	3,18 ± 0,009	3,22 ± 0,057	3,23 ± 0,058	79	4,04	3,87 ± 0,013	3,87 ± 0,094	3,87 ± 0,083
12	3,52	3,35 ± 0,009	3,32 ± 0,325	3,32 ± 0,055	46	3,49	3,36 ± 0,014	3,34 ± 0,058	3,31 ± 0,034	80	4,6	4,71 ± 0,022	4,65 ± 0,086	4,62 ± 0,076
13	3,48	3,42 ± 0,007	3,40 ± 0,043	3,37 ± 0,044	47	3,32	3,24 ± 0,008	3,25 ± 0,055	3,26 ± 0,035	81	3,49	3,65 ± 0,011	3,63 ± 0,139	3,58 ± 0,116
14	4,19	4,03 ± 0,011	4,08 ± 0,048	4,08 ± 0,056	48	3,56	3,52 ± 0,007	3,51 ± 0,073	3,48 ± 0,049	82	2,95	3,04 ± 0,013	3,04 ± 0,102	3,10 ± 0,060
15	4,05	4,12 ± 0,014	4,19 ± 0,049	4,20 ± 0,043	49	3,42	3,34 ± 0,014	3,32 ± 0,055	3,28 ± 0,043	83	3,39	3,47 ± 0,005	3,43 ± 0,062	3,40 ± 0,036
16	3,09	3,15 ± 0,010	3,20 ± 0,094	3,23 ± 0,059	50	3,99	3,95 ± 0,010	3,98 ± 0,038	3,99 ± 0,037	84	2,91	3,08 ± 0,012	3,11 ± 0,092	3,14 ± 0,044
17	3,21	3,29 ± 0,012	3,28 ± 0,052	3,26 ± 0,040	51	3,13	3,27 ± 0,007	3,29 ± 0,056	3,27 ± 0,049	85	3,41	3,44 ± 0,012	3,44 ± 0,069	3,38 ± 0,057
18	3,27	3,32 ± 0,014	3,32 ± 0,082	3,28 ± 0,069	52	3,26	3,38 ± 0,011	3,37 ± 0,052	3,32 ± 0,060	86	3,31	3,48 ± 0,006	3,38 ± 0,112	3,40 ± 0,087
19	2,81	2,85 ± 0,018	2,95 ± 0,216	3,06 ± 0,088	53	4,03	4,16 ± 0,013	4,23 ± 0,047	4,23 ± 0,049	87	3,58	3,66 ± 0,013	3,67 ± 0,079	3,63 ± 0,072
20	3,61	3,53 ± 0,006	3,50 ± 0,052	3,47 ± 0,040	54	4,2	4,48 ± 0,016	4,51 ± 0,068	4,53 ± 0,062	88	4,89	5,02 ± 0,024	4,73 ± 0,166	4,67 ± 0,088
21	4,25	3,96 ± 0,009	3,98 ± 0,058	4,01 ± 0,031	55	3,66	3,65 ± 0,006	3,63 ± 0,058	3,60 ± 0,039	89	3,63	3,80 ± 0,008	3,81 ± 0,067	3,80 ± 0,041
22	4,32	4,29 ± 0,018	4,33 ± 0,081	4,32 ± 0,118	56	3,86	3,76 ± 0,008	3,76 ± 0,059	3,73 ± 0,036	90	4,2	4,34 ± 0,016	4,38 ± 0,124	4,39 ± 0,079
23	3,42	3,44 ± 0,006	3,42 ± 0,052	3,39 ± 0,046	57	3,81	3,87 ± 0,010	3,89 ± 0,043	3,89 ± 0,036	91	4,04	4,09 ± 0,016	4,14 ± 0,111	4,21 ± 0,101
24	3,99	3,91 ± 0,010	3,95 ± 0,050	3,95 ± 0,042	58	3,92	3,94 ± 0,012	3,96 ± 0,100	4,00 ± 0,097	92	4,78	4,68 ± 0,020	4,64 ± 0,080	4,61 ± 0,052
25	4,76	4,72 ± 0,022	4,64 ± 0,079	4,62 ± 0,067	59	3,66	3,60 ± 0,008	3,57 ± 0,060	3,55 ± 0,039	93	4,16	4,14 ± 0,013	4,21 ± 0,047	4,23 ± 0,038
26	3,96	3,96 ± 0,010	3,98 ± 0,041	4,00 ± 0,041	60	3,87	3,85 ± 0,009	3,86 ± 0,061	3,87 ± 0,045	94	3,39	3,61 ± 0,011	3,59 ± 0,174	3,59 ± 0,110
27	3,82	3,77 ± 0,008	3,76 ± 0,054	3,75 ± 0,036	61	3,78	3,82 ± 0,009	3,80 ± 0,068	3,81 ± 0,049	95	3,68	3,61 ± 0,006	3,56 ± 0,058	3,56 ± 0,038
28	4,36	4,21 ± 0,013	4,22 ± 0,090	4,20 ± 0,079	62	4,47	4,43 ± 0,186	3,75 ± 1,142	4,21 ± 0,569	96	4,35	4,38 ± 0,017	4,45 ± 0,057	4,45 ± 0,058
29	3,19	2,94 ± 0,018	2,92 ± 0,306	3,03 ± 0,059	63	4,5	4,17 ± 0,012	4,16 ± 0,067	4,24 ± 0,061	97	4,29	4,19 ± 0,013	4,26 ± 0,063	4,28 ± 0,038
30	4,64	4,41 ± 0,015	4,45 ± 0,051	4,47 ± 0,051	64	4,86	4,75 ± 0,020	4,65 ± 0,080	4,63 ± 0,060	98	4,58	4,59 ± 0,020	4,59 ± 0,091	4,58 ± 0,068
31	4,04	4,08 ± 0,013	4,14 ± 0,051	4,14 ± 0,047	65	4,03	3,90 ± 0,009	3,92 ± 0,063	3,94 ± 0,027	99	3,85	3,83 ± 0,008	3,80 ± 0,068	3,82 ± 0,065
32	3,34	3,37 ± 0,006	3,35 ± 0,076	3,32 ± 0,028	66	2,99	3,03 ± 0,012	3,05 ± 0,122	3,10 ± 0,071	100	4,62	4,61 ± 0,020	4,61 ± 0,090	4,59 ± 0,046
33	3,23	3,19 ± 0,009	3,16 ± 0,102	3,19 ± 0,050	67	4,61	5,02 ± 0,024	4,86 ± 0,144	4,69 ± 0,059					
34	4,04	4,00 ± 0,011	4,04 ± 0,040	4,05 ± 0,029	68	3,85	3,83 ± 0,009	3,84 ± 0,054	3,81 ± 0,041					

Tabla 4.5: Valores de pH reales medidos y estimados para la variedad de naranja *Bam* utilizando los métodos de regresión SVM, MLP y ANN-ICA.

Blood #	true pH	SVM pH	MLP pH	ANN-ICA pH	Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH	Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH
1	3,88	3,79 ± 0,013	3,77 ± 0,051	3,76 ± 0,061	35	4,23	4,17 ± 0,018	4,23 ± 0,056	4,22 ± 0,058	69	3,81	3,87 ± 0,011	3,84 ± 0,035	3,83 ± 0,053
2	3,25	3,15 ± 0,022	3,21 ± 0,105	3,29 ± 0,068	36	4,01	4,00 ± 0,014	4,03 ± 0,057	4,05 ± 0,037	70	3,91	3,88 ± 0,014	3,89 ± 0,043	3,89 ± 0,031
3	3,81	3,70 ± 0,011	3,68 ± 0,047	3,65 ± 0,034	37	4,12	4,22 ± 0,011	4,21 ± 0,059	4,22 ± 0,036	71	3,54	3,64 ± 0,010	3,62 ± 0,043	3,61 ± 0,047
4	3,24	3,44 ± 0,012	3,41 ± 0,045	3,41 ± 0,040	38	3,18	3,36 ± 0,015	3,34 ± 0,055	3,35 ± 0,050	72	3,77	3,98 ± 0,013	4,00 ± 0,257	3,94 ± 0,071
5	4,26	4,34 ± 0,013	4,35 ± 0,062	4,34 ± 0,043	39	3,56	3,70 ± 0,014	3,69 ± 0,040	3,65 ± 0,040	73	3,39	3,38 ± 0,013	3,37 ± 0,061	3,35 ± 0,060
6	4,22	4,64 ± 0,017	4,56 ± 0,136	4,50 ± 0,076	40	3,57	3,66 ± 0,009	3,63 ± 0,075	3,62 ± 0,045	74	3,78	3,70 ± 0,011	3,67 ± 0,067	3,67 ± 0,044
7	3,62	3,68 ± 0,011	3,66 ± 0,062	3,64 ± 0,046	41	3,82	3,86 ± 0,013	3,87 ± 0,042	3,87 ± 0,055	75	4,59	4,49 ± 0,028	4,38 ± 0,092	4,41 ± 0,072
8	4,79	4,66 ± 0,025	4,59 ± 0,147	4,52 ± 0,102	42	3,5	3,54 ± 0,013	3,49 ± 0,055	3,51 ± 0,050	76	4,14	4,35 ± 0,012	4,37 ± 0,051	4,35 ± 0,053
9	4,72	4,65 ± 0,026	4,61 ± 0,126	4,55 ± 0,083	43	3,48	3,69 ± 0,009	3,68 ± 0,077	3,67 ± 0,071	77	3,3	3,40 ± 0,014	3,36 ± 0,055	3,37 ± 0,049
10	3,26	3,53 ± 0,016	3,55 ± 0,050	3,53 ± 0,046	44	3,65	3,74 ± 0,011	3,73 ± 0,039	3,69 ± 0,043	78	3,7	3,72 ± 0,013	3,71 ± 0,035	3,68 ± 0,047
11	4,56	5,09 ± 0,022	4,68 ± 0,351	4,62 ± 0,062	45	3,67	3,42 ± 0,015	3,43 ± 0,054	3,43 ± 0,036	79	3,82	3,74 ± 0,012	3,73 ± 0,044	3,72 ± 0,039
12	3,47	3,61 ± 0,009	3,60 ± 0,070	3,55 ± 0,055	46	4,38	4,33 ± 0,018	4,33 ± 0,058	4,35 ± 0,042	80	3,68	3,66 ± 0,013	3,65 ± 0,044	3,63 ± 0,045
13	3,56	3,45 ± 0,011	3,43 ± 0,066	3,40 ± 0,057	47	3,34	3,31 ± 0,015	3,33 ± 0,104	3,32 ± 0,052	81	3,15	3,35 ± 0,020	3,40 ± 0,077	3,38 ± 0,057
14	4,34	4,37 ± 0,021	4,37 ± 0,100	4,38 ± 0,092	48	3,26	3,52 ± 0,012	3,50 ± 0,053	3,47 ± 0,036	82	3,39	3,37 ± 0,014	3,34 ± 0,044	3,36 ± 0,066
15	3,81	3,73 ± 0,012	3,71 ± 0,054	3,69 ± 0,031	49	3,83	3,85 ± 0,012	3,88 ± 0,046	3,87 ± 0,044	83	4,01	4,09 ± 0,011	4,15 ± 0,175	4,12 ± 0,062
16	3,72	3,47 ± 0,015	3,46 ± 0,049	3,47 ± 0,055	50	3,92	3,88 ± 0,011	3,85 ± 0,042	3,86 ± 0,037	84	4,03	3,74 ± 0,011	3,71 ± 0,042	3,71 ± 0,033
17	3,74	3,62 ± 0,012	3,60 ± 0,051	3,58 ± 0,031	51	3,62	3,72 ± 0,011	3,71 ± 0,044	3,68 ± 0,035	85	3,97	3,83 ± 0,012	3,84 ± 0,050	3,82 ± 0,035
18	2,93	3,08 ± 0,026	3,24 ± 0,100	3,29 ± 0,072	52	3,63	3,56 ± 0,010	3,51 ± 0,054	3,51 ± 0,037	86	3,9	3,84 ± 0,008	3,79 ± 0,147	3,77 ± 0,100
19	3,93	3,78 ± 0,012	3,78 ± 0,056	3,75 ± 0,042	53	3,62	3,90 ± 0,010	3,89 ± 0,042	3,89 ± 0,037	87	3,46	3,45 ± 0,015	3,46 ± 0,058	3,44 ± 0,037
20	3,42	3,37 ± 0,020	3,39 ± 0,054	3,40 ± 0,063	54	3,75	3,84 ± 0,011	3,86 ± 0,230	3,85 ± 0,050	88	3,27	3,27 ± 0,015	3,25 ± 0,091	3,31 ± 0,058
21	3,94	3,84 ± 0,011	3,81 ± 0,044	3,81 ± 0,046	55	3,89	4,09 ± 0,010	4,07 ± 0,054	4,10 ± 0,055	89	3,35	3,14 ± 0,021	3,18 ± 0,114	3,28 ± 0,076
22	3,57	3,42 ± 0,016	3,42 ± 0,056	3,43 ± 0,038	56	3,56	3,80 ± 0,010	3,81 ± 0,037	3,79 ± 0,052	90	3,3	3,36 ± 0,020	3,38 ± 0,057	3,40 ± 0,045
23	3,91	3,95 ± 0,011	3,92 ± 0,042	3,94 ± 0,041	57	3,86	3,79 ± 0,012	3,78 ± 0,043	3,74 ± 0,053	91	3,31	3,23 ± 0,014	3,25 ± 0,065	3,25 ± 0,098
24	3,28	3,30 ± 0,015	3,29 ± 0,068	3,33 ± 0,058	58	3,36	3,53 ± 0,014	3,54 ± 0,049	3,51 ± 0,031	92	3,57	3,39 ± 0,012	3,38 ± 0,063	3,36 ± 0,063
25	3,71	3,71 ± 0,012	3,70 ± 0,037	3,67 ± 0,033	59	4,11	4,16 ± 0,014	4,13 ± 0,058	4,16 ± 0,049	93	3,91	3,84 ± 0,012	3,85 ± 0,036	3,84 ± 0,038
26	4,47	4,39 ± 0,016	4,37 ± 0,102	4,39 ± 0,068	60	4,46	4,31 ± 0,019	4,27 ± 0,123	4,34 ± 0,082	94	3,07	3,13 ± 0,020	3,25 ± 0,131	3,25 ± 0,091
27	4,5	4,59 ± 0,015	4,62 ± 0,116	4,49 ± 0,090	61	4,09	3,94 ± 0,014	3,98 ± 0,080	3,96 ± 0,051	95	3,57	3,71 ± 0,012	3,70 ± 0,043	3,66 ± 0,035
28	4,38	4,22 ± 0,019	4,25 ± 0,060	4,26 ± 0,064	62	3,84	3,98 ± 0,011	4,02 ± 0,046	4,02 ± 0,067	96	3,72	3,10 ± 0,100	3,84 ± 1,006	3,57 ± 0,504
29	4,58	4,37 ± 0,017	4,36 ± 0,066	4,33 ± 0,054	63	4	3,86 ± 0,011	3,83 ± 0,064	3,83 ± 0,041	97	3,97	4,08 ± 0,010	4,05 ± 0,051	4,06 ± 0,056
30	3,83	3,85 ± 0,012	3,88 ± 0,051	3,86 ± 0,048	64	3,83	3,83 ± 0,010	3,81 ± 0,039	3,79 ± 0,041	98	4,33	4,36 ± 0,020	4,37 ± 0,073	4,37 ± 0,088
31	4,37	4,30 ± 0,017	4,30 ± 0,063	4,31 ± 0,059	65	4,11	4,07 ± 0,012	4,02 ± 0,054	4,07 ± 0,049	99	3,29	3,31 ± 0,016	3,28 ± 0,061	3,32 ± 0,057
32	4,39	4,32 ± 0,023	4,34 ± 0,072	4,33 ± 0,106	66	4,43	4,46 ± 0,025	4,40 ± 0,104	4,38 ± 0,087	100	3,54	3,52 ± 0,013	3,48 ± 0,060	3,49 ± 0,046
33	3,16	3,26 ± 0,021	3,31 ± 0,082	3,34 ± 0,058	67	3,91	3,94 ± 0,013	3,91 ± 0,046	3,93 ± 0,040					
34	3,95	3,73 ± 0,017	3,96 ± 0,997	4,14 ± 0,453	68	4,36	4,28 ± 0,015	4,27 ± 0,059	4,25 ± 0,049					

Tabla 4.6: Valores de pH reales medidos y estimados para la variedad de naranja *Blood* utilizando los métodos de regresión SVM, MLP y ANN-ICA.

Thomson #	true pH	SVM pH	MLP pH	ANN-ICA pH	Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH	Bam #	true pH	SVM pH	MLP pH	ANN-ICA pH
1	3,95	3,93 ± 0,021	3,93 ± 0,060	3,96 ± 0,058	35	5,54	5,48 ± 0,026	5,49 ± 0,094	5,58 ± 0,096	69	5,33	5,44 ± 0,021	5,42 ± 0,112	5,50 ± 0,092
2	4,26	4,06 ± 0,021	4,09 ± 0,084	4,04 ± 0,078	36	3,79	3,92 ± 0,018	3,88 ± 0,083	3,92 ± 0,066	70	4,93	4,91 ± 0,018	4,93 ± 0,082	4,92 ± 0,086
3	4,24	4,20 ± 0,012	4,19 ± 0,109	4,14 ± 0,090	37	5,61	5,39 ± 0,024	5,44 ± 0,163	5,40 ± 0,113	71	5,00	5,01 ± 0,020	5,04 ± 0,080	5,05 ± 0,093
4	3,75	4,58 ± 0,016	4,52 ± 0,088	4,54 ± 0,094	38	5,17	5,17 ± 0,024	5,15 ± 0,168	5,17 ± 0,167	72	3,83	3,83 ± 0,027	3,81 ± 0,128	3,93 ± 0,095
5	4,46	4,16 ± 0,019	4,12 ± 0,096	4,08 ± 0,112	39	4,49	4,50 ± 0,013	4,31 ± 0,143	4,36 ± 0,076	73	5,18	5,31 ± 0,018	5,32 ± 0,073	5,39 ± 0,056
6	4,14	4,10 ± 0,014	4,10 ± 0,087	4,06 ± 0,111	40	3,69	3,54 ± 0,032	3,76 ± 0,269	3,84 ± 0,104	74	4,62	4,54 ± 0,010	4,52 ± 0,072	4,51 ± 0,092
7	3,96	4,54 ± 0,018	4,43 ± 0,164	4,40 ± 0,100	41	4,98	4,93 ± 0,012	4,95 ± 0,090	4,95 ± 0,096	75	5,17	5,08 ± 0,016	5,09 ± 0,072	5,10 ± 0,072
8	4,43	4,52 ± 0,015	4,35 ± 0,160	4,40 ± 0,071	42	4,74	4,84 ± 0,024	4,57 ± 0,260	4,66 ± 0,101	76	4,05	4,15 ± 0,010	4,19 ± 0,102	4,13 ± 0,073
9	4,42	3,76 ± 0,017	3,84 ± 0,103	3,86 ± 0,079	43	4,19	4,27 ± 0,017	4,25 ± 0,097	4,17 ± 0,100	77	4,46	4,37 ± 0,012	4,34 ± 0,175	4,30 ± 0,103
10	3,72	3,79 ± 0,017	3,98 ± 0,125	3,96 ± 0,087	44	4,06	4,13 ± 0,020	4,14 ± 0,105	4,09 ± 0,098	78	3,61	3,80 ± 0,022	3,82 ± 0,085	3,91 ± 0,099
11	5,65	5,85 ± 0,037	5,81 ± 0,351	5,74 ± 0,177	45	3,88	3,94 ± 0,012	4,10 ± 0,113	4,01 ± 0,098	79	5,01	4,81 ± 0,012	4,85 ± 0,048	4,81 ± 0,063
12	4,37	4,08 ± 0,021	4,07 ± 0,063	4,06 ± 0,092	46	4,33	4,25 ± 0,015	4,16 ± 0,127	4,18 ± 0,097	80	5,71	5,83 ± 0,028	5,88 ± 0,191	5,84 ± 0,068
13	4,32	4,18 ± 0,016	4,20 ± 0,060	4,16 ± 0,045	47	4,12	4,05 ± 0,016	4,07 ± 0,133	4,05 ± 0,099	81	4,33	4,51 ± 0,011	4,49 ± 0,067	4,49 ± 0,073
14	5,21	5,00 ± 0,011	5,04 ± 0,044	5,02 ± 0,067	48	4,41	4,31 ± 0,012	4,28 ± 0,080	4,25 ± 0,074	82	3,66	3,67 ± 0,023	3,69 ± 0,146	3,80 ± 0,093
15	5,02	5,14 ± 0,016	5,16 ± 0,073	5,21 ± 0,086	49	4,24	4,21 ± 0,011	4,19 ± 0,119	4,19 ± 0,140	83	4,21	4,27 ± 0,018	4,27 ± 0,054	4,22 ± 0,056
16	3,84	3,82 ± 0,020	3,81 ± 0,108	3,88 ± 0,076	50	4,96	4,89 ± 0,014	4,95 ± 0,056	4,90 ± 0,038	84	3,61	3,73 ± 0,020	3,79 ± 0,103	3,88 ± 0,081
17	3,98	4,04 ± 0,019	4,03 ± 0,058	4,03 ± 0,060	51	3,89	3,99 ± 0,018	4,01 ± 0,061	3,98 ± 0,069	85	4,23	4,24 ± 0,016	4,25 ± 0,070	4,19 ± 0,065
18	4,06	4,11 ± 0,022	4,10 ± 0,092	4,07 ± 0,095	52	4,05	4,18 ± 0,014	4,17 ± 0,067	4,16 ± 0,058	86	4,10	4,31 ± 0,021	4,29 ± 0,099	4,28 ± 0,092
19	3,48	3,52 ± 0,022	3,85 ± 0,178	3,88 ± 0,095	53	5,00	5,18 ± 0,015	5,21 ± 0,060	5,26 ± 0,093	87	4,45	4,52 ± 0,011	4,54 ± 0,072	4,50 ± 0,061
20	4,48	4,34 ± 0,014	4,34 ± 0,057	4,30 ± 0,050	54	5,21	5,55 ± 0,025	5,60 ± 0,346	5,53 ± 0,173	88	6,31	6,16 ± 0,029	5,96 ± 0,319	5,91 ± 0,114
21	5,27	4,92 ± 0,013	4,94 ± 0,054	4,94 ± 0,064	55	4,54	4,51 ± 0,013	4,50 ± 0,057	4,51 ± 0,066	89	4,50	4,70 ± 0,009	4,73 ± 0,053	4,67 ± 0,077
22	5,35	5,35 ± 0,019	5,36 ± 0,076	5,42 ± 0,075	56	4,79	4,74 ± 0,019	4,45 ± 0,318	4,59 ± 0,153	90	5,21	5,39 ± 0,019	5,39 ± 0,084	5,45 ± 0,082
23	4,24	4,23 ± 0,014	4,21 ± 0,058	4,17 ± 0,065	57	4,72	4,80 ± 0,012	4,82 ± 0,055	4,81 ± 0,075	91	5,01	5,10 ± 0,016	5,14 ± 0,057	5,16 ± 0,067
24	4,96	4,84 ± 0,013	4,89 ± 0,051	4,88 ± 0,051	58	4,86	4,90 ± 0,011	4,95 ± 0,053	4,95 ± 0,047	92	5,93	5,82 ± 0,027	5,84 ± 0,163	5,82 ± 0,074
25	6,26	5,85 ± 0,023	5,84 ± 0,145	5,81 ± 0,079	59	4,53	4,51 ± 0,013	4,35 ± 0,136	4,39 ± 0,063	93	5,16	5,15 ± 0,014	5,18 ± 0,059	5,20 ± 0,060
26	4,91	4,90 ± 0,017	4,96 ± 0,062	4,89 ± 0,071	60	4,80	4,79 ± 0,010	4,80 ± 0,056	4,81 ± 0,058	94	4,20	4,47 ± 0,018	4,46 ± 0,092	4,48 ± 0,082
27	4,73	4,65 ± 0,014	4,67 ± 0,065	4,66 ± 0,062	61	4,69	4,72 ± 0,011	4,74 ± 0,050	4,72 ± 0,054	95	4,57	4,44 ± 0,016	4,45 ± 0,080	4,43 ± 0,074
28	5,41	5,32 ± 0,030	5,36 ± 0,169	5,39 ± 0,169	62	5,54	5,56 ± 0,021	5,55 ± 0,106	5,60 ± 0,059	96	5,40	5,44 ± 0,022	5,44 ± 0,078	5,54 ± 0,085
29	3,95	3,64 ± 0,018	3,81 ± 0,158	3,85 ± 0,094	63	5,58	5,18 ± 0,015	5,17 ± 0,050	5,22 ± 0,053	97	5,33	5,21 ± 0,015	5,22 ± 0,065	5,26 ± 0,074
30	5,76	5,57 ± 0,026	5,63 ± 0,223	5,67 ± 0,169	64	6,23	5,90 ± 0,025	5,88 ± 0,174	5,86 ± 0,089	98	5,68	5,68 ± 0,024	5,70 ± 0,087	5,73 ± 0,075
31	5,01	5,16 ± 0,025	5,17 ± 0,146	5,14 ± 0,138	65	5,00	4,85 ± 0,011	4,86 ± 0,056	4,86 ± 0,070	99	4,78	4,74 ± 0,017	4,77 ± 0,056	4,74 ± 0,070
32	4,14	4,14 ± 0,015	4,12 ± 0,060	4,12 ± 0,034	66	3,71	3,68 ± 0,032	3,67 ± 0,168	3,87 ± 0,113	100	5,73	5,72 ± 0,025	5,73 ± 0,109	5,77 ± 0,068
33	4,00	3,97 ± 0,013	4,10 ± 0,149	4,02 ± 0,120	67	5,72	6,22 ± 0,033	6,27 ± 0,298	6,00 ± 0,139					
34	5,01	4,95 ± 0,013	5,00 ± 0,053	4,99 ± 0,045	68	4,77	4,75 ± 0,017	4,77 ± 0,059	4,77 ± 0,063					

Tabla 4.7: Valores de pH reales medidos y estimados para la variedad de naranja *Thomson* utilizando los métodos de regresión SVM, MLP y ANN-ICA.

5

Conclusiones y lineas futuras

5.1. Conclusiones

La investigación realizada en el presente proyecto ha permitido obtener unos determinados resultados a partir de dos lineas de análisis distintos: por un lado, el empleo de diferentes métodos de regresión no intrusivos utilizados para estimar el grado de acidez (pH) de las naranjas que conforman la base de datos propuesta, y, por otro lado, el análisis de las distintas variedades de naranjas disponibles (*Bam*, *Blood* y *Thomson*) ha permitido deducir cuáles de estas variedades ofrecen mejores resultados a la hora de estimar el grado de acidez de las muestras de cada una de ellas.

En este capítulo se expondrán las pertinentes conclusiones a partir del conjunto de resultados obtenidos en el presente proyecto.

Métodos de regresión

Para realizar la estimación del grado de acidez de las muestras de naranjas disponibles en la base de datos propuesta se han seleccionado tres mecanismos de regresión no intrusivos distintos: SVM, MLP y ANN-ICA.

- El análisis de los métodos de regresión propuestos ha sido realizado en base a dos criterios: los coeficientes de error propuestos y el error en la estimación del pH definido a través de la ecuación (4.7). Para ambos criterios se puede observar que los boxplots representados para los cuatro grupos de narajas estudiados (*Bam*, *Blood*, *Thomson* y *3-Variety*) son muy similares entre sí, pero con una diferencia significativa: el grado de dispersión del error representado por los outliers. Para la red neuronal MLP la cantidad de outliers representados es muy alta en comparación con los otros dos métodos de regresión propuestos, esto implica que la dispersión del error (desviación típica) es más elevada y, por tanto, el error cometido en las

diferentes iteraciones realizadas para la red es más elevado, lo que provocará que la estimación del pH sea incorrecta.

- Respecto a los métodos de regresión SVM y ANN-ICA los resultados obtenidos en términos de errores cometidos son muy similares entre sí, lo que implicará que la estimación del pH también será adecuada.

- En cuanto a la carga computacional y al tiempo de simulación de cada uno de los métodos de regresión propuestos, SVM y MLP poseen unos tiempos de simulación muy similares y reducidos en comparación al método de regresión ANN-ICA. Esto se debe, en gran medida a la función de entrenamiento utilizada para cada una de los métodos, en concreto, la función de entrenamiento para ANN-ICA es el propio algoritmo ICA, que posee una carga computacional muy elevada.

Variedades de naranja

El análisis de las distintas variedades de naranjas ha sido dividido en los tres grupos principales que forman la base de datos: *Bam*, *Blood* y *Thomson*, más un conjunto adicional formado por la combinación de los tres grupos anteriores (*3-Variety*).

- El conjunto combinación de variedades *3-Variety* analizado a partir de los coeficientes de error propuestos ofrece unos resultados significativamente peores para los métodos de regresión estudiados en la sección de resultados 4.2 en comparación al análisis de variedades de forma individual debido a que las muestras de distintas variedades de naranjas pueden tener el mismo valor del grado de acidez (pH) para características totalmente diferentes, esto implica que a la hora de realizar una estimación del pH se produzcan un mayor número de errores.

- Para las variedades de naranjas analizadas individualmente los resultados obtenidos tanto en la estimación como en el análisis de errores son correctos para los tres métodos de regresión SVM, MLP y ANN-ICA ya que el coeficiente de correlación lineal (R) y de determinación (R^2) es cercano a 1, y el resto de coeficientes de error se sitúan cercanos a cero, mientras que para el error en la estimación en el pH, para los métodos SVM y ANN-ICA la mayoría de las muestras representadas por boxplots sitúan su media en torno a cero y su dispersión en el error en un rango de [-0.5 0.5] lo que implica que el error cometido en la estimación para el conjunto de simulaciones realizadas no es muy elevado, en cambio la red neuronal MLP introduce mayor dispersión en los valores estimados para el conjunto de experimentos realizados, por lo que el error en la estimación será más elevado. Existen excepciones para determinadas muestras de cada variedad: para la variedad

Bam, las muestras 40 y 62 producen un error estimado muy elevado en comparación al resto de muestras disponibles; para la variedad *Blood*, las muestras 34 y 96; y para la variedad *Thomson* la muestra 4. Estos errores en la estimación del pH para estas muestras en concreto se deben a que las imágenes disponibles poseen defectos en el momento de su captura, como pueden ser movimiento de las muestras o brillos, o también errores a la hora de realizar la segmentación de la imagen, lo que hace que las características obtenidas sobre cada una de ellas no sean adecuadas para el valor de pH real medido.

- Atendiendo a los valores estimados del grado de acidez para cada una de las variedades de naranjas propuestas, la variedad de naranja *Thomson* ofrece los mejores resultados en términos del coeficiente de correlación lineal R ya que su valor es el más cercano a 1 (R = 0.95) para los métodos de regresión utilizados en la sección 4.3. La variedad de naranja *Bam* ofrece, igualmente, buenos resultados en la estimación (R = 0.94) para los métodos SVM y ANN-ICA, pero un valor más reducido (R = 0.93) para la red neuronal MLP; mientras que los resultados del coeficiente de correlación lineal R para la variedad *Blood* son de 0.92 para el método de regresión SVM y de 0.94 para MLP y ANN-ICA. Esto implica que la estimación del pH para las dos últimas variedades ofrece notables variaciones en función del método de regresión utilizado, por lo que no está claro cuál de las variedades ofrece mejores resultados en la estimación del grado de acidez.

- Resumiendo, la variedad de naranja *Thomson* ofrece unos resultados en la estimación del grado de acidez mejores que el resto de variedades que conforman la base de datos en términos del coeficiente de correlación lineal R. Además el análisis de las variedades de naranjas de forma individual frente al conjunto combinación de variedades (*3-Variety*) proporciona unos resultados mucho más eficientes en relación con los coeficientes de error propuestos en la sección 4.1.1.

5.2. Lineas futuras

- **Ampliación BBDD**. La base de datos disponible en el presente proyecto dispone de 300 imágenes, 100 por cada variedad de naranja estudiada. La ampliación del número de imágenes, así como una mejora en la calidad en la captura de las mismas, puede resultar de gran ayuda a la hora de generalizar los resultados de la estimación del pH para las diferentes variedades de naranjas disponibles.

- **Conjunto de características definidas**. La extracción de un gran número de características de cada imagen de la base de datos proporciona una definición más

precisa de cada una de las muestras disponibles. La extracción de nuevos grupos de características, como los momentos de Zernike o los coeficientes Wavelet pueden ayudar a mejorar la descripción de estas muestras.

- **Empleo de algoritmos de selección de características**. La etapa de reducción de la dimensionalidad de la matriz de características $\times$ muestras es una de las etapas más importantes a la hora de garantizar un funcionamiento correcto y efectivo de los métodos de regresión propuestos. La utilización de algoritmos capaces de seleccionar las características más discriminantes de todo el conjunto propuesto puede ser una buena forma de incrementar el rendimiento de los diferentes mecanismos de regresión no intrusivos estudiados.

- **Utilización de redes convolucionales (*Convolutional Neural Network*, CNN)**. Los métodos de regresión propuestos en el presente proyecto trabajan a partir de matrices de características reducidas obtenidas a partir de las imágenes que componen la base de datos. La utilización de redes convolucionales permitiría la obtención del pH estimado a partir del procesamiento de la imagen directamente, eliminando las etapas de extracción de características y reducción de la dimensionalidad.

Agradecimientos

En primer lugar, agradecer a D. Sajab Sabzi por proporcionar la base de datos de naranjas para realizar el presente proyecto y parte del código de MATLAB utilizado, y por su ayuda con las dudas que me han ido surgiendo a lo largo de estos meses de trabajo.

Por otro lado, expreso mi más sincero agradecimiento a D. Juan Ignacio Arribas Sánchez por la tutela tanto de este Trabajo de Fin de Máster como del Trabajo de Fin de Grado realizado anteriormente, destacando su disponibilidad y su capacidad para resolver todos los problemas que han ido apareciendo en este periodo de tiempo de dos años que hemos compartido. Muchas gracias por todo.

Fernando García Gómez.

Bibliografía

[1] Jing Hu, Daoliang Li, Qingling Duan, Yueqi Han, Guifen Chen, and Xiuli Si. Fish species classification by color, texture and multi-class support vector machine using computer vision. *Computers and electronics in agriculture*, 88:133–140, 2012.

[2] Meijun Sun, Dong Zhang, Li Liu, and Zheng Wang. How to predict the sugariness and hardness of melons: A near-infrared hyperspectral imaging method. *Food chemistry*, 218:413–421, 2017.

[3] Yun Zhao, Yong He, and Xing Xu. A novel algorithm for damage recognition on pest-infested oilseed rape leaves. *Computers and electronics in agriculture*, 89:41–50, 2012.

[4] Slamet Riyadi, Mohd Marzuki Mustafa, Aini Hussain, and Azman Hamzah. Papaya fruit grading based on size using image analysis. In *Proc. International Conference on Electrical Engineering and Informatics*, pages 645–648, 2007.

[5] Akira Mizushima and Renfu Lu. An image segmentation method for apple sorting and grading using support vector machine and otsu's method. *Computers and electronics in agriculture*, 94:29–37, 2013.

[6] Sajad Sabzi, Yousef Abbaspour-Gilandeh, and Juan Ignacio Arribas. Non-intrusive image processing thompson orange grading methods. In *2017 56th FITCE Congress*, pages 35–39. IEEE, 2017.

[7] Sajad Sabzi, Yousef Abbaspour-Gilandeh, and Juan Ignacio Arribas. A new method based on computer vision for non-intrusive orange peel sorting. In *2017 56th FITCE Congress*, pages 16–19. IEEE, 2017.

[8] Sajad Sabzi and Juan Ignacio Arribas. A visible-range computer-vision system for automated, non-intrusive assessment of the ph value in thomson oranges. *Computers in Industry*, 99:69–82, 2018.

[9] Manuel Agustí. *Citricultura*. Number 634.3 A3. 2003.

[10] Pedro Miguel Chomé Fuster. Las variedades de cítricos: retos y soluciones. *Agricultura: Revista agropecuaria*, (957):828–832, 2012.

[11] S Zaragoza, A Pina Lorca, M Forner, L Navarro, A Medina, G Soler, and PM Chomé. Las variedades de cítricos. *El material vegetal y el registro de variedades comerciales de España. Spain. Ministerio de Medio Ambiente, Medio Rural y Marino*, 2011.

[12] J Soler Aznar. *Reconocimiento de variedades de cítricos en campo.* Generalitat Valenciana. Conselleria de Agricultura, Pesca y Alimentación, 1999.

[13] Mery Solis. Determinación de las propiedades físico-químicas del limón (citrus limus). B.S. thesis, 2006.

[14] Magaly Gómez Ugarte, Lesly Eliana Peralta Mamani, Aleyda Abigail Martinez Ayala, Dayana Lilian Paredes Fernandez, Yessica Conde Camacho, et al. Propiedades medicinales del pomelo, beneficios nutricionales y su aplicación en la estética. *Revista de Investigación e Información en Salud*, 10:49, 2015.

[15] James Saunt. *Variedades de cítricos del mundo: guia ilustrada.* Number 634.304 S3Y. Edipublic, 1991.

[16] Roger Gordon Bates et al. Determination of ph: theory and practice. *Determination of pH: theory and practice.*, 1964.

[17] Raymond Chang. Principios esenciales de química general, raymond chang. 2006.

[18] RP Buck, S Rondinini, AK Covington, FGK Baucke, Christopher MA Brett, MF Camoes, MJT Milton, T Mussini, Renate Naumann, KW Pratt, et al. Measurement of ph. definition, standards, and procedures (iupac recommendations 2002). *Pure and applied chemistry*, 74(11):2169–2200, 2002.

[19] S Karastogianni, S Girousi, and S Sotiropoulos. ph: Principles and measurement. *The Oxford: Academic Press. Encyclopedia of Food and Health*, pages 333–8, 2016.

[20] Enrique Vera López. uso de métodos electroquímicos como herramientas para evaluar parámetros de interfase en sistemas heterogéneos metal/medio acuoso. *Universidad Pedagógica y Tecnológica de Colombia UPTC, Tunja, Colombia*, 2010.

[21] Hossein Javadikia, Sajad Sabzi, and Hekmat Rabbani. Machine vision based expert system to estimate orange mass of three varieties. *International Journal of Agricultural and Biological Engineering*, 10(2):132–139, 2017.

[22] Sajad Sabzi, Yousef Abbaspour-Gilandeh, and Ginés García-Mateos. A new approach for visual identification of orange varieties using neural networks and metaheuristic algorithms. *Information processing in agriculture*, 5(1):162–172, 2018.

[23] G García-Mateos, JL Hernández-Hernández, D Escarabajal-Henarejos, S Jaen-Terrones, and JM Molina-Martínez. Study and comparison of color models for automatic image analysis in irrigation management applications. *Agricultural water management*, 151:158–166, 2015.

[24] Sajad Sabzi, Yousef Abbaspour-Gilandeh, Ginés García-Mateos, Antonio Ruiz-Canales, and José Molina-Martínez. Segmentation of apples in aerial images under sixteen different lighting conditions using color and texture for optimal irrigation. *Water*, 10(11):1634, 2018.

[25] Bencheriet Chemesse Ennehar, Oudjani Brahim, and Tebbikh Hicham. An appropriate color space to improve human skin detection. *INFOCOMP*, 9(4):1–10, 2010.

[26] Congcong Zhang, Xiaoyan Xiao, Xiaomei Li, Ying-Jie Chen, Wu Zhen, Jun Chang, Chengyun Zheng, and Zhi Liu. White blood cell segmentation by color-space-based k-means clustering. *Sensors*, 14(9):16128–16147, 2014.

[27] George H Joblove and Donald Greenberg. Color spaces for computer graphics. In *ACM siggraph computer graphics*, volume 12, pages 20–25. ACM, 1978.

[28] Khamar Basha Shaik, P Ganesan, V Kalist, BS Sathish, and J Merlin Mary Jenitha. Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, 57:41–48, 2015.

[29] Gurjinder Singh Sahdra and Kamaljot Singh Kailey. Detection of contaminants in cotton by using ydbdr color space. *Int. J. Computer Technology & Applications*, 3(3):1118–1124, 2012.

[30] Baisa L Gunjal and Suresh N Mali. Comparative performance analysis of dwt-svd based color image watermarking technique in yuv, rgb and yiq color spaces. *International Journal of Computer Theory and Engineering*, 3(6):714, 2011.

[31] Bangalore S Manjunath, J-R Ohm, Vinod V Vasudevan, and Akio Yamada. Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):703–715, 2001.

[32] RC Gonzalez. Digital image processing using matlab-gonzalez woods & eddins. pdf. education. 2004.

[33] David F Rogers. Procedural elements for computer graphics mcgraw-hill. *New York*, 1985.

[34] Dany Vijay D'Souza. *An fMRI study of chromatic processing in humans*. PhD thesis, Faculty of Biology, Georg-August-Universität Göttingen, 2009.

[35] John F Shroder. *Biological and Environmental Hazards, Risks, and Disasters*. Elsevier, 2015.

[36] David M Woebbecke, George E Meyer, K Von Bargen, and DA Mortensen. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995.

[37] GE Meyer, T Mehta, MF Kocher, DA Mortensen, and A Samal. Textural imaging and discriminant analysis for distinguishingweeds for spot spraying. *Transactions of the ASAE*, 41(4):1189, 1998.

[38] Takashi Kataoka, Toshihiro Kaneko, Hiroshi Okamoto, and S Hata. Crop growth estimation system using machine vision. In *Proceedings 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003)*, volume 2, pages b1079–b1083. IEEE, 2003.

[39] George E Meyer and João Camargo Neto. Verification of color vegetation indices for automated crop imaging applications. *Computers and electronics in agriculture*, 63(2):282–293, 2008.

[40] David M Woebbecke, George E Meyer, Kenneth Von Bargen, and David A Mortensen. Plant species identification, size, and enumeration using machine vision techniques on near-binary images. In *Optics in Agriculture and Forestry*, volume 1836, pages 208–220. International Society for Optics and Photonics, 1993.

[41] Mahmood R Golzarian and Ross A Frick. Classification of images of wheat, ryegrass and brome grass species at early growth stages using principal component analysis. *Plant Methods*, 7(1):28, 2011.

[42] Alaa Eleyan and Hasan Demirel. Co-occurrence matrix and its statistical features as a new approach for face recognition. *Turkish Journal of Electrical Engineering & Computer Sciences*, 19(1):97–107, 2011.

[43] ES Gadelmawla, AE Eladawi, OB Abouelatta, and IM Elewa. Investigation of the cutting conditions in milling operations using image texture features. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 222(11):1395–1404, 2008.

[44] Agus Eko Minarno, Yuda Munarko, Arrie Kurniawardhani, Fitri Bimantoro, and Nanik Suciati. Texture feature extraction using co-occurrence matrices of sub-band image for batik image classification. In *2014 2nd International Conference on Information and Communication Technology (ICoICT)*, pages 249–254. IEEE, 2014.

[45] David A Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of remote sensing*, 28(1):45–62, 2002.

[46] Michael A Wirth. Shape analysis and measurement. *University of Guelph. CIS*, 6320, 2001.

[47] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[48] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303, 2008.

[49] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[50] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[51] Vojislav Kecman. Support vector machines–an introduction. In *Support vector machines: theory and applications*, pages 1–47. Springer, 2005.

[52] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[53] Tania Kleynhans, Matthew Montanaro, Aaron Gerace, and Christopher Kanan. Predicting top-of-atmosphere thermal radiance using merra-2 atmospheric data with deep learning. *Remote Sensing*, 9(11):1133, 2017.

[54] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[55] Rodrigo Torán Sierra. Estudio de simulación con r para una mejor comprensión de los modelos multicapa. 2017.

[56] Esmaeil Atashpaz-Gargari and Caro Lucas. Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. In *2007 IEEE congress on evolutionary computation*, pages 4661–4667. IEEE, 2007.

[57] Arys Carrasquilla Batista, Alfonso Chacón Rodríguez, Kattia Núñez Montero, Olman Gómez Espinoza, Johnny Valverde Cerdas, and Maritza Guerrero Barrantes. Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Tecnología en Marcha*, 30(5):33–45, 2016.

[58] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[59] Kristin Potter, Hans Hagen, Andreas Kerren, and Peter Dannenmann. Methods for presenting statistical information: The box plot. *Visualization of large and unstructured data sets*, 4:97–106, 2006.

Apéndice A

Base de datos de las variedades de naranjas propuestas

En este apéndice se presentarán las diferentes muestras de imágenes segmentadas que conforman la base de datos propuesta en [22], además del conjunto de pH reales medidos para cada una de las variedades de naranjas estudiadas: *Bam*, *Blood* y *Thomson*. El banco de datos propuesto está formado por las siguientes imágenes segmentadas:

- Conjunto de 100 imágenes correspondientes a las 100 muestras recogidas de la variedad de naranja *Bam* perteneciente al grupo Blancas (Figura A.1).
- Conjunto de 100 imágenes correspondientes a las 100 muestras recogidas de la variedad de naranja *Blood* perteneciente al grupo Sangre (Figura A.2).
- Conjunto de 100 imágenes correspondientes a las 100 muestras recogidas de la variedad de naranja *Thomson* perteneciente al grupo Navel (Figura A.3).

A.1. Conjunto de naranjas de la variedad Bam

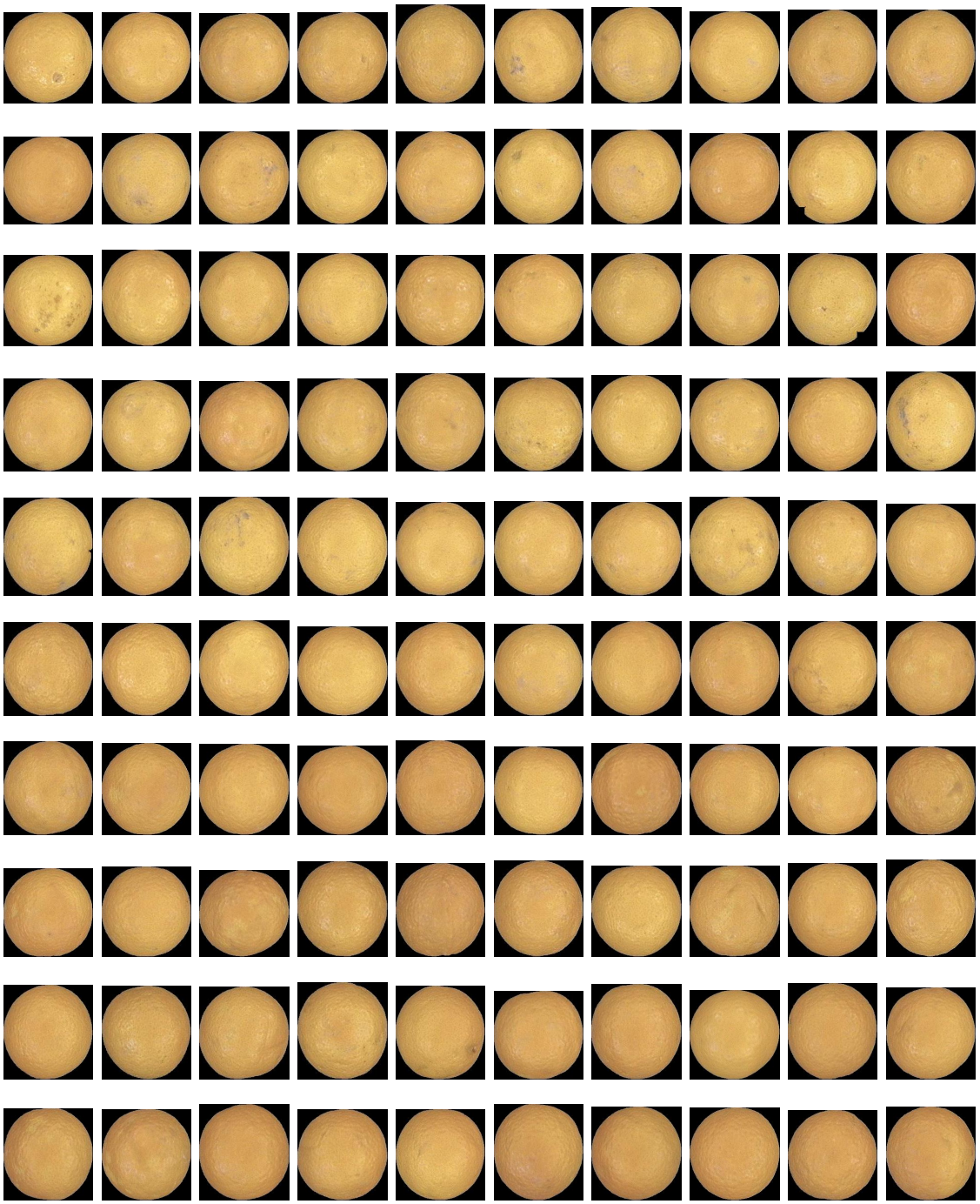


Figura A.1: De izquierda a derecha y de arriba hacia abajo. Conjunto de 100 muestras (1 a 100) de la variedad de naranja *Bam*.

A.2. Conjunto de naranjas de la variedad Blood

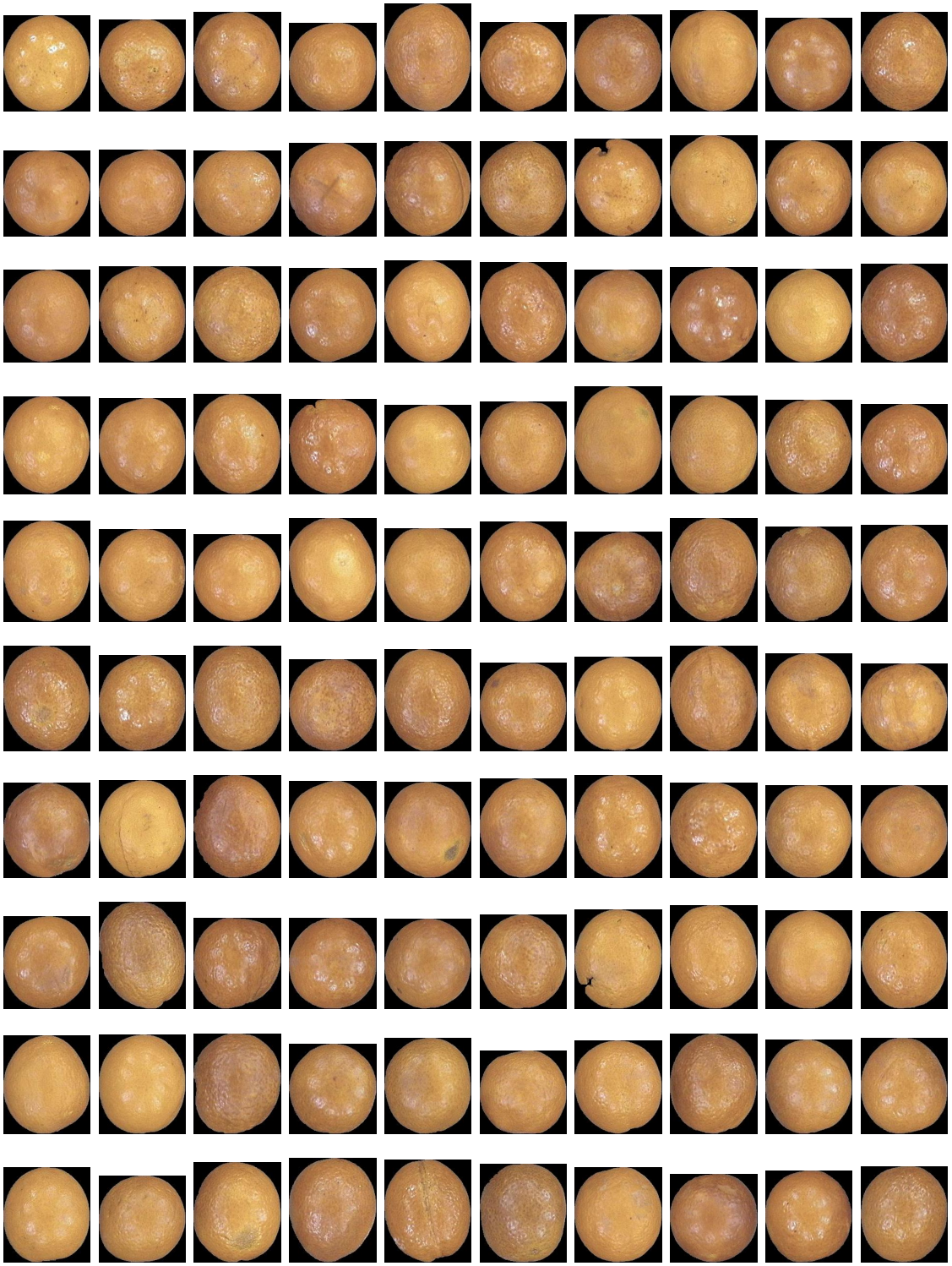


Figura A.2: De izquierda a derecha y de arriba hacia abajo. Conjunto de 100 muestras (1 a 100) de la variedad de naranja *Blood*.

A.3. Conjunto de naranjas de la variedad Thomson

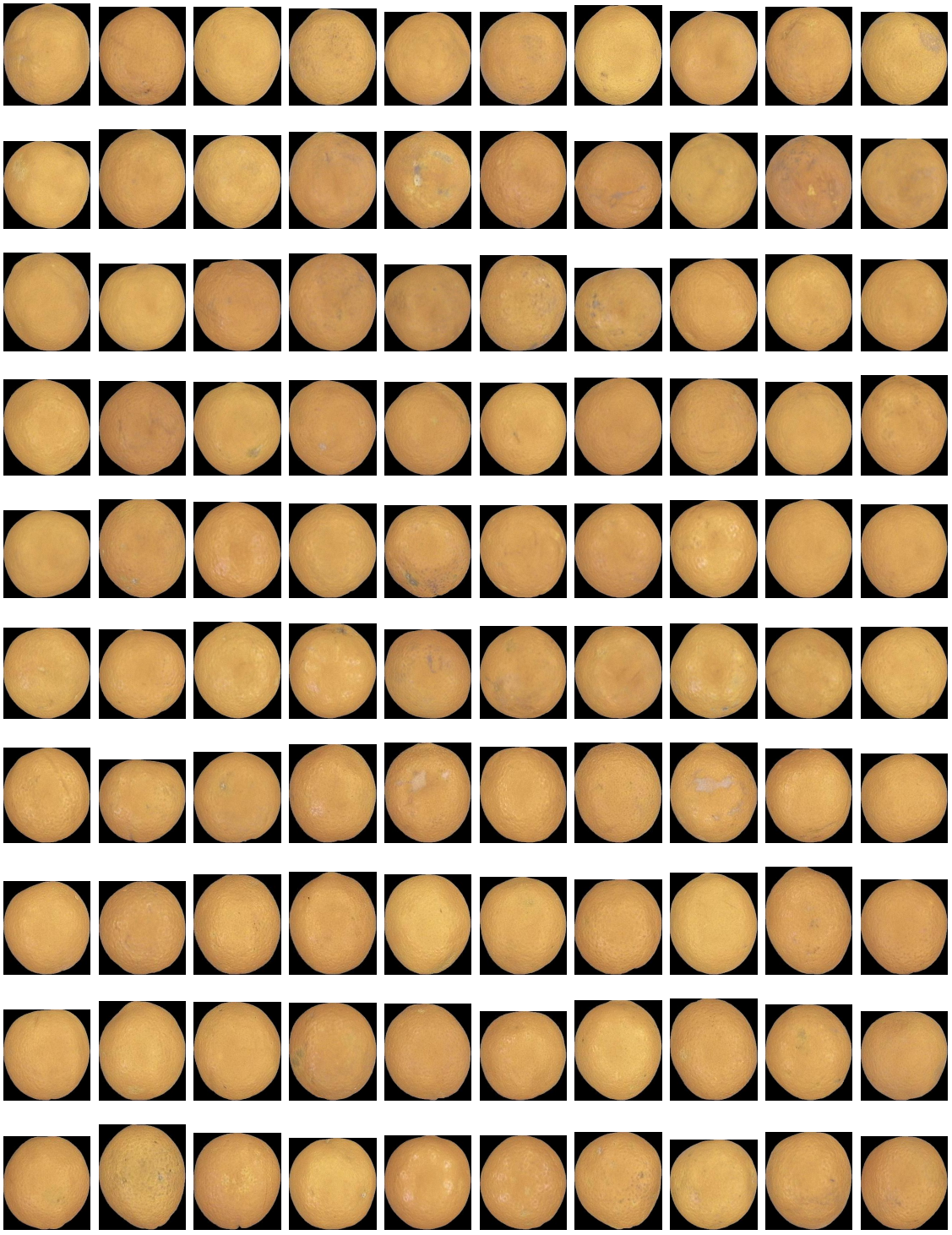


Figura A.3: De izquierda a derecha y de arriba hacia abajo. Conjunto de 100 muestras (1 a 100) de la variedad de naranja *Thomson*.

A.4. Valores de pH reales medidos para cada una de las variedades de naranjas

Orange #	Ph Bam	Ph Blood	Ph Thomson	Orange #	Ph Bam	Ph Blood	Ph Thomson
1	3.19	3.88	3.95	51	3.13	3.62	3.89
2	3.43	3.25	4.26	52	3.26	3.63	4.05
3	3.42	3.81	4.24	53	4.03	3.62	5
4	3.03	3.24	3.75	54	4.2	3.75	5.21
5	3.59	4.26	4.46	55	3.66	3.89	4.54
6	3.34	4.22	4.14	56	3.86	3.56	4.79
7	3.19	3.62	3.96	57	3.81	3.86	4.72
8	3.57	4.79	4.43	58	3.92	3.36	4.86
9	3.56	4.72	4.42	59	3.66	4.11	4.53
10	3	3.26	3.72	60	3.87	4.46	4.8
11	4.56	4.56	5.65	61	3.78	4.09	4.69
12	3.52	3.47	4.37	62	4.47	3.84	5.54
13	3.48	3.56	4.32	63	4.5	4	5.58
14	4.19	4.34	5.21	64	4.86	3.83	6.23
15	4.05	3.81	5.02	65	4.03	4.11	5
16	3.09	3.72	3.84	66	2.99	4.43	3.71
17	3.21	3.74	3.98	67	4.61	3.91	5.72
18	3.27	2.93	4.06	68	3.85	4.36	4.77
19	2.81	3.93	3.48	69	4.3	3.81	5.33
20	3.61	3.42	4.48	70	3.98	3.91	4.93
21	4.25	3.94	5.27	71	4.03	3.54	5
22	4.32	3.57	5.35	72	3.08	3.77	3.83
23	3.42	3.91	4.24	73	4.18	3.39	5.18
24	3.99	3.28	4.96	74	3.72	3.78	4.62
25	4.76	3.71	6.26	75	4.17	4.59	5.17
26	3.96	4.47	4.91	76	3.26	4.14	4.05
27	3.82	4.5	4.73	77	3.6	3.3	4.46
28	4.36	4.38	5.41	78	2.91	3.7	3.61
29	3.19	4.58	3.95	79	4.04	3.82	5.01
30	4.64	3.83	5.76	80	4.6	3.68	5.71
31	4.04	4.37	5.01	81	3.49	3.15	4.33
32	3.34	4.39	4.14	82	2.95	3.39	3.66
33	3.23	3.16	4	83	3.39	4.01	4.21
34	4.04	3.95	5.01	84	2.91	4.03	3.61
35	4.47	4.23	5.54	85	3.41	3.97	4.23
36	3.06	4.01	3.79	86	3.31	3.9	4.1
37	4.52	4.12	5.61	87	3.58	3.46	4.45
38	4.17	3.18	5.17	88	4.89	3.27	6.31
39	3.62	3.56	4.49	89	3.63	3.35	4.5
40	2.97	3.57	3.69	90	4.2	3.3	5.21
41	4.02	3.82	4.98	91	4.04	3.31	5.01
42	3.82	3.5	4.74	92	4.78	3.57	5.93
43	3.38	3.48	4.19	93	4.16	3.91	5.16
44	3.28	3.65	4.06	94	3.39	3.07	4.2
45	3.13	3.67	3.88	95	3.68	3.57	4.57
46	3.49	4.38	4.33	96	4.35	3.72	5.40
47	3.32	3.34	4.12	97	4.29	3.97	5.33
48	3.56	3.26	4.41	98	4.58	4.33	5.68
49	3.42	3.83	4.24	99	3.85	3.29	4.78
50	3.99	3.92	4.96	100	4.62	3.54	5.73

Tabla A.1: Valores de pH medidos para cada muestra de naranja de la base de datos.

Apéndice B

Ampliación de resultados

El siguiente apéndice está formado por el conjunto de resultados obtenidos en la sección 4.2 desglosados con el objetivo de que el lector tenga una mayor facilidad para su interpretación.

B.1. Boxplots de pH

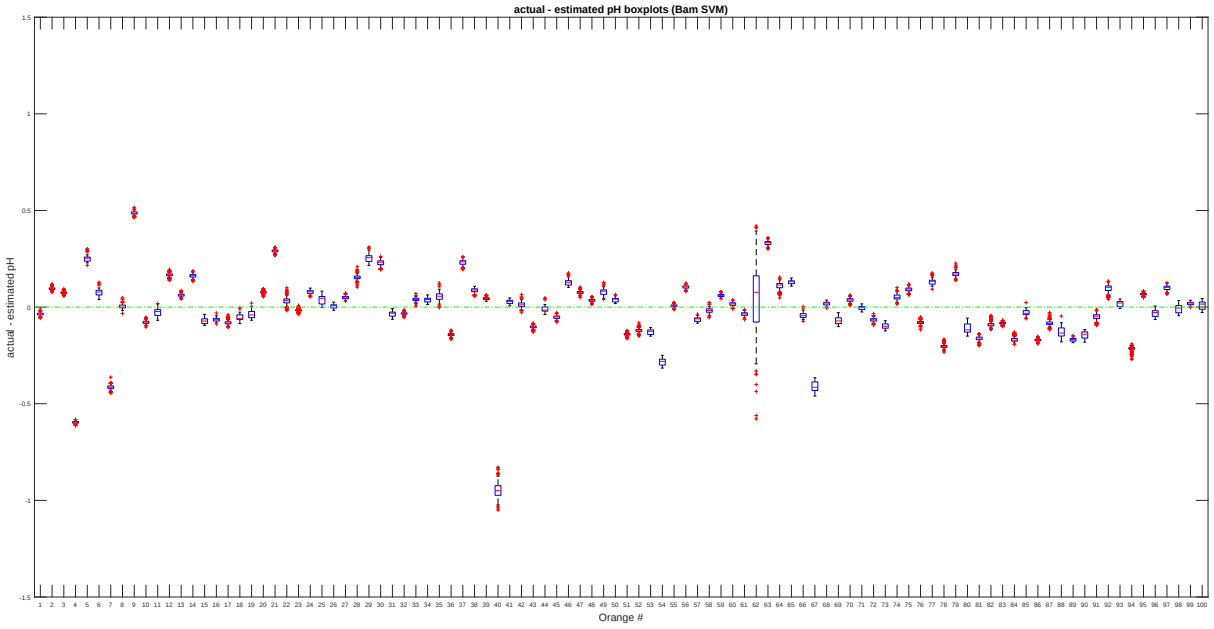


Figura B.1: Boxplots relativos a las 100 muestras de la variedad de naranja *Bam* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión SVM. ($n_{cic} = 1000$, 200 valores de media por muestra).

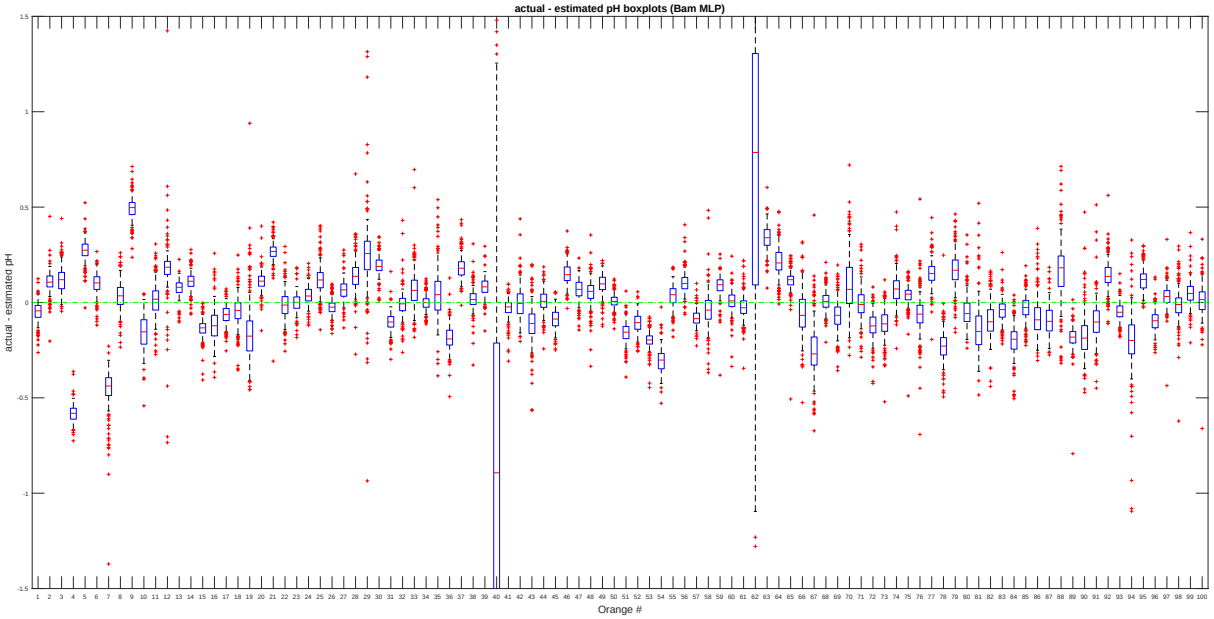


Figura B.2: Boxplots relativos a las 100 muestras de la variedad de naranja *Bam* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión MLP. ($n_{cic} = 1000$, 200 valores de media por muestra).

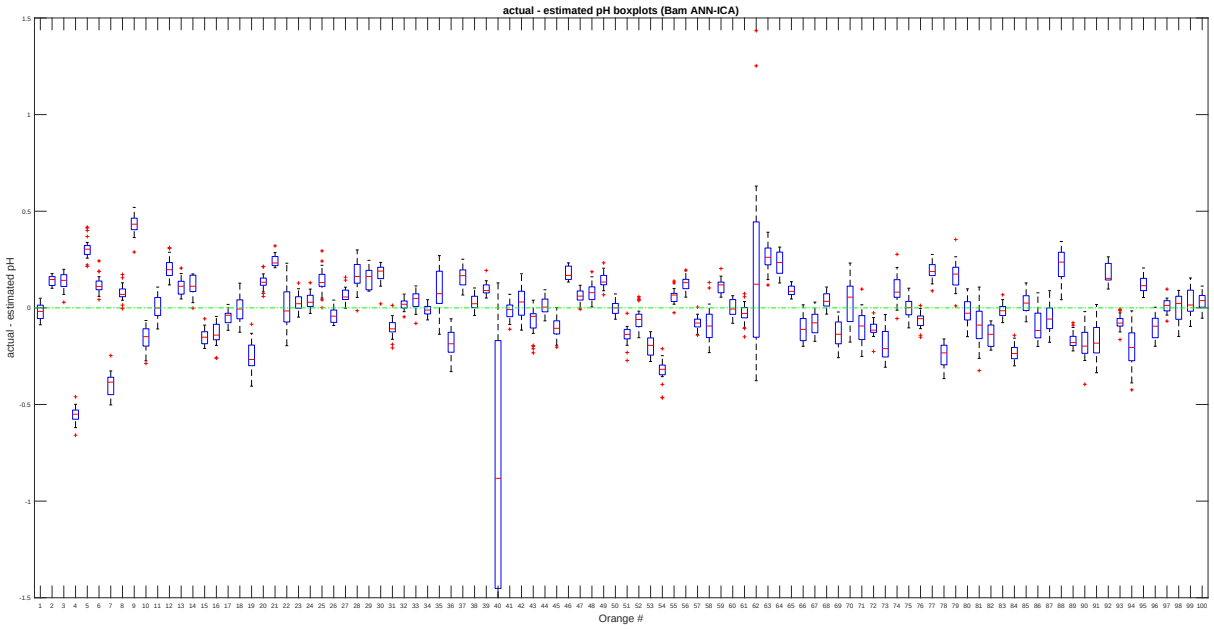


Figura B.3: Boxplots relativos a las 100 muestras de la variedad de naranja *Bam* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión ANN-ICA. ($n_{cic} = 1000$, 200 valores de media por muestra).

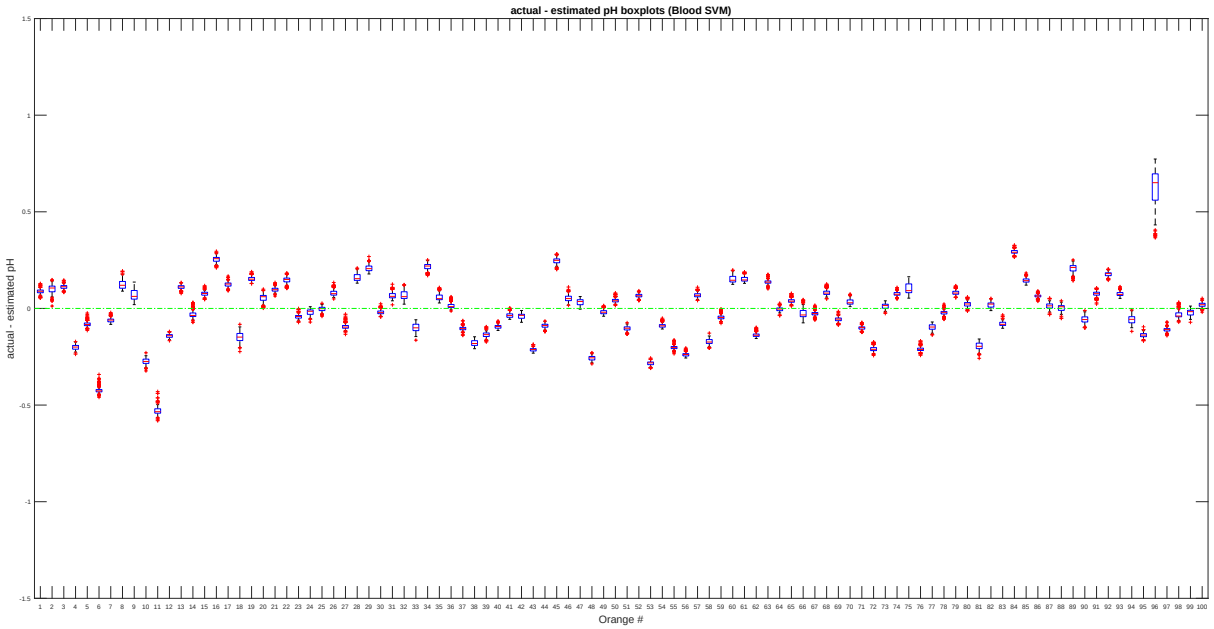


Figura B.4: Boxplots relativos a las 100 muestras de la variedad de naranja *Blood* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión SVM. ($n_{cic} = 1000$, 200 valores de media por muestra).

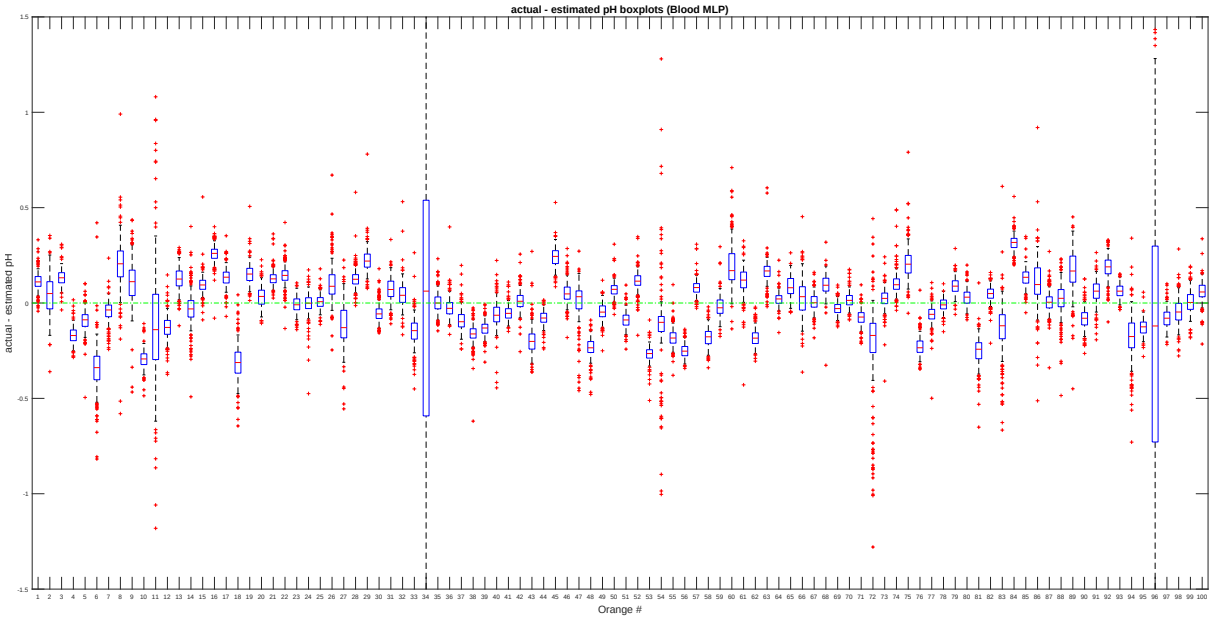


Figura B.5: Boxplots relativos a las 100 muestras de la variedad de naranja *Blood* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión MLP. ($n_{cic} = 1000$, 200 valores de media por muestra).

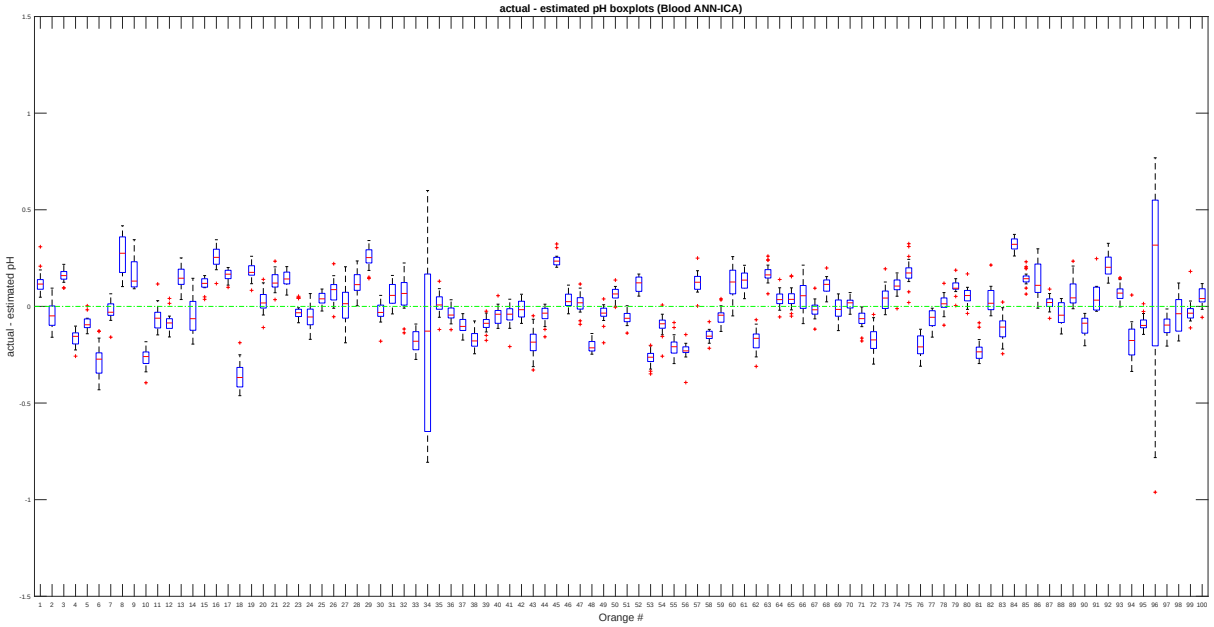


Figura B.6: Boxplots relativos a las 100 muestras de la variedad de naranja *Blood* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión ANN-ICA. ($n_{cic} = 1000$, 200 valores de media por muestra).

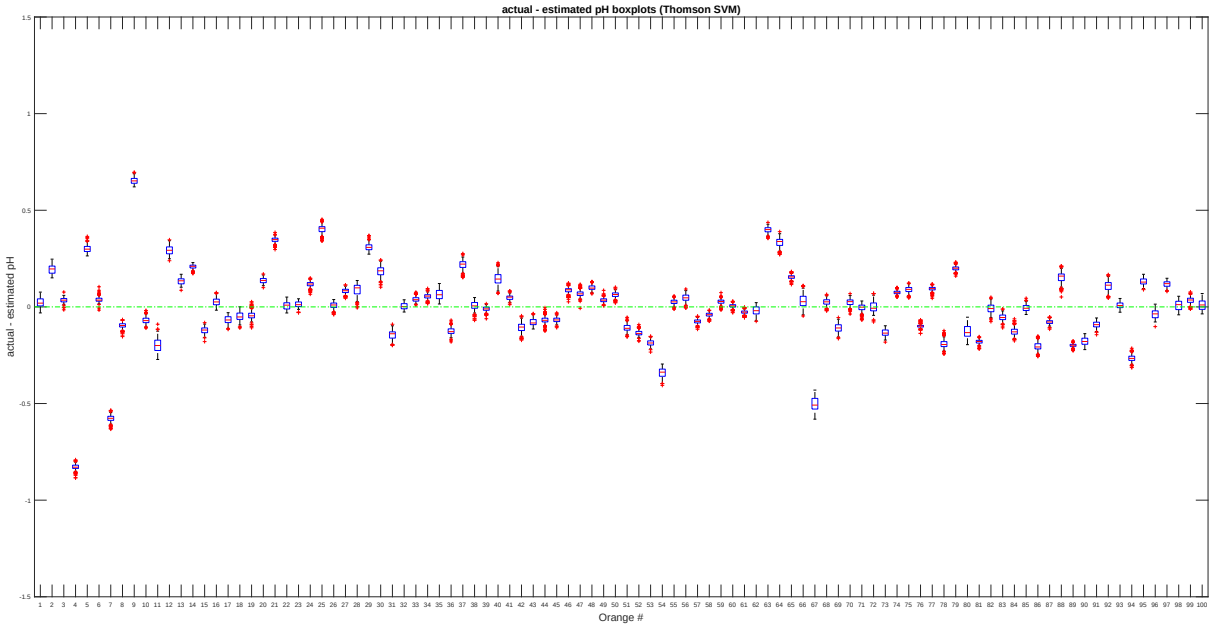


Figura B.7: Boxplots relativos a las 100 muestras de la variedad de naranja *Thomson* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión SVM. ($n_{cic} = 1000$, 200 valores de media por muestra).

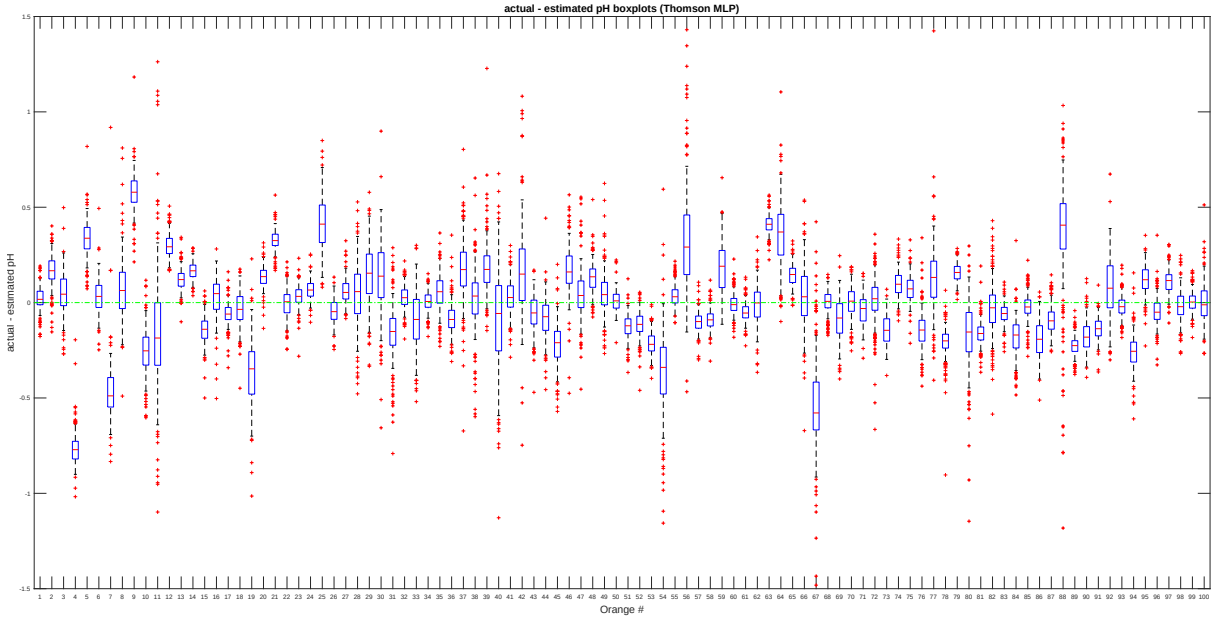


Figura B.8: Boxplots relativos a las 100 muestras de la variedad de naranja *Thomson* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión MLP. ($n_{cic} = 1000$, 200 valores de media por muestra).

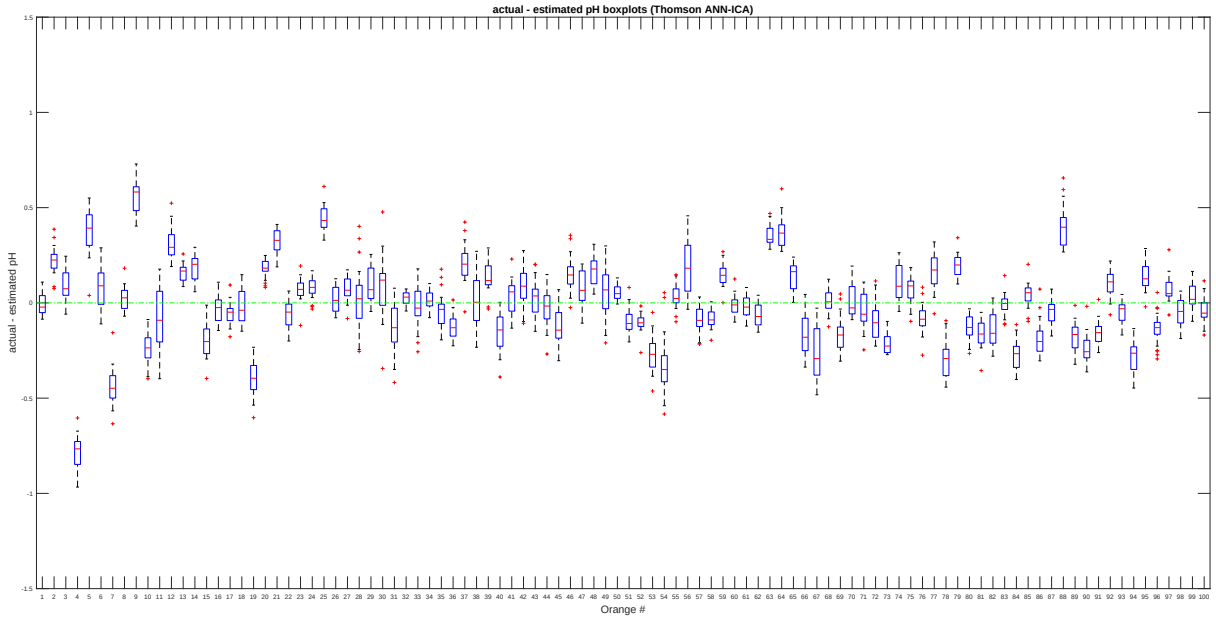


Figura B.9: Boxplots relativos a las 100 muestras de la variedad de naranja *Thomson* indicativos de la diferencia entre el pH real y el pH estimado para el método de regresión ANN-ICA. ($n_{cic} = 1000$, 200 valores de media por muestra).

B.2. Boxplots de errores

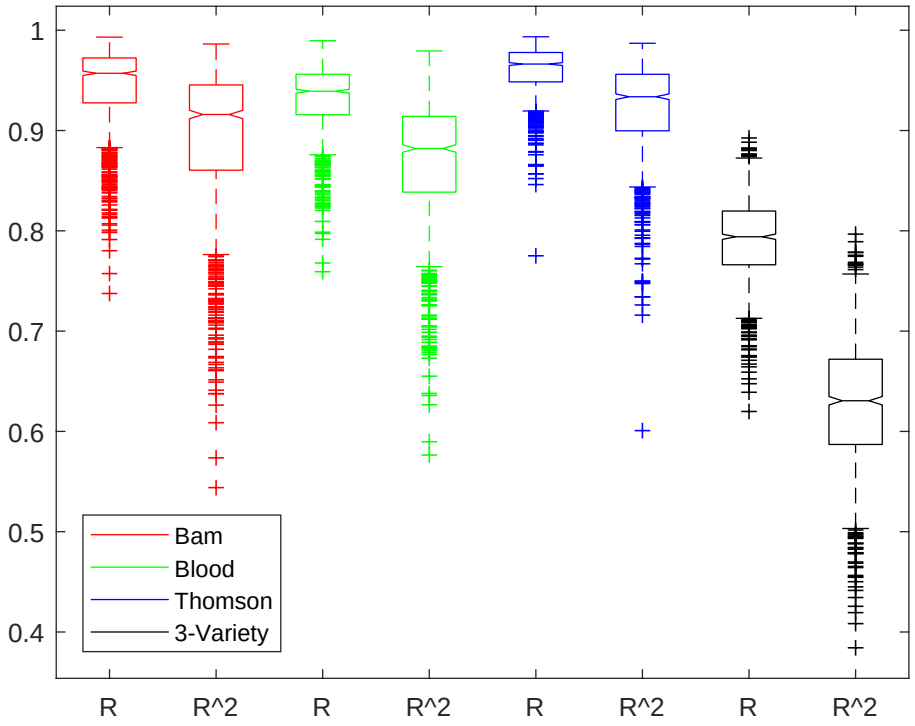


Figura B.10: Boxplots de los coeficientes de error R y R^2 para el método de regresión SVM ($n_{cic} = 1000$).

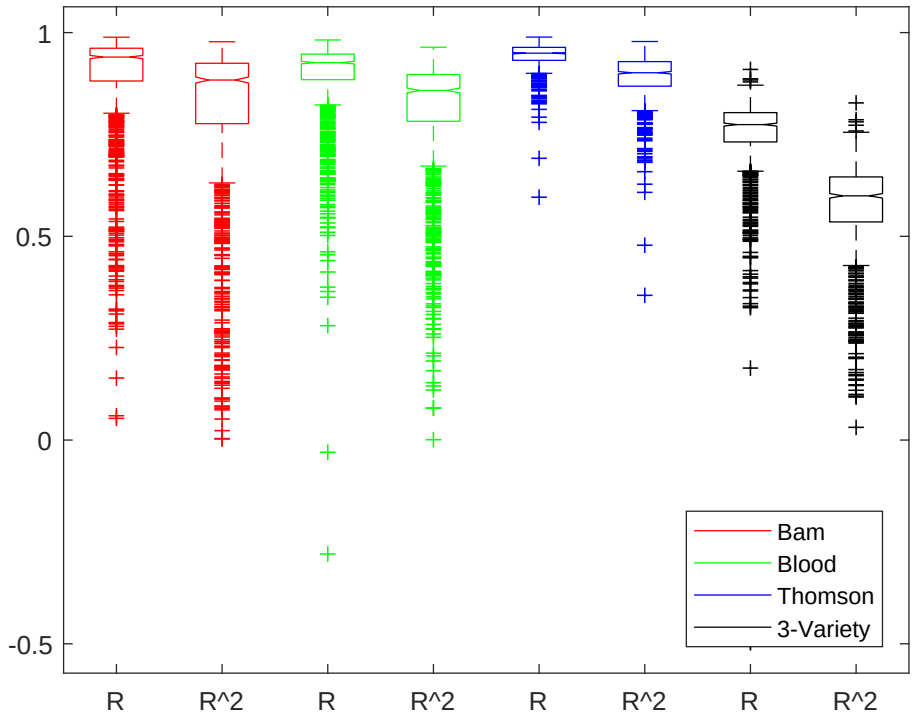


Figura B.11: Boxplots de los coeficientes de error R y R^2 para el método de regresión MLP ($n_{cic} = 1000$).

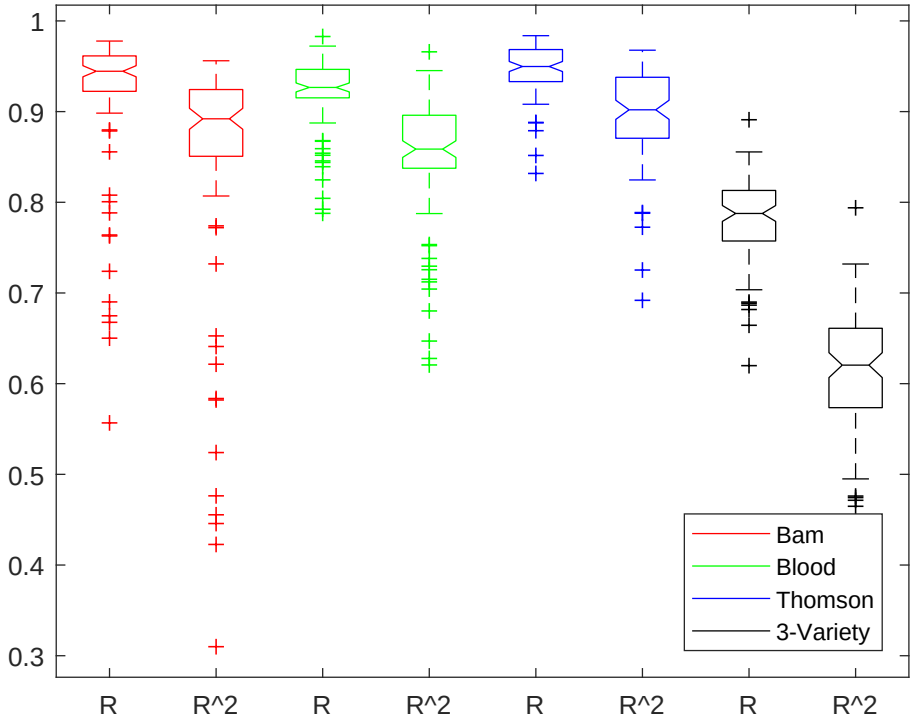


Figura B.12: Boxplots de los coeficientes de error R y R^2 para el método de regresión ANN-ICA ($n_{cic} = 1000$).

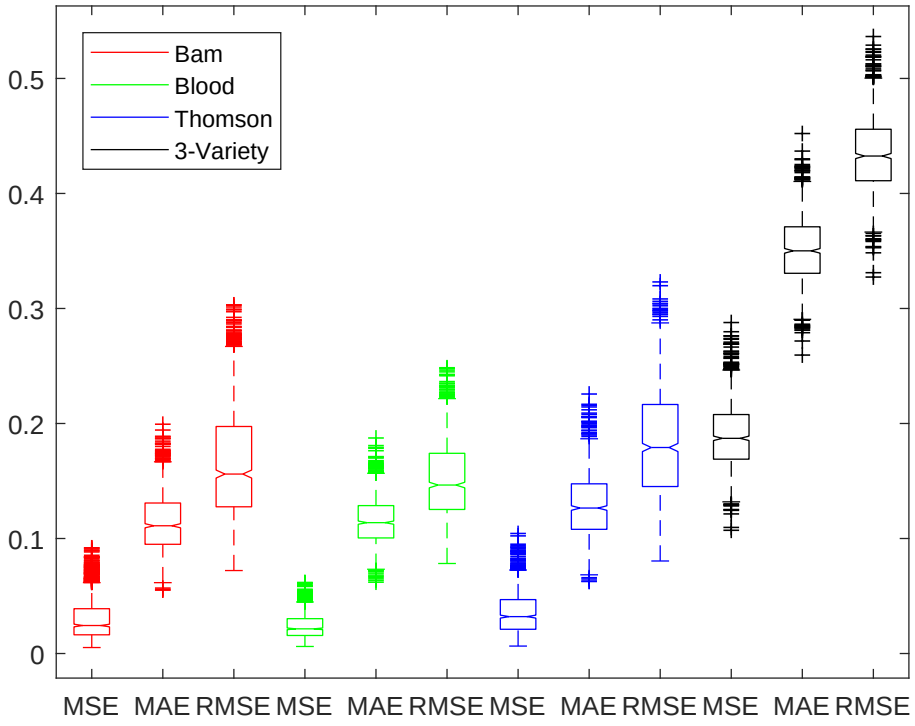


Figura B.13: Boxplots de los coeficientes de error MSE, MAE y RMSE para el método de regresión SVM ($n_{cic} = 1000$).

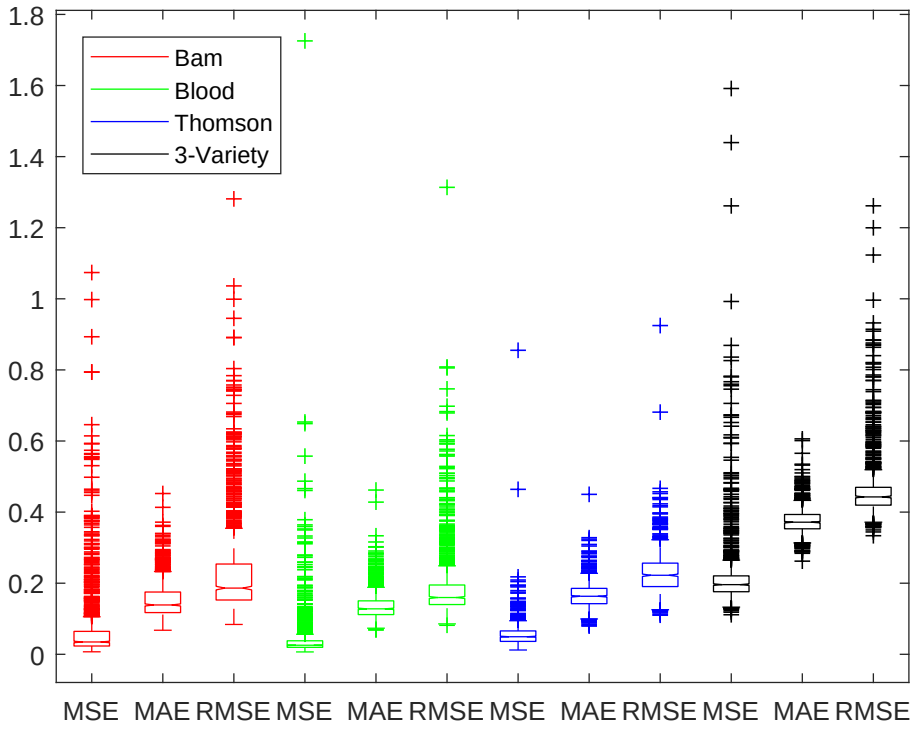


Figura B.14: Boxplots de los coeficientes de error MSE, MAE y RMSE para el método de regresión MLP ($n_{cic} = 1000$).

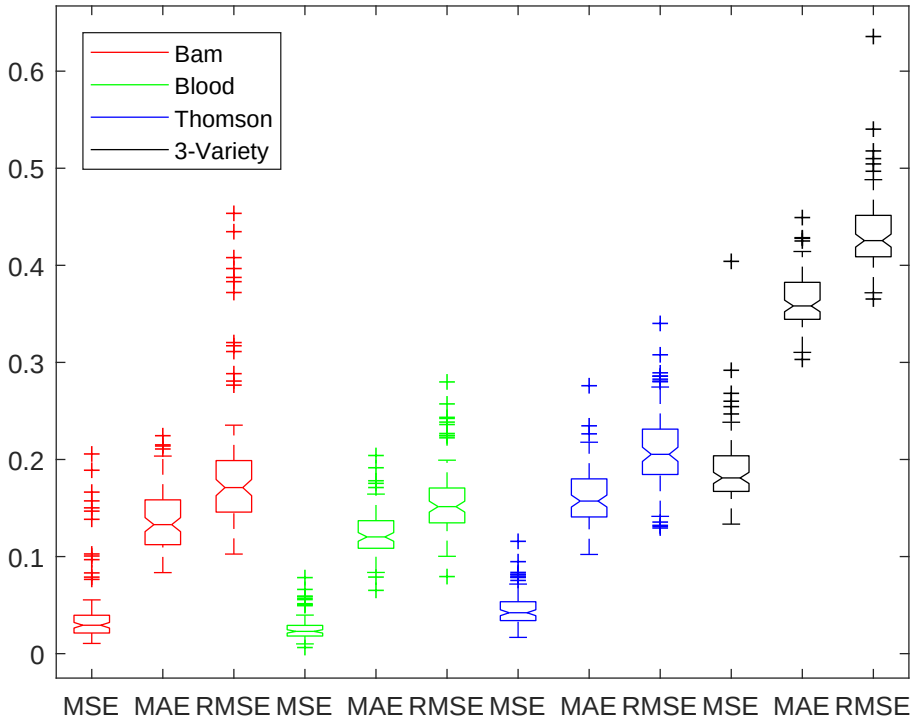


Figura B.15: Boxplots de los coeficientes de error MSE, MAE y RMSE para el método de regresión ANN-ICA ($n_{cic} = 1000$).

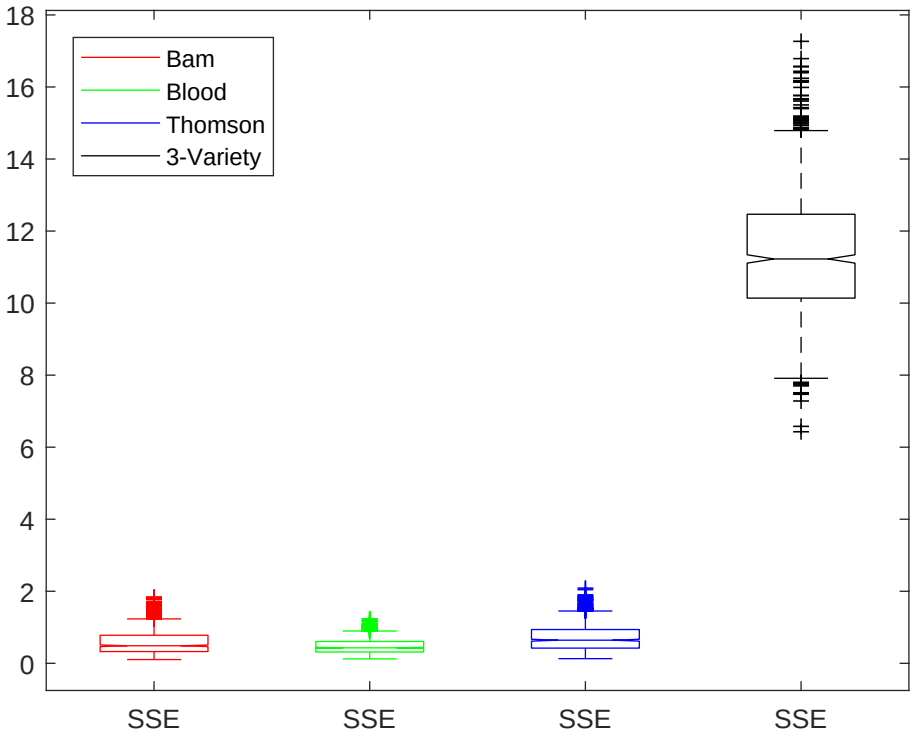


Figura B.16: Boxplots del coeficiente de error SSE para el método de regresión SVM ($n_{cic} = 1000$).

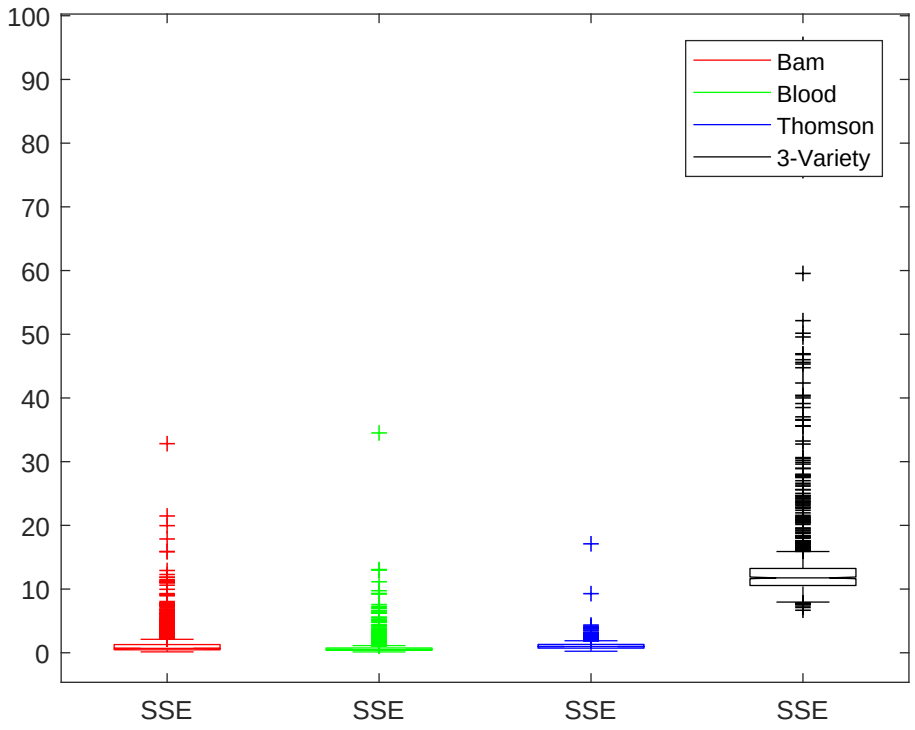


Figura B.17: Boxplots del coeficiente de error SSE para el método de regresión MLP ($n_{cic} = 1000$).

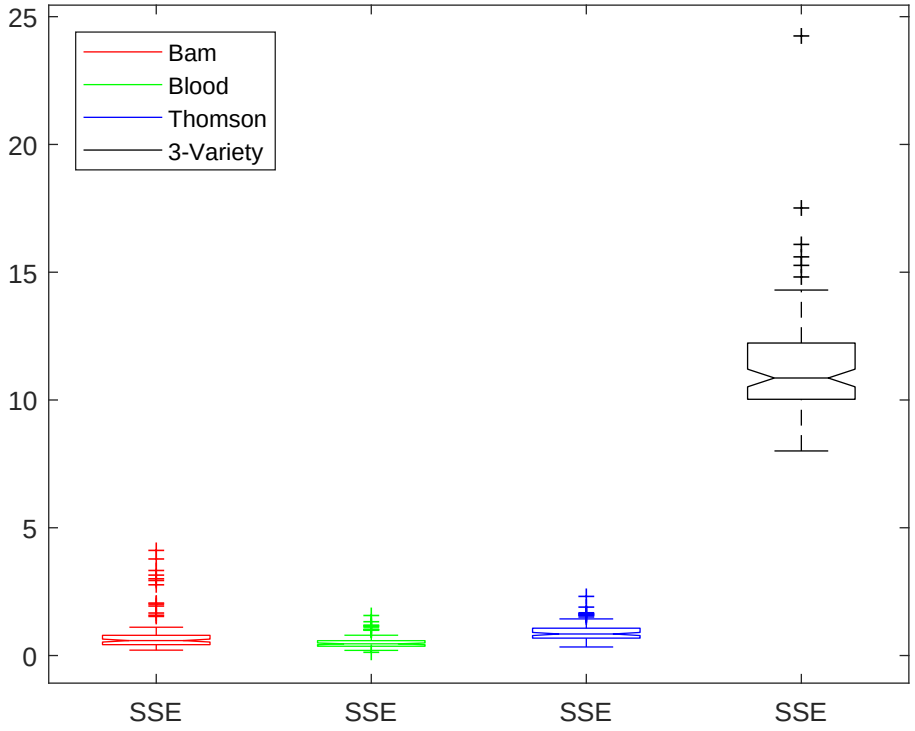


Figura B.18: Boxplots del coeficiente de error SSE para el método de regresión ANN-ICA ($n_{cic} = 1000$).

B.3. pH real vs pH estimado

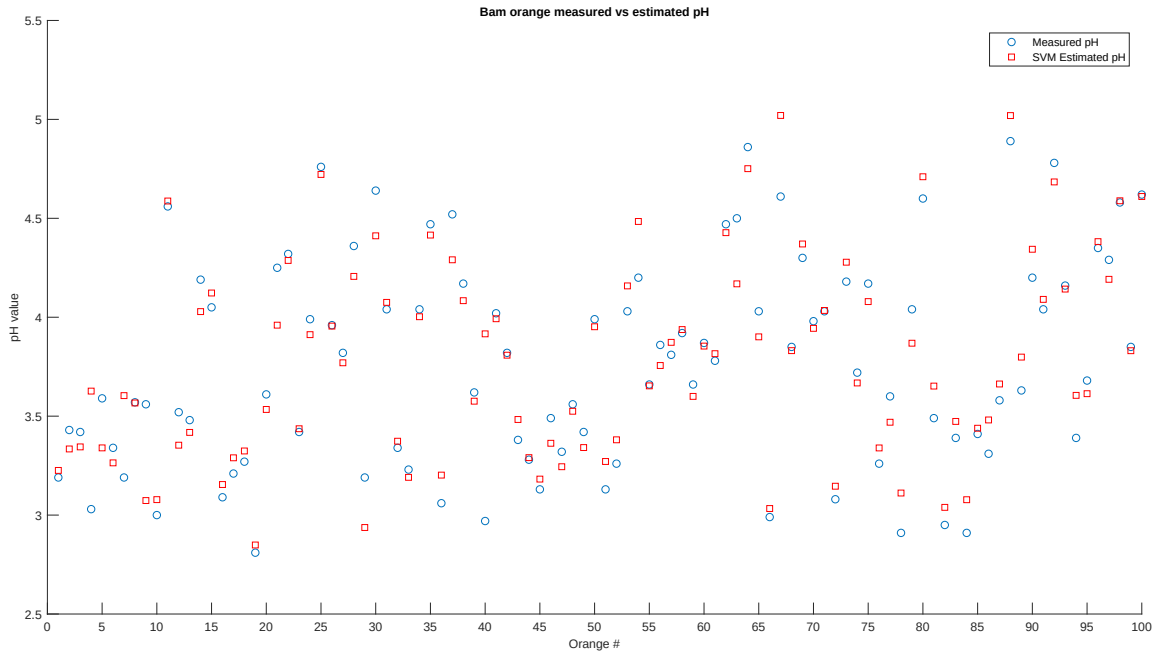


Figura B.19: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión SVM (rojo) para la variedad de naranja *Bam*. ($n_{cic} = 1000$, 200 valores de media por muestra).

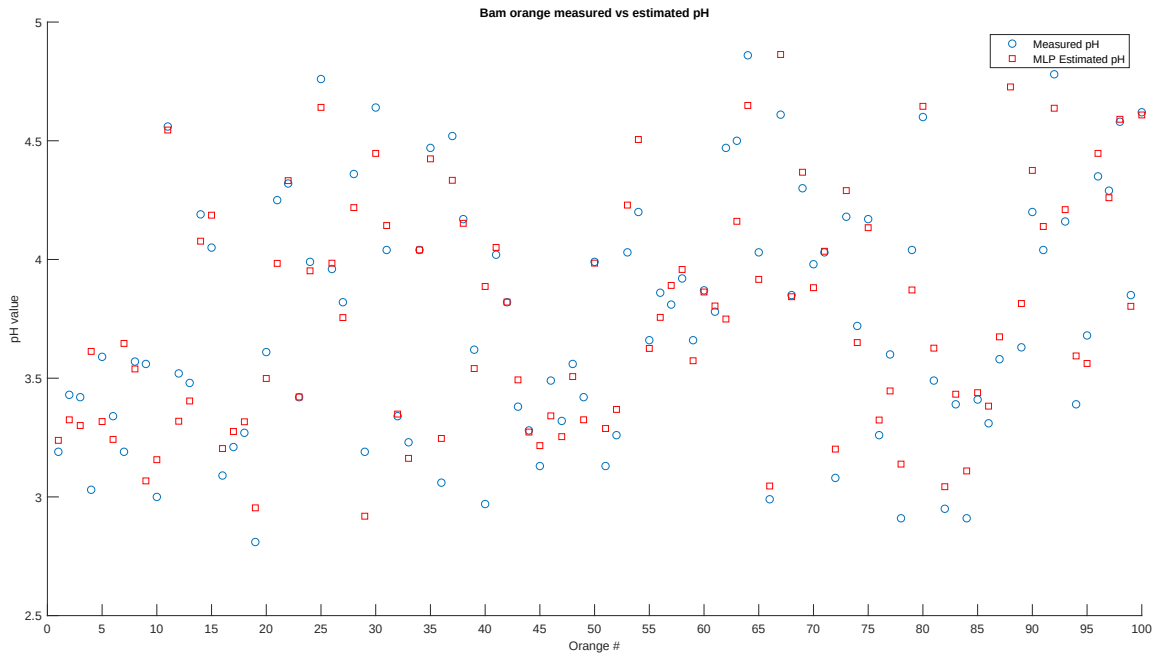


Figura B.20: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión MLP (rojo) para la variedad de naranja *Bam.* ($n_{cic} = 1000$, 200 valores de media por muestra).

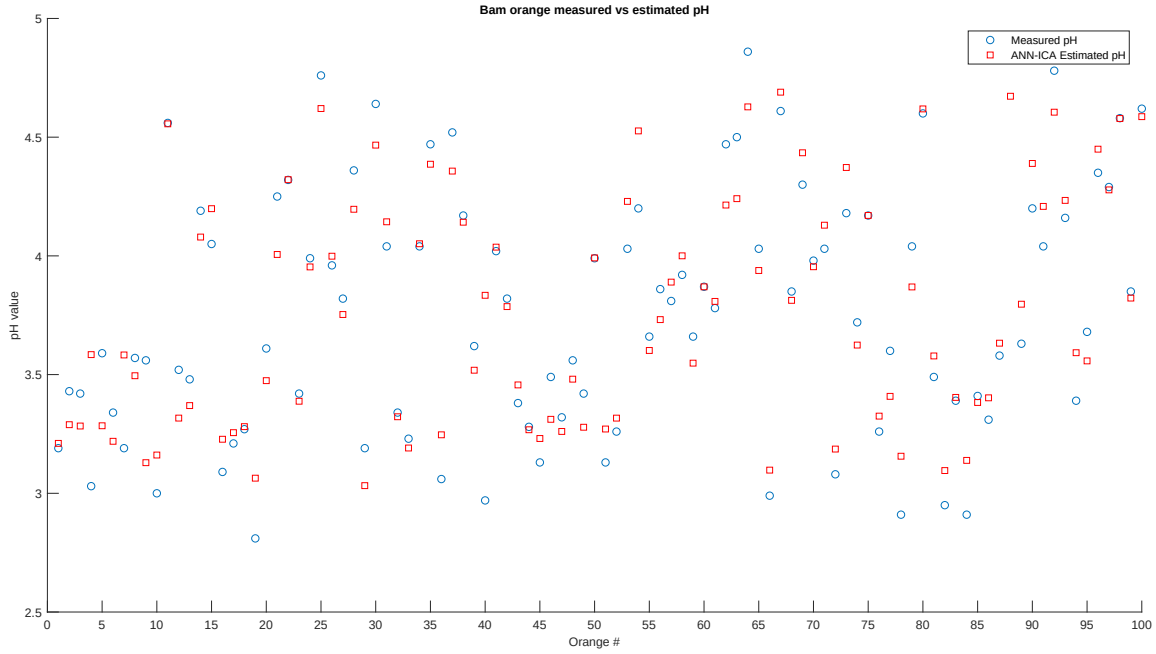


Figura B.21: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión ANN-ICA (rojo) para la variedad de naranja *Bam.* ($n_{cic} = 1000$, 200 valores de media por muestra).

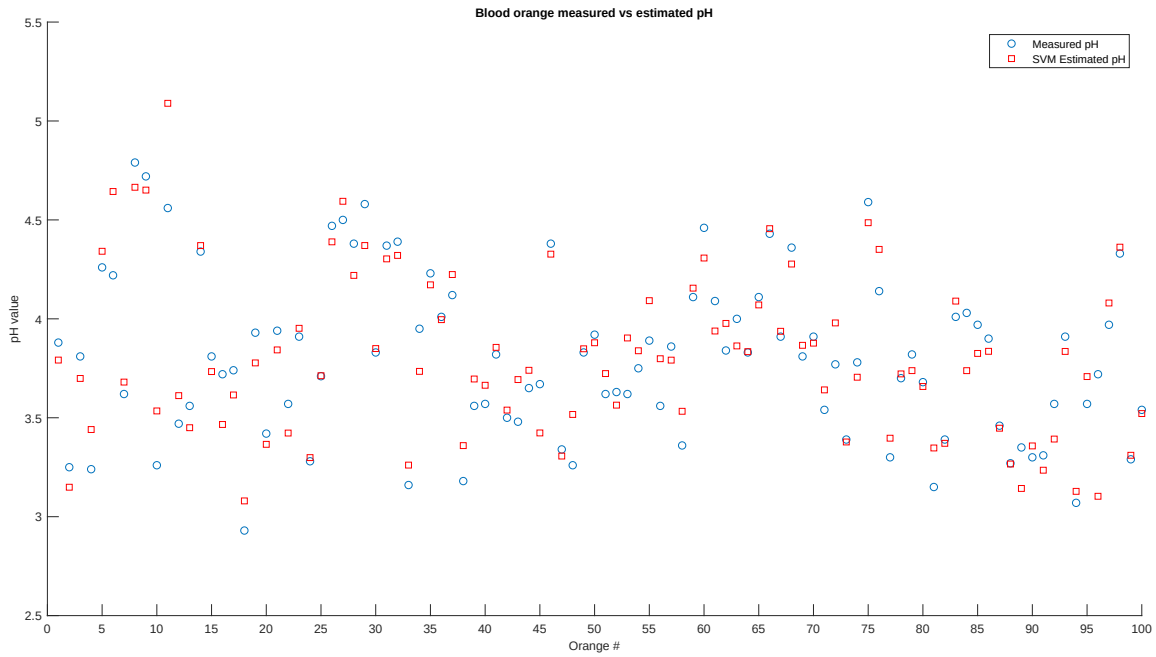


Figura B.22: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión SVM (rojo) para la variedad de naranja *Blood*. ($n_{cic} = 1000$, 200 valores de media por muestra).

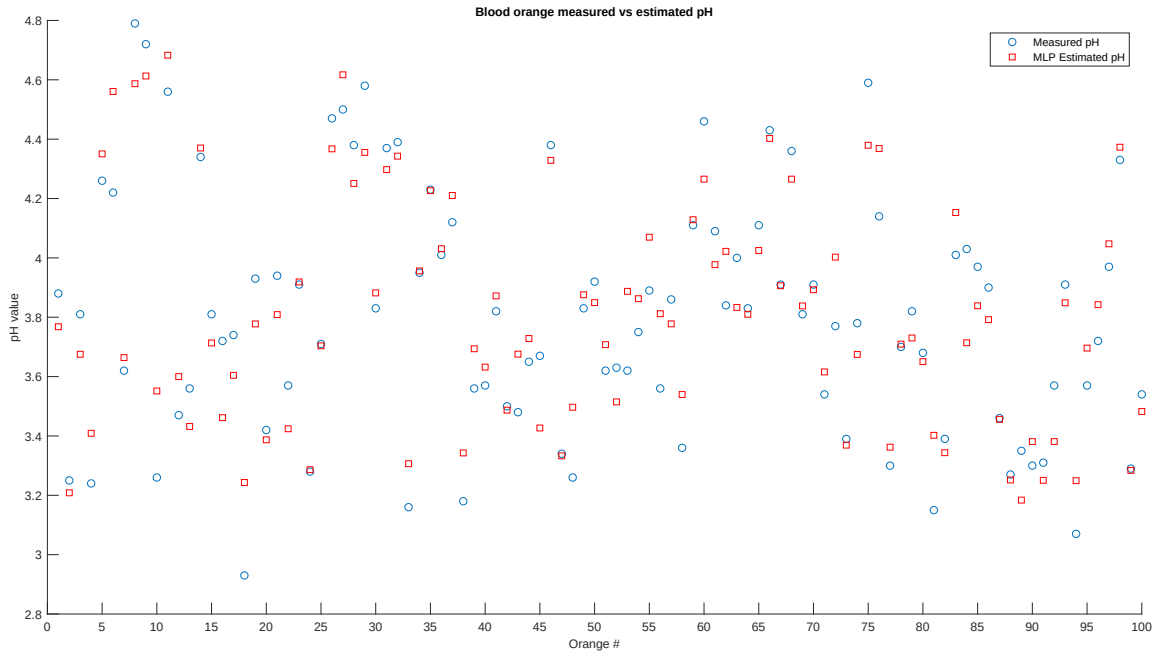


Figura B.23: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión MLP (rojo) para la variedad de naranja *Blood*. ($n_{cic} = 1000$, 200 valores de media por muestra).

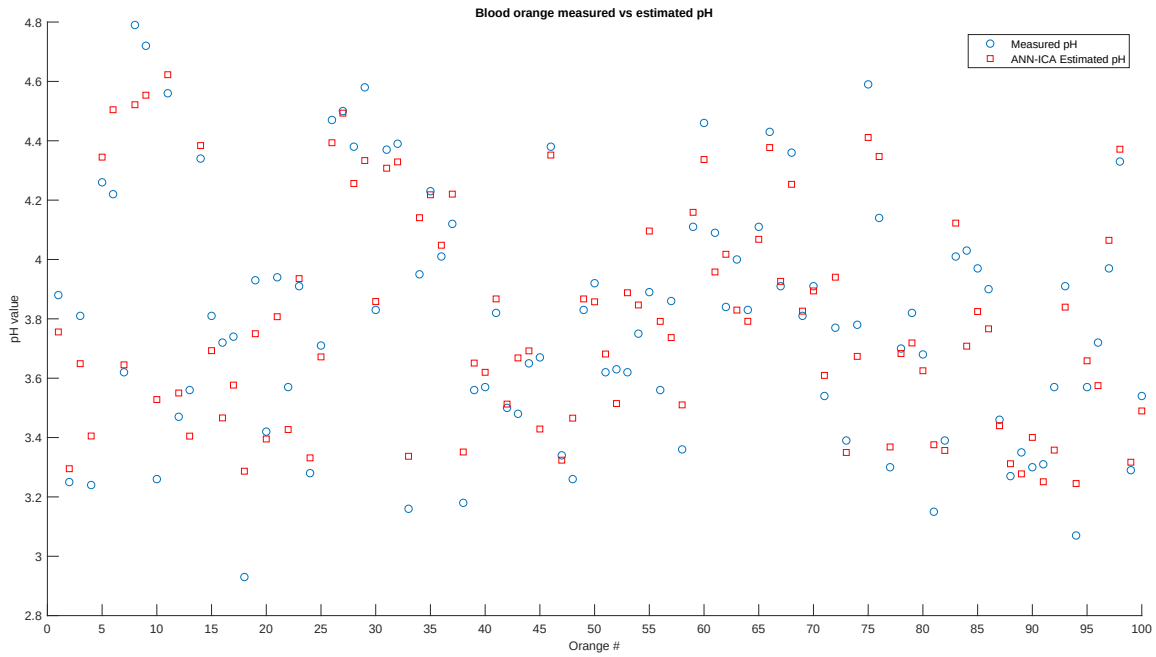


Figura B.24: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión ANN-ICA (rojo) para la variedad de naranja *Blood*. ($n_{cic} = 1000$, 200 valores de media por muestra).

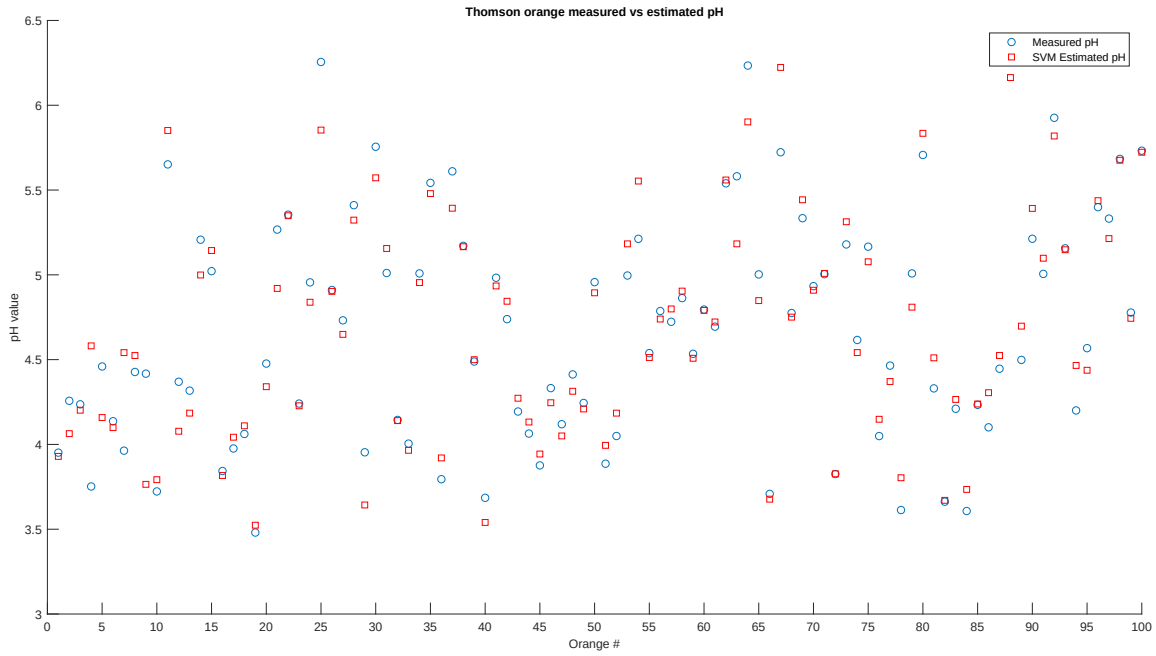


Figura B.25: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión SVM (rojo) para la variedad de naranja *Thomson*. ($n_{cic} = 1000$, 200 valores de media por muestra).

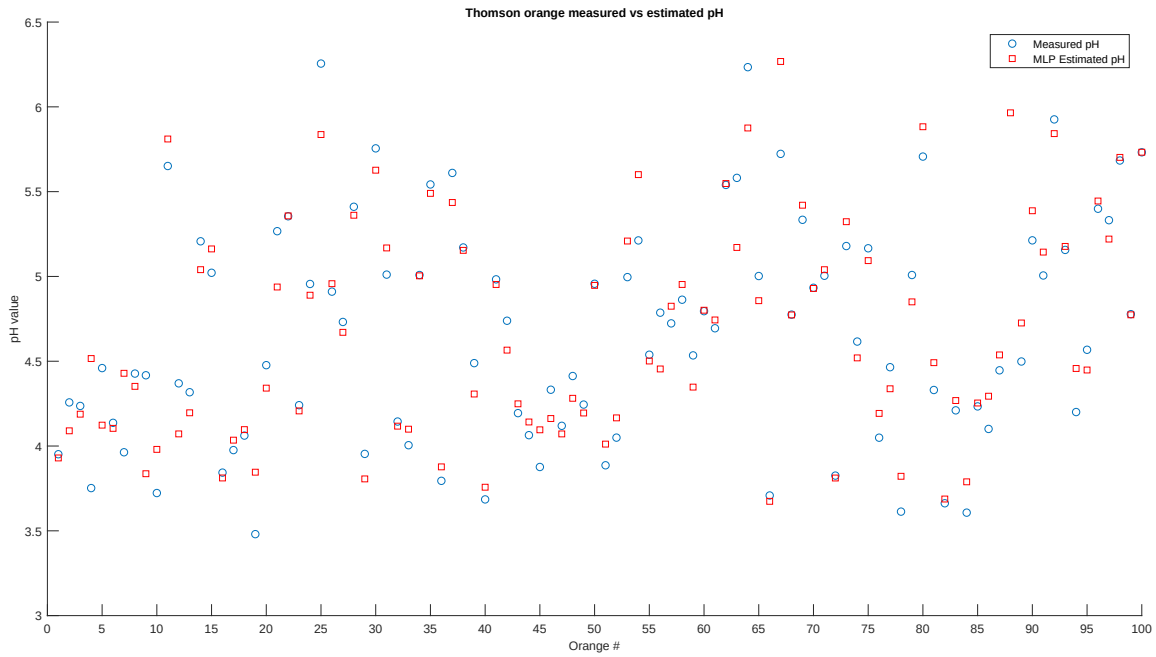


Figura B.26: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión MLP (rojo) para la variedad de naranja *Thomson*. ($n_{cic} = 1000$, 200 valores de media por muestra).

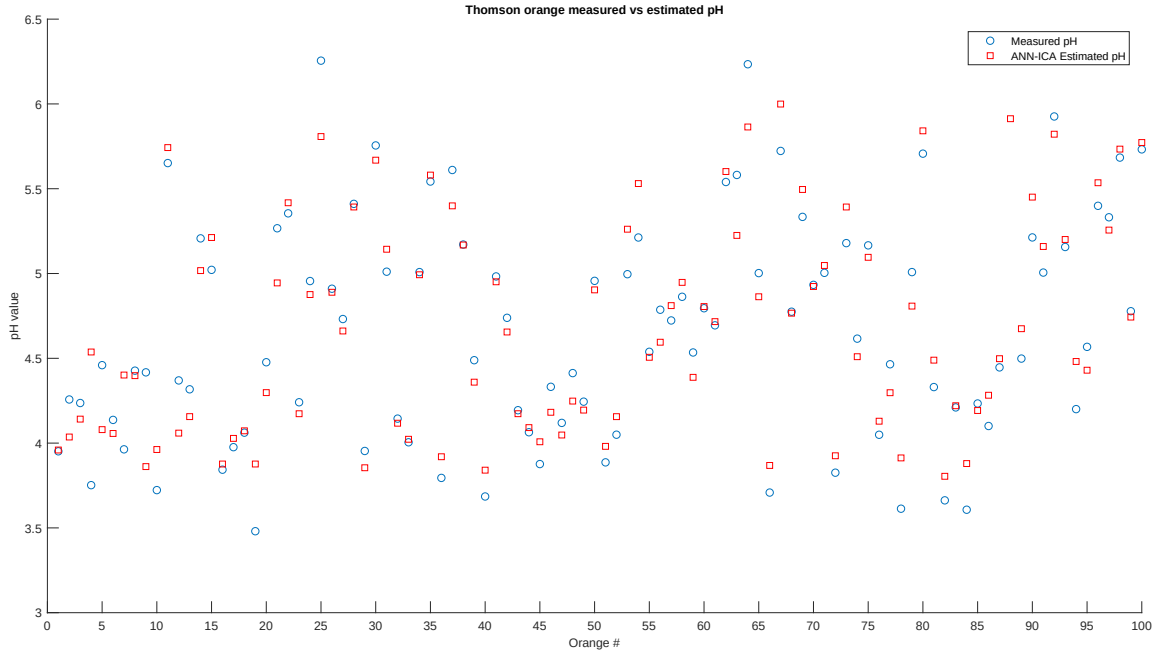


Figura B.27: Comparación del valor de pH real (azul) con los valores estimados medios de pH obtenidos a partir del método de regresión ANN-ICA (rojo) para la variedad de naranja *Thomson*. ($n_{cic} = 1000$, 200 valores de media por muestra).

Apéndice C

Funciones y *Scripts* de MATLAB

C.1. Diagrama de dependencias entre funciones

El conjunto de código de MATLAB empleado en el presente proyecto se divide en cuatro bloques diferenciados: La extracción de características se realiza a partir del procesamiento de las imágenes en formato *jpg* que conforman la base de datos. La reducción de la dimensionalidad del conjunto de características obtenido es llevada a cabo a través del algoritmo PCA ejecutado mediante el método de los autovalores y autovectores. Una vez obtenidas las componentes principales y reducida la dimensionalidad del banco de datos se implementan los tres métodos de regresión propuestos teniendo en cuenta la normalización de los valores de entrada y la utilización del pH real medido. Por último, a partir de las salidas obtenidas de los métodos de regresión, pH estimado y coeficientes de error, se obtienen los resultados (ver capítulo 4) más concluyentes de este experimento.

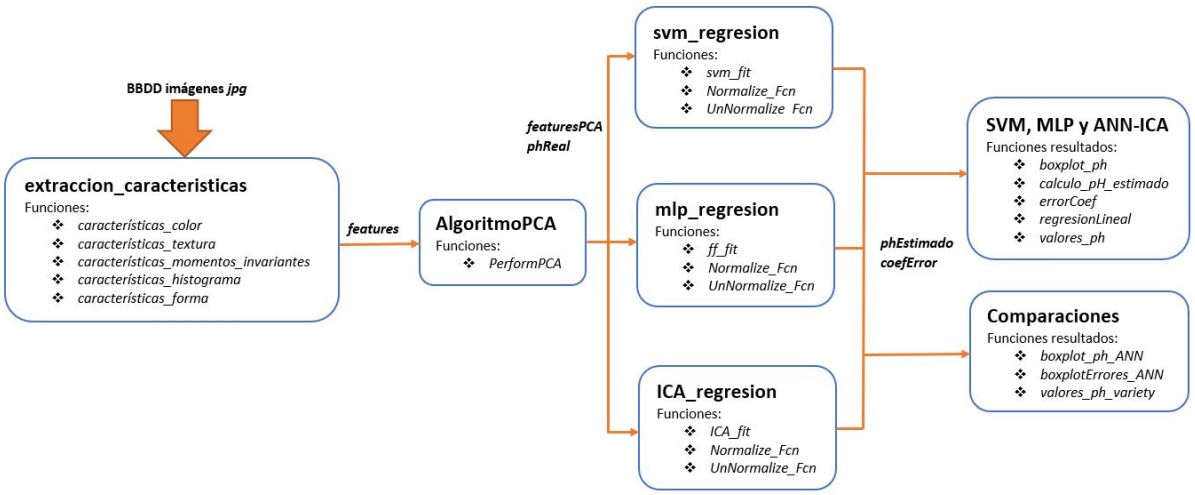


Figura C.1: Diagrama de dependencias entre las funciones diseñadas en MATLAB para obtener los resultados del presente proyecto.

C.2. Extracción de características

Este *script* se encarga de la generación de las matrices de características × muestras para cada una de las variedades de naranjas propuestas en la base de datos. Las llamadas a las funciones *caracteristicas_color*, *caracteristicas_textura*, *caracteristicas_histograma*, *caracteristicas_momentos_invariantes* y *caracteristicas_forma* generan el conjunto de 452 características para cada muestra de naranja de la base de datos.

```
%Extracción de características de las imagenes de los tres tipos de
%naranjas: Bam, Blood y Thomson

% Script de ejecución de las funciones de extracción de los
% diferentes grupos de características para formar el vector de salida
% correspondiente a las muestras de cada tipo de naranja.
%   Funciones:
%       caracteristicas_color
%       caracteristicas_textura
%       caracteristicas_histrograma
%       caracteristicas_momentos_invariantes
%       caracteristicas_forma
%
%   Entrada: imagen en formato rgb: "rgb".
%   Salida: vector de caracteristicas: "feature".

clear all
clc
close all

%Inicialización de variables
N = 452;
M = 100;

%Bucle for para analizar el conjunto de imagenes
for i=1:M

    %Cadena que contiene el nombre de la imagen
    n = num2str(i);
    name = ['img',n,'.jpg'];

    %Imagen cargada
    rgb = imread(name);

    %Ejecucion de las funciones de extracción de grupos de características
    color = caracteristicas_color(rgb); %336
    textura = caracteristicas_textura(rgb); %80
    hist = caracteristicas_histograma(rgb); %6
    momentos = caracteristicas_momentos_invariantes(rgb); %10
    forma = caracteristicas_forma(rgb); %20

    %Vector de características: 452
    featuresThomson(:,i) = [color;textura;hist;momentos;forma];
    featuresBam(:,i) = [color;textura;hist;momentos;forma];
    featuresBlood(:,i) = [color;textura;hist;momentos;forma];

end

save('featuresThomson.mat','featuresThomson');
save('featuresBam.mat','featuresBam');
save('featuresBlood.mat','featuresBlood');
```

C.3. Análisis de Componentes Principales (PCA)

El siguiente *script* de MATLAB representa la función de extracción de los autovalores y autovectores a partir de la matriz de caracteristicas × muestras.

```
function [Q lambda]=PerformPCA(X)
% Función para el cálculo de las componentes principales mediante el método
% de los autovalores y autovectores (eigenvalue)
%     Entradas:    X-> matriz de caracteristicas (filas: carac, col: muestras)
%     Salidas:     Q-> autovectores
%                  lambda-> autovalores

    C=cov(X');

    [Q LAMBDA]=eig(C);

    lambda=diag(LAMBDA);
    [lambda SortOrder]=sort(lambda,'descend');

    Q=Q(:,SortOrder);

end
```

C.4. Métodos de regresión

El siguiente conjunto de *scripts* representan las funciones definitorias de los tres métodos de regresión utilizados: SVM, MLP y ANN-ICA. Se puede observar la utilización de funciones de normalización y desnormalización, *Normalice_Fcn* y *UnNormalize_Fcn*, necesarias para el correcto funcionamiento del método de regresión y que la salida estimada se pueda comparar con los valores originales del grado de acidez (pH).

```
function [svmStruct, ind_test, targets, predicted_target] = svm_fit(CARAC,DIAG)
% Función para generar una SVM utilizada como regresor lineal
% Entradas:   CARAC-> matriz de componentes principales/muestras
%             DIAG-> valores de pH medidos para cada muestra
% Salidas:    svmStruct-> estructura svm generada a partir de "fitrsvm"
%             ind_test-> índices del conjunto de valores de test
%             targets-> valores de pH medidos para cada muestra
%             predicted_target-> pH estimado para el conjunto de test

inputs = CARAC;
targets = DIAG';

%Normalización
Min_inputs = min(inputs);
Max_inputs = max(inputs);

Min_targets = min(targets);
Max_targets = max(targets);

for ii = 1:size(inputs,2)
    inputs_norm(:,ii) = Normalize_Fcn(inputs(:,ii),Min_inputs(ii),Max_inputs(ii));
end

for ii = 1:size(targets,2)
    targets_norm(:,ii) = Normalize_Fcn(targets(:,ii),Min_targets(ii),Max_targets(ii));
end

% Dividir los parámetros para entrenar y testear (un 80% de train y un 20%
% de testing)
DataNum = size(inputs_norm,1);
```

```

TrPercent = 80;
TrNum = round(DataNum * TrPercent / 100);
TsNum = DataNum - TrNum;

R = randperm(DataNum);
trIndex = R(1 : TrNum);
tsIndex = R(1+TrNum : end);

inputsTrain = inputs_norm(trIndex,:);
targetsTrain = targets_norm(trIndex,:);

inputsTest = inputs_norm(tsIndex,:);
targetsTest = targets_norm(tsIndex,:);

%Regresión con un SVM
svmStruct = fitrsvm(inputsTrain,targetsTrain);
predicted_target_norm = predict(svmStruct, inputsTest);

%Indices test
ind_test = tsIndex;

%Desnormalización
for ii = 1:size(predicted_target_norm,2)
    predicted_target(:,ii) = UnNormalize_Fcn(predicted_target_norm(:,ii),Min_targets(ii),Max_targets(ii));
end

end
```

```
function [indTest, targets, y] = ff_fit(features, pH)
% Función para generar una red MLP utilizada como regresor lineal
% Entradas:   features-> matriz de componentes principales/muestras
%             pH-> valores de pH medidos para cada muestra
% Salidas:    indTest-> índices del conjunto de valores de test
%             targets-> valores de pH medidos para cada muestra
%             y-> pH estimado para el conjunto de test

inputs = features;
targets = pH';

%Normalización
Min_inputs = min(inputs);
Max_inputs = max(inputs);

Min_targets = min(targets);
Max_targets = max(targets);

for ii = 1:size(inputs,2)
    inputs_norm(:,ii) = Normalize_Fcn(inputs(:,ii),Min_inputs(ii),Max_inputs(ii));
end

for ii = 1:size(targets,2)
    targets_norm(:,ii) = Normalize_Fcn(targets(:,ii),Min_targets(ii),Max_targets(ii));
end

% Dividir los parámetros para entrenar y testear (un 80% de train y un 20%
% de testing)
DataNum = size(inputs_norm,1);

TrPercent = 80;
TrNum = round(DataNum * TrPercent / 100);
TsNum = DataNum - TrNum;

R = randperm(DataNum);
trIndex = R(1 : TrNum);
tsIndex = R(1+TrNum : end);

inputsTrain = inputs_norm(trIndex,:);
targetsTrain = targets_norm(trIndex,:);

```

```
inputsTest = inputs_norm(tsIndex,:);
targetsTest = targets_norm(tsIndex,:);

% Creación de la red neuronal
net = feedforwardnet([5 4],'trainlm');

net.layers{1}.transferFcn = 'tansig';
net.layers{2}.transferFcn = 'tansig';

net.trainParam.showWindow = 0;

%Entrenamiento
[net,tr] = train(net, inputsTrain', targetsTrain');
[net,tr] = train(net, inputsTrain', targetsTrain');

%Salida normalizada
y_norm = net(inputsTest');
y_norm = y_norm';

indTest = tsIndex;

%Desnormalización
for ii = 1:size(y_norm,2)
    y(:,ii) = UnNormalize_Fcn(y_norm(:,ii),Min_targets(ii),Max_targets(ii));
end

end
```

```
function [predicted, tsIndex, Y] = ICA_fit(features, pH)
% Función para generar una red MLP utilizada como regresor lineal
% Entradas:   features-> matriz de componentes principales/muestras
%             pH-> valores de pH medidos para cada muestra
% Salidas:    tsIndex-> índices del conjunto de valores de test
%             Y-> valores de pH medidos para cada muestra
%             predicted-> pH estimado para el conjunto de test

tic

X = features;
Y = pH';

DataNum = size(X,1);
InputNum = size(X,2);
OutputNum = size(Y,2);

%% Normalizar
MinX = min(X);
MaxX = max(X);

MinY = min(Y);
MaxY = max(Y);

XN = X;
YN = Y;

for ii = 1:InputNum
    XN(:,ii) = Normalize_Fcn(X(:,ii),MinX(ii),MaxX(ii));
end

for ii = 1:OutputNum
    YN(:,ii) = Normalize_Fcn(Y(:,ii),MinY(ii),MaxY(ii));
end

%% Conjuntos de entrenamiento (80 %) y test (20 %)
TrPercent = 80;
TrNum = round(DataNum * TrPercent / 100);
TsNum = DataNum - TrNum;

R = randperm(DataNum);
trIndex = R(1 : TrNum);
```

```
tsIndex = R(1+TrNum : end);

Xtr = XN(trIndex,:);
Ytr = YN(trIndex,:);

Xts = XN(tsIndex,:);
Yts = YN(tsIndex,:);

%%Estructura de la red
pr = [-1 1];
PR = repmat(pr,InputNum,1);

Network = newff(PR,[5  4 OutputNum],{'tansig' 'tansig' 'tansig'});

%%Entrenamiento
Network = TrainUsing_ICA_Fcn(Network,Xtr,Ytr);

%%Estimación
YtsNet = sim(Network,Xts')';

%%Desnormalizar
for ii = 1:size(YtsNet,2)
    predicted(:,ii) = UnNormalize_Fcn(YtsNet(:,ii),MinY(ii),MaxY(ii));
end

toc
end
```

C.5. Obtención de resultados

En este conjunto de *scripts* se aglutinan los resultados básicos del presente proyecto: regresión lineal, boxplots de error del pH, boxplots de coeficientes de error y comparaciones de valores de pH medido y estimado.

Los resultados obtenidos para estos *scripts* son extrapolables para las diferentes variedades de naranjas (*Bam*, *Blood* y *Thomson*) y mecanismos de regresión (SVM, MLP y ANN-ICA) utilizados en este proyecto.

```
%Script para obtener la regresion lineal de los pH estimados para la red
%neuronal SVM

clear all
clc
close all

%Carga de datos
load phReal.mat
load phEstimado_svm.mat

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Bam
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[R_Bam,m_Bam,b_Bam] = regression(phBam_real, phBam_estimado_svm);

figure(1);
plotregression(phBam_real, phBam_estimado_svm, 'Bam SVM Linear Regression')
xlabel('actual pH', 'FontSize', 10, 'FontWeight', 'bold')
ylabel(['mean estimated pH ~= ' num2str(m_Bam),'*actual pH + (',num2str(b_Bam),')'],...
    'FontSize', 10, 'FontWeight', 'bold')

%Guardar las figuras
nombre = strcat('regresionBam_svm', '.eps');
```

```
nombre2 = strcat('regresionBam_svm', '.jpeg');

saveas(gcf, nombre, 'epsc')
saveas(gcf, nombre2, 'jpg')
```

```
%Script para representar los boxplots actual ph - estimated ph de la
%variedad de naranja Bam

clear all
clc
close all

%Carga de datos
load phReal.mat
load ResultadosBam_svm.mat
load ResultadosBam_mlp.mat
load ResultadosBam_ica.mat

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Bam
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%svm
for i=1:size(phBam_real,2)
   idx = find(i == indices_Bam_svm);
   val = salidas_Bam_svm(idx);
   A{i} = phBam_real(i) - val;
end

%mlp
for i=1:size(phBam_real,2)
   idx = find(i == indices_Bam_mlp);
   val = salidas_Bam_mlp(idx);
   B{i} = phBam_real(i) - val;
end

%ann-ica
for i=1:size(phBam_real,2)
   idx = find(i == indices_Bam_ica);
   val = salidas_Bam_ica(idx);
   C{i} = phBam_real(i) - val;
end

tamano=get(0,'ScreenSize');
figure('position',[tamano(1) tamano(2) tamano(3) tamano(4)]);
y = num2cell(1:numel(A));
x = cellfun(@(x, y) [x(:) y*ones(size(x(:)))], A, y, 'UniformOutput', 0);
X = vertcat(x{:});

y = num2cell(1:numel(B));
x = cellfun(@(x, y) [x(:) y*ones(size(x(:)))], B, y, 'UniformOutput', 0);
Y = vertcat(x{:});

y = num2cell(1:numel(B));
x = cellfun(@(x, y) [x(:) y*ones(size(x(:)))], C, y, 'UniformOutput', 0);
Z = vertcat(x{:});

boxplot(X(:,1), X(:,2),'notch','off','Colors','r','whisker',1,'Symbol','+','OutlierSize',3)
hold on
boxplot(Y(:,1), Y(:,2),'notch','off','Colors','g','whisker',1,'Symbol','+','OutlierSize',3)
hold on
boxplot(Z(:,1), Z(:,2),'notch','off','Colors','b','whisker',1,'Symbol','+','OutlierSize',3)
ylim([-1.5 1.5])

ax = gca; outerpos = ax.OuterPosition;
ti = ax.TightInset;
left = outerpos(1) + 2*ti(1); bottom = outerpos(2) + 2*ti(2);
ax_width = outerpos(3) - 2*ti(1) - 2*ti(3); ax_height = outerpos(4) - 3*ti(2) - 3*ti(4);
ax.Position = [left bottom ax_width ax_height];
set(ax,'fontsize',6);
```

```
title('actual — estimated pH boxplots (Bam)','FontSize',10)
ylabel('actual — estimated pH','FontSize',10)
xlabel('Orange #','FontSize',10)
hold on
hline = refline([0 0]);
hline.Color = 'k';
hline.LineStyle = '—.';

boxes = findobj(gca,'Tag','Box');
legend([boxes(201,1),boxes(101,1),boxes(1,1)], {'SVM','MLP','ANN—ICA'},...
    'FontSize',10,'Location','northeast');

%Guardar las figuras
nombre = strcat('boxplotPH_Bam_ann', '.eps');
nombre2 = strcat('boxplotPH_Bam_ann', '.jpeg');

saveas(gcf, nombre, 'epsc')
saveas(gcf, nombre2, 'jpg')
```

```
%Valores y boxplots de errores MSE, MAE, R2, R, RMSE y SSE

clear all
clc
close all

%Carga de datos
load ResultadosBam_svm.mat
load ResultadosBlood_svm.mat
load ResultadosThomson_svm.mat
load ResultadosTotal_svm.mat
load ResultadosBam_mlp.mat
load ResultadosBlood_mlp.mat
load ResultadosThomson_mlp.mat
load ResultadosTotal_mlp.mat
load ResultadosBam_ica.mat
load ResultadosBlood_ica.mat
load ResultadosThomson_ica.mat
load ResultadosTotal_ica.mat

R2_Bam_svm = R_Bam_svm.^2;
RMSE_Bam_svm = sqrt(MSE_Bam_svm);
R2_Bam_mlp = R_Bam_mlp.^2;
RMSE_Bam_mlp = sqrt(MSE_Bam_mlp);
R2_Bam_ica = R_Bam_ica.^2;
RMSE_Bam_ica = sqrt(MSE_Bam_ica);

R2_Blood_svm = R_Blood_svm.^2;
RMSE_Blood_svm = sqrt(MSE_Blood_svm);
R2_Blood_mlp = R_Blood_mlp.^2;
RMSE_Blood_mlp = sqrt(MSE_Blood_mlp);
R2_Blood_ica = R_Blood_ica.^2;
RMSE_Blood_ica = sqrt(MSE_Blood_ica);

R2_Thomson_svm = R_Thomson_svm.^2;
RMSE_Thomson_svm = sqrt(MSE_Thomson_svm);
R2_Thomson_mlp = R_Thomson_mlp.^2;
RMSE_Thomson_mlp = sqrt(MSE_Thomson_mlp);
R2_Thomson_ica = R_Thomson_ica.^2;
RMSE_Thomson_ica = sqrt(MSE_Thomson_ica);

R2_Total_svm = R_Total_svm.^2;
RMSE_Total_svm = sqrt(MSE_Total_svm);
R2_Total_mlp = R_Total_mlp.^2;
RMSE_Total_mlp = sqrt(MSE_Total_mlp);
R2_Total_ica = R_Total_ica.^2;
RMSE_Total_ica = sqrt(MSE_Total_ica);

%Boxplot de errores

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
%R, R2
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

A(:,1) = R_Bam_svm;
A(:,2) = R2_Bam_svm;
A(:,3) = R_Blood_svm;
A(:,4) = R2_Blood_svm;
A(:,5) = R_Thomson_svm;
A(:,6) = R2_Thomson_svm;
A(:,7) = R_Total_svm;
A(:,8) = R2_Total_svm;

B(:,1) = R_Bam_mlp;
B(:,2) = R2_Bam_mlp;
B(:,3) = R_Blood_mlp;
B(:,4) = R2_Blood_mlp;
B(:,5) = R_Thomson_mlp;
B(:,6) = R2_Thomson_mlp;
B(:,7) = R_Total_mlp;
B(:,8) = R2_Total_mlp;

C(:,1) = R_Bam_ica;
C(:,2) = R2_Bam_ica;
C(:,3) = R_Blood_ica;
C(:,4) = R2_Blood_ica;
C(:,5) = R_Thomson_ica;
C(:,6) = R2_Thomson_ica;
C(:,7) = R_Total_ica;
C(:,8) = R2_Total_ica;

%Generacion del boxplot
tamano=get(0,'ScreenSize');
figure('position',[tamano(1) tamano(2) tamano(3) tamano(4)]);
boxplot(A(:,:),'notch','off','Colors','r','Labels',{'R(Bam)', 'R^2(Bam)',...
    'R(Blood)', 'R^2(Blood)', 'R(Thomson)', 'R^2(Thomson)', 'R(3-Variety)',...
    'R^2(3-Variety)'}, 'whisker', 1, 'Symbol','+');
hold on
boxplot(B(:,:),'notch','off','Colors','g','Labels',{'R(Bam)', 'R^2(Bam)',...
    'R(Blood)', 'R^2(Blood)', 'R(Thomson)', 'R^2(Thomson)', 'R(3-Variety)',...
    'R^2(3-Variety)'}, 'whisker', 1, 'Symbol','+');
hold on
boxplot(C(:,:),'notch','off','Colors','b','Labels',{'R(Bam)', 'R^2(Bam)',...
    'R(Blood)', 'R^2(Blood)', 'R(Thomson)', 'R^2(Thomson)', 'R(3-Variety)',...
    'R^2(3-Variety)'}, 'whisker', 1, 'Symbol','+');
boxes = findobj(gca,'Tag','Box');
legend([boxes(19,1),boxes(9,1),boxes(1,1)], {'SVM','MLP','ANN-ICA'},'Location','southeast');

%Guardar las figuras
nombre = strcat('boxplotErrores1_ann', '.eps');
nombre2 = strcat('boxplotErrores1_ann', '.jpeg');

saveas(gcf, nombre, 'epsc')
saveas(gcf, nombre2, 'jpg')

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%MSE, MAE y RMSE
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

D(:,1) = MSE_Bam_svm;
D(:,2) = MAE_Bam_svm;
D(:,3) = RMSE_Bam_svm;
D(:,4) = MSE_Blood_svm;
D(:,5) = MAE_Blood_svm;
D(:,6) = RMSE_Blood_svm;
D(:,7) = MSE_Thomson_svm;
D(:,8) = MAE_Thomson_svm;
D(:,9) = RMSE_Thomson_svm;
D(:,10) = MSE_Total_svm;
D(:,11) = MAE_Total_svm;
D(:,12) = RMSE_Total_svm;
```

```

E(:,1) = MSE_Bam_mlp;
E(:,2) = MAE_Bam_mlp;
E(:,3) = RMSE_Bam_mlp;
E(:,4) = MSE_Blood_mlp;
E(:,5) = MAE_Blood_mlp;
E(:,6) = RMSE_Blood_mlp;
E(:,7) = MSE_Thomson_mlp;
E(:,8) = MAE_Thomson_mlp;
E(:,9) = RMSE_Thomson_mlp;
E(:,10) = MSE_Total_mlp;
E(:,11) = MAE_Total_mlp;
E(:,12) = RMSE_Total_mlp;

F(:,1) = MSE_Bam_ica;
F(:,2) = MAE_Bam_ica;
F(:,3) = RMSE_Bam_ica;
F(:,4) = MSE_Blood_ica;
F(:,5) = MAE_Blood_ica;
F(:,6) = RMSE_Blood_ica;
F(:,7) = MSE_Thomson_ica;
F(:,8) = MAE_Thomson_ica;
F(:,9) = RMSE_Thomson_ica;
F(:,10) = MSE_Total_ica;
F(:,11) = MAE_Total_ica;
F(:,12) = RMSE_Total_ica;

%Generacion del boxplot
tamano=get(0,'ScreenSize');
figure('position',[tamano(1) tamano(2) tamano(3) tamano(4)]);
boxplot(D(:,:),'notch','off','Colors','r','Labels',{'MSE(Bam)', 'MAE(Bam)',...
    'RMSE(Bam)', 'MSE(Blood)', 'MAE(Blood)', 'RMSE(Blood)', 'MSE(Thomson)',...
    'MAE(Thomson)', 'RMSE(Thomson)', 'MSE(3-Variety)', 'MAE(3-Variety)',...
    'RMSE(3-Variety)'}, 'whisker', 1, 'Symbol','+');
hold on
boxplot(E(:,:),'notch','off','Colors','g','Labels',{'MSE(Bam)', 'MAE(Bam)',...
    'RMSE(Bam)', 'MSE(Blood)', 'MAE(Blood)', 'RMSE(Blood)', 'MSE(Thomson)',...
    'MAE(Thomson)', 'RMSE(Thomson)', 'MSE(3-Variety)', 'MAE(3-Variety)',...
    'RMSE(3-Variety)'}, 'whisker', 1, 'Symbol','+');
hold on
boxplot(F(:,:),'notch','off','Colors','b','Labels',{'MSE(Bam)', 'MAE(Bam)',...
    'RMSE(Bam)', 'MSE(Blood)', 'MAE(Blood)', 'RMSE(Blood)', 'MSE(Thomson)',...
    'MAE(Thomson)', 'RMSE(Thomson)', 'MSE(3-Variety)', 'MAE(3-Variety)',...
    'RMSE(3-Variety)'}, 'whisker', 1, 'Symbol','+');
boxes = findobj(gca,'Tag','Box');
legend([boxes(26,1),boxes(13,1),boxes(1,1)], {'SVM','MLP','ANN-ICA'},'Location','southeast');

%Guardar las figuras
nombre = strcat('boxplotErrores2_ann', '.eps');
nombre2 = strcat('boxplotErrores2_ann', '.jpeg');

saveas(gcf, nombre, 'epsc')
saveas(gcf, nombre2, 'jpg')

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%SSE
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

G(:,1) = SSE_Bam_svm;
G(:,2) = SSE_Blood_svm;
G(:,3) = SSE_Thomson_svm;
G(:,4) = SSE_Total_svm;

H(:,1) = SSE_Bam_mlp;
H(:,2) = SSE_Blood_mlp;
H(:,3) = SSE_Thomson_mlp;
H(:,4) = SSE_Total_mlp;

I(:,1) = SSE_Bam_ica;
I(:,2) = SSE_Blood_ica;
```

```
I(:,3) = SSE_Thomson_ica;
I(:,4) = SSE_Total_ica;

%Generacion del boxplot
figure(3);
boxplot(G(:,:), 'notch', 'off','Colors','r','Labels', {'SSE(Bam)', 'SSE(Blood)',...
    'SSE(Thomson)', 'SSE(3-Variety)'}, 'whisker', 1, 'Symbol','+');
hold on
boxplot(H(:,:), 'notch', 'off','Colors','g','Labels', {'SSE(Bam)', 'SSE(Blood)',...
    'SSE(Thomson)', 'SSE(3-Variety)'}, 'whisker', 1, 'Symbol','+');
hold on
boxplot(I(:,:), 'notch', 'off','Colors','b','Labels', {'SSE(Bam)', 'SSE(Blood)',...
    'SSE(Thomson)', 'SSE(3-Variety)'}, 'whisker', 1, 'Symbol','+');
boxes = findobj(gca,'Tag','Box');
legend([boxes(9,1),boxes(5,1),boxes(1,1)], {'SVM','MLP','ANN-ICA'},'Location','northeast');

%Guardar las figuras
nombre = strcat('boxplotErrores3_ann', '.eps');
nombre2 = strcat('boxplotErrores3_ann', '.jpeg');

saveas(gcf, nombre, 'epsc')
saveas(gcf, nombre2, 'jpg')
```

```
%Script para comparar los valores de ph real y estimado (media) para la
%variedad de naranja Bam

clear all
clc
close all

%Carga de datos
load phReal.mat
load phEstimado_svm.mat
load phEstimado_mlp.mat
load phEstimado_ica.mat

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Bam
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Representacion comparativa de los valores real y estimado del pH
tamano=get(0,'ScreenSize');
figure('position',[tamano(1) tamano(2) tamano(3) tamano(4)]);
scatter([1:100], phBam_real(1:100), 'k')
hold on
scatter([1:100], phBam_estimado_svm(1:100),'r', 's')
hold on
scatter([1:100], phBam_estimado_mlp(1:100),'g', '*')
hold on
scatter([1:100], phBam_estimado_ica(1:100),'b', 'd')

xticks([0:5:100])
title('Bam orange measured vs estimated pH','FontSize',10)
ylabel('pH value','FontSize',10)
xlabel('Orange #','FontSize',10)
legend(findobj(gca, 'Tag', 'Box'), 'Measured pH', 'SVM Estimated pH',...
    'MLP Estimated pH''ANN-ICA Estimated pH', 'Location', 'northeast')

%Guardar las figuras
nombre = strcat('valoresPH_Bam', '.eps');
nombre2 = strcat('valoresPH_Bam', '.jpeg');

saveas(gcf, nombre, 'epsc')
saveas(gcf, nombre2, 'jpg')
```