ANÁLISIS DE DATOS AGROPECUARIOS

IVÁN RAMÍREZ-MORALES / BERTHA MAZON-OLIVO

Editorial UTMACH

REDES 2017
COLECCIÓN EDITORIAL

Análisis de Datos Agropecuarios

Iván Ramírez-Morales
Bertha Mazon-Olivo

Coordinadores

Primera edición en español, 2018

Este texto ha sido sometido a un proceso de evaluación por pares externos con base en la normativa editorial de la UTMACH

Ediciones UTMACH

Gestión de proyectos editoriales universitarios

302 pag; 22X19cm - (Colección REDES 2017)

Título: Análisis de Datos Agropecuarios. / Iván Ramírez-Morales / Bertha Mazon-Olivo (Coordinadores)

ISBN: 978-9942-24-120-7

Publicación digital

Título del libro: Análisis de Datos Agropecuarios.

ISBN: 978-9942-24-120-7

Comentarios y sugerencias: editorial@utmachala.edu.ec

Diseño de portada: MZ Diseño Editorial

Diagramación: MZ Diseño Editorial

Diseño y comunicación digital: Jorge Maza Córdova, Ms.

Km. 5 1/2 Vía Machala Pasaje

www.utmachala.edu.ec

Machala - Ecuador

Índice

Introducción

El análisis de datos es un proceso complejo que trata de encontrar patrones útiles y relaciones entre los datos a fin de obtener información sobre un problema específico y de esta manera tomar decisiones acertadas para su solución.

Las técnicas de análisis de datos que son exploradas en el presente libro son actualmente utilizadas en diversos sectores de la economía. En un inicio, fueron empleadas por las grandes empresas a fin de incrementar sus rendimientos financieros.

El libro se basa en la aplicación de la especialización inteligente, de este modo, gracias al trabajo colaborativo, se combina al sector agropecuario con las tecnologías, matemáticas, estadística y las ciencias computacionales, para la optimización de los procesos productivos.

La idea de descubrir la información oculta en las relaciones entre los datos, incentiva a encontrar aplicaciones para el sector agropecuario, por ejemplo los obtenidos de una producción avícola, o los datos que se generan durante los procesos de fermentación, los parámetros físicos y químicos del suelo, del agua y de las plantas, los datos de sensores, de espectrometría, entre otros.

En la actualidad, este sector se ha mantenido con su producción habitual sin un destacado repunte ni diferenciación, a pesar de existir herramientas científicas que han permitido desarrollar dispositivos tecnológicos y sus aplicaciones.

Este libro ha sido el resultado de la sistematización de las experiencias individuales de un equipo humano con objetivos comunes y una historia académica multidisciplinar, cuyos hallazgos de investigación han sido publicados en revistas científicas y conferencias de alto impacto. El área temática sobre la que se centra este texto es en técnicas de extracción, procesamiento y análisis de datos del ámbito agropecuario, se combinan para entregar al lector una obra de calidad y alto valor científico.

Así, el presente libro está concebido desde diferentes puntos de vista de profesionales agrónomos, informáticos, electrónicos, matemáticos, estadísticos y empresarios. Todos buscan un objetivo en común: "descubrir el conocimiento oculto en los datos que proporcione una ventaja competitiva". Se aborda el ciclo completo del proceso de obtención de conocimiento a partir de datos crudos del sector agropecuario, con la finalidad de apoyar la toma de decisiones. Este ciclo involucra procesos de: selección de los datos (extracción, comunicación, almacenamiento), pre-procesamiento, transformación, aplicación de modelos y/o técnicas de análisis, presentación e interpretación de resultados. El enfoque temático del libro es el siguiente:

Capítulo 1: Ciencia de Datos en el sector Agropecuario.- En este capítulo se aborda una revisión desde los inicios del análisis de datos en el sector agropecuario hasta el progreso actual que se ha dado en esta área del conocimiento que se considera como la nueva revolución en la agricultura y la ganadería de precisión.

Capítulo 2: Obtención de datos en sistemas agropecuarios.- El enfoque del capítulo es la generación de datos crudos en los sistemas agropecuarios, aplicando métodos y técnicas básicas donde se registran información de: número de unidades producidas, cantidad de nutrientes, variables climáticas, muestreo y monitoreo de organismos vivos, entre otros.

Capítulo 3: Internet de las cosas (IoT).- Este capítulo aborda los sistemas de telemetría para obtención de datos y control de dispositivos, aplicando tecnologías como: redes de sensores inalámbricos (dispositivos electrónicos, sensores, actuadores y puertas de enlace), protocolos de comunicación, centros de procesamiento de datos (cloud computing) y aplicaciones IoT para el sector agropecuario.

Capítulo 4: Matemáticas aplicadas al sector agropecuario.- Este capítulo explica los procedimientos para la creación de modelos matemáticos determinísticos que representen procesos asociados al sector agropecuario, como una alternativa de solución en la ingeniería.

Capítulo 5: Estadística básica con datos agropecuarios.- El capítulo se enfoca en los atributos, escalas de medición de las variables, su influencia en la elección del procedimiento estadístico a desarrollar, así como, el papel de las medidas de resumen, estimación puntual y prueba de hipótesis en la investigación científica.

Capítulo 6: Estadística predictiva con datos agropecuarios.- El capítulo considera las principales técnicas de la estadística avanzada aplicada al sector agropecuario, con el propósito de establecer predicciones que permita tomar mejores decisiones.

Capítulo 7: Inteligencia de negocios en el sector agropecuario.- El capítulo comprende la obtención de conocimiento a partir de datos crudos con la finalidad de apoyar la toma de decisiones en empresas del sector agropecuario. Involucra procesos de extracción, transformación y almacenamiento de datos en nuevos almacenes (Data warehouse - Big Data), distribución y análisis de la información con técnicas: multidimensional OLAP y tableros de control (dashboards).

Capítulo 8: Inteligencia Artificial aplicada a datos agropecuarios.- El capítulo trata sobre las principales técnicas de machine learning aplicadas a los datos agropecuarios, entre éstas se destacan: las redes de neuronas artificiales, máquinas de soporte de vectores, vecinos más cercanos, análisis de componentes principales, entre otros.

08 Capítulo

Inteligencia Artificial aplicada a datos agropecuarios

Iván Ramírez-Morales, Eduardo Tusa; Daniel Rivero

La inteligencia artificial (IA), podría sonar aún como un término de ciencia ficción, sin embargo muchas personas no se percatan de que están utilizándose cada vez más en actividades de la vida cotidiana. Los asistentes personales como

Iván Ramírez-Morales: Doctor en Medicina Veterinaria y Zootecnia por la Universidad Agraria de la Habana, Máster en Desarrollo Comunitario por la Universidad Nacional de Loja y está finalizando su Doctorado en TIC por la Universidade A Coruña, ha realizado varios cursos en Brasil, Japón, Perú y Argentina. Fué Oficial de Territorio del Programa Marco ART/PNUD de la ONU, y Director de Planificación del Gobierno Provincial de El Oro. Actualmente es Profesor Titular en la Universidad Técnica de Machala, su área de investigación se centra en el uso de tecnologías para el mejoramiento de la productividad agropecuaria, Cuenta a la fecha más de 10 publicaciones indexadas, varias de ellas en revistas de alto impacto en los índices de JCR y SJR.

Eduardo Tusa: Ingeniero Electrónico (Magna Cum Laude) con una Subespecialización en Matemáticas de la Universidad San Francisco de Quito. Su cuarto año de formación de pregrado fue realizado en la Universidad de Illinois en Urbana - Champaign, USA. Máster en Visión, Imagen y Robótica (con distinción) de la Universidad de Borgoña (Francia), la Universidad de Girona (España) y la Universidad Heriot-Watt (Reino Unido). Es docente de la Unidad Académica de Ingeniería Civil de la Universidad Técnica de Machala, donde ha impartido las asignaturas de Programación en MATLAB, Informática, Nuevas Tecnologías de la Información y Comunicación, Cálculo Integral, Ecuaciones Diferenciales, Matemática Avanzada, Probabilidad y Estadístic

Daniel Rivero: Ingeniero en Informática por la Universidad de A Coruña, y Doctor en el área de conocimiento en Ciencias de la Información e Inteligencia Artificial. Trabaja como Profesor Contratado Doctor en el Departamento de Tecnologías de la Información y las Comunicaciones de la citada Universidad. Su área de investigación incluye las Redes de Neuronas Artificiales, Computación Evolutiva, y en general la aplicación de técnicas de Machine Learning en distintos entornos. Fruto de este trabajo de investigación, ha publicado una gran cantidad de artículos en distintas revistas indexadas con índice de impacto JCR, así como distintos libros, como autor, editor o autor de capítulos, y comunicaciones a congresos. Por otra parte, ha colaborado también en un gran número de proyectos de investigación financiados de forma competitiva.

Siri o Cortana, los vehículos autónomos, predicción de fraudes, o de condiciones idóneas de mercado, recomendaciones sobre tendencias, son entre otras aplicaciones de uso común.

Una rama de la inteligencia artificial conocida como aprendizaje automático o aprendizaje máquina (machine learning - ML), se refiere a los algoritmos computacionales que son capaces de realizar acciones complejas, sin que éstos hayan sido explícitamente programados para ello, si no que estos algoritmos son más bien, entrenados para realizar esta tarea.

En tareas muy complejas, esta particularidad es indispensable para que un computador sea capaz de realizar una tarea; por ejemplo, programar un juego como el ajedrez implica una complejidad que fue calculada por Shannon (1950) como 10^{120}, como punto de comparación, se dice que todos los átomos del universo son 6^{79}, en estos casos es mejor utilizar algoritmos que aprenden a partir de ejemplos.

Los algoritmos de aprendizaje máquina tienen la capacidad de generalizar un comportamiento de respuesta a partir de una información suministrada en forma de ejemplos, durante el proceso de entrenamiento.

Actualmente se está utilizando algoritmos de aprendizaje profundo que permite a las máquinas aprender de una manera muy similar a como lo hacen los humano.

Aunque estas tecnologías no están alejadas de la potencial aplicación en el sector agropecuario, existe todavía un rezago en cuanto a su uso por parte de los profesionales de este importante sector de la economía. En referencia al objetivo principal de este libro, en este capítulo se explora la aplicabilidad de algunas técnicas de IA enfocadas al análisis de información de ámbito agropecuario.

Para los profesionales que se desenvuelven en el ámbito agropecuario, es común utilizar estadísticas descriptivas para el procesamiento de sus datos. Estas técnicas permiten una adecuada comprensión de la información existente en el dato, sin embargo no siempre son capaces de extraer con

suficiente exhaustividad, la información relevante que apoye a la toma de decisiones por parte de los administradores y técnicos de producción.

En las fincas agropecuarias cada día se genera una gran cantidad de información, aunque por la naturaleza misma de los medios con los que se obtiene, esta información generalmente tiene mucho ruido, es decir, datos erróneos, o irregulares, que pueden enmascarar el conocimiento contenido en la relación de la información. Es por ello que se hace necesario utilizar nuevas técnicas de análisis bajo un enfoque de aprendizaje automático.

Tipos de aprendizaje automático

Aprendizaje no supervisado

El aprendizaje automático no supervisado, consiste en asignar una máquina la tarea de inferir una función que describa la estructura oculta de los datos, dado que éstos no han sido previamente etiquetados. En este caso no se cuenta con la posibilidad de evaluar fácilmente la exactitud del resultado de la función inferida.

En este tipo de algoritmos, la salida se asocia con el grado de similitud entre las características de entrada, es decir que el aprendizaje se centra en las asociaciones que ocurren en un conjunto de datos tratando de encontrar cualquier tipo de regularidad en los datos.

Estas técnicas suelen ser utilizadas para agrupar datos según su criterio de similitud. Además son muy utilizadas para visualización de datos ya que permiten reducir a dos o tres dimensiones, datos multidimensionales. Precisamente por esta propiedad, los algoritmos no supervisados suelen ser utilizados para extracción de características previo al entrenamiento con alguno de los algoritmos de aprendizaje supervisado.

Aprendizaje supervisado

El aprendizaje supervisado consiste en el descubrimiento de patrones válidos a partir de conjuntos de datos de entrenamiento que han sido previamente etiquetados. En el aprendizaje supervisado, cada ejemplo tiene un un objeto de entrada y un valor de salida deseada.

Un algoritmo de aprendizaje supervisado analiza los datos de entrenamiento y produce una función inferida, que puede ser utilizado para el mapeo de nuevos ejemplos. Una correcta selección de ejemplos, permitirá el algoritmo para determinar correctamente las etiquetas de clase para nuevas instancias. Esta capacidad de inferir la clase de datos nuevos, se conoce como generalización.

Para entrenar un algoritmo con técnicas de aprendizaje supervisado, es necesario en primer lugar identificar el conjunto de datos para el entrenamiento. Este tiene que ser representativo del universo de datos y debe haber sido etiquetado y revisado por expertos en el área.

La precisión va depender en gran medida de las características del vector de entrada, estas características deben contener suficiente información sobre el patrón de entrada para que sea capaz de predecir con precisión la salida deseada. Debido a un efecto que se denomina la "maldición de la multidimensionalidad", el vector de entrada no debe tener demasiadas características.

Los algoritmos de aprendizaje automático suelen tener varios parámetros que deben ser ajustados durante el proceso de entrenamiento, estos parámetros permiten modelar de mejor manera y elevan la precisión de la función aprendida.

Existe una gran variedad de algoritmos de aprendizaje supervisado. No existe uno que sea válido para todos los problemas, cada uno tiene sus particularidades. La selección del algoritmo idóneo se realiza habitualmente en un proceso que es empírico y requiere de muchas pruebas cuyo resultado final es la optimización del modelo.

Los problemas que se abordan con aprendizaje supervisado suelen ser generalmente de dos tipos: clasificación y regresión. Es importante diferenciar ambos tipos, ya que su comprensión es básica para entender el funcionamiento de las técnicas. La clasificación consiste en la asignación de una clase o tipo de acuerdo a sus características; por ejemplo de qué raza es un animal, o de qué variedad es una planta. La regresión por otro lado, busca predecir un valor cuantificable, por ejemplo, cuántos litros de leche producirá una vaca, o cuántos quintales por hectárea se obtendrán de una parcela. Como se puede apreciar, la clasificación tiene el objetivo de predecir valores discretos, mientras que la regresión predice valores continuos.

Técnicas de ML más utilizadas

En este apartado se realizará una breve descripción de las técnicas más comunes utilizadas en aprendizaje automático. Además se presentan varios ejemplos de aplicaciones reales en las que los autores han implementado estas técnicas, así como otros ejemplos ilustrativos llevados a cabo por otros autores.

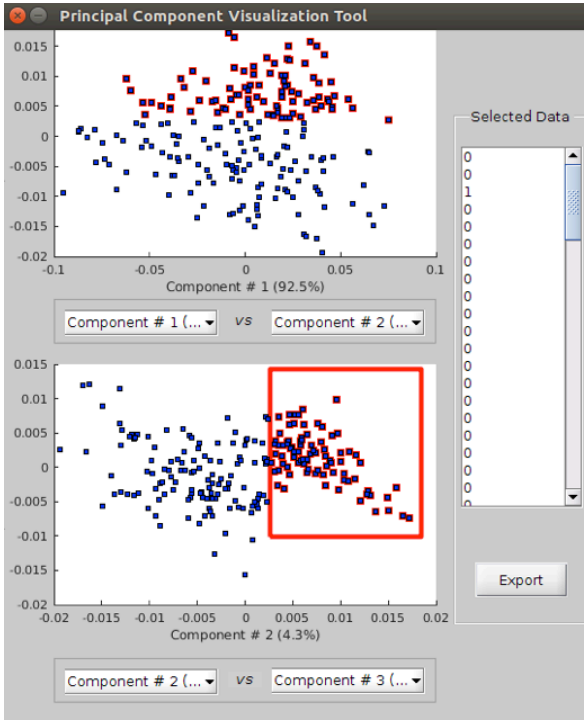
Análisis de componentes principales

La técnica de análisis de componentes principales (Principal Component Analysis - PCA) es una técnica de aprendizaje no supervisado. Es comúnmente utilizada para reducir la dimensión de un conjunto de datos. La variabilidad de los datos se conserva y el número de variables o características se reduce.

Con esta técnica se busca la mejor representación de los mínimos cuadrados de los datos, de esta manera cuando en un patrón de datos existen algunas variables posiblemente correlacionadas, el algoritmo devuelve un nuevo conjunto de valores que no tienen una correlación lineal entre sí, a estos valores se les llama componentes principales.

El análisis de componentes principales, retiene los valores numéricos que explican en mayor proporción la varianza del conjunto de datos, e ignora aquellos que tienen menor influencia en la varianza. Comúnmente los componentes principales contienen lo más importante de la información original, sin embargo se pierde la representación original del dato.

Gráfico 8.1 Gráfico de un análisis de componentes principales para el diagnóstico de mastitis bovina utilizando espectrometría NIR.



En el Gráfico 8.1 se observa un gráfico del componente principal 1 versus el componente principal 2 y otro del componente principal 2 versus el componente principal 3. Se puede apreciar que los datos seleccionados pertenecen a la misma clase 0 (Infección Negativa) mientras que los datos no seleccionados pertenecen a la clase 1 (Infección Positiva). En la imagen también se observa que entre los seleccionados de la clase 0, aparece un dato mal clasificado.

En la agricultura de precisión se cuenta con grandes volúmenes de información georeferenciada, el análisis PCA clásico aplicado a este tipo de datos es capaz de evaluar las propiedades del suelo y el rendimiento del cultivo. De esta manera se detectan correlaciones entre variables que permiten la consolidación y homogeneización de zonas dentro de los lotes. Existen varias aplicaciones de esta técnica que serán abordadas más adelante, en general es una técnica que por su sencillez ha logrado una buena difusión entre la comunidad científica.

K Vecinos Más Cercanos

También conocida como k-NN por su traducción en inglés (k Nearest Neighbors) Esta es una técnica muy versátil puesto que puede ser utilizada tanto para aprendizaje no supervisado, como para aprendizaje supervisado. En el caso no supervisado, se ha utilizado en clasificación y en regresión.

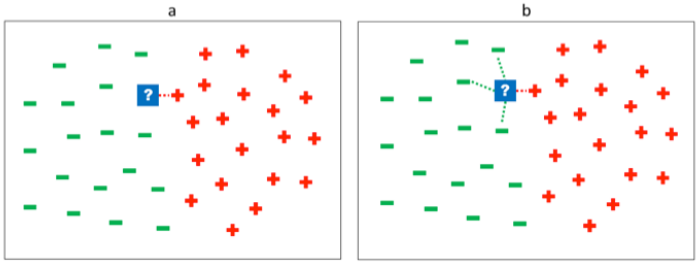
Su aplicación más habitual es en clasificación. Su funcionamiento consiste en asignar la clase a una patrón desconocido, según la clase que tengan los patrones conocidos más cercanos. El número de vecinos k normalmente es ajustado en un proceso de optimización con la finalidad de mejorar la precisión del clasificador, este ajuste especialmente importante cuando la muestra desconocida está rodeada de muestras conocidas que tienen diferentes clases.

Se genera una regla para la clasificación de acuerdo a la que la clase asignada será aquella que tenga la mayor parte de sus k vecinos más próximos.

El Gráfico 8.2 muestra la regla de decisión k-NN para k = 1 (a) y para k = 4 (b). El conjunto de datos ha sido etiquetado de tal manera que la clase negativa corresponde a la ausencia de una enfermedad y la clase positiva a la presencia de esta. En esta figura se ilustra cómo influye en la decisión el número k en la decisión de la asignación de una clase. En el primer caso que se puede observar en el Gráfico 8.2a, la muestra desconocida está fue clasificada con sólo un vecino más cercano, por lo tanto que de acuerdo a esta regla de

decisión, pertenecería a la clase positiva. En el segundo caso que se observa en el Gráfico 8.2b, se utilizan cuatro vecinos del conjunto de entrenamiento para clasificar la misma muestra desconocida, en esta ocasión, tres de los vecinos más cercanos pertenecen a la clase negativa.

Gráfico 8.2 Reglas de decisión en k-NN de acuerdo al número k



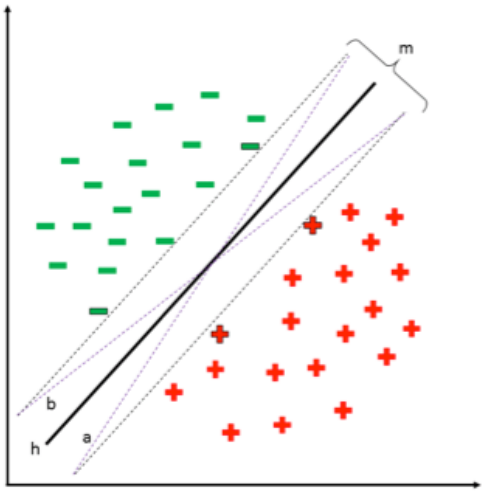
Esta técnica ha sido utilizada en clasificación y regresión debido en una amplia variedad de campos de la ciencia debido a su simplicidad y precisión. Se ha empleado en la predicción de enfermedades, pronóstico del clima, en la detección de deficiencias de nutrientes, entre otras aplicaciones.

Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (Support Vector Machine - SVM), están entre las técnicas más utilizadas en aprendizaje máquina. Se relacionan principalmente con la resolución de problemas de clasificación y regresión. Los principios de las SVM fueron desarrollados por Vapnik y colaboradores (1997). El enfoque original estaba dirigido a resolver problemas de clasificación binaria, sin embargo su aplicación se ha extendido a tareas de clasificación múltiple, aprendizaje no supervisado y regresión.

Las SVM tratan de obtener modelos que minimicen el riesgo estructural de cometer errores ante datos futuros. Su funcionamiento básico consiste en la separación del conjunto de datos en dos clases distintas gracias a un hiperplano definido en un espacio adecuado.

Gráfico 8.3 Representación del hiperplano y margen óptimos del modelo (m).

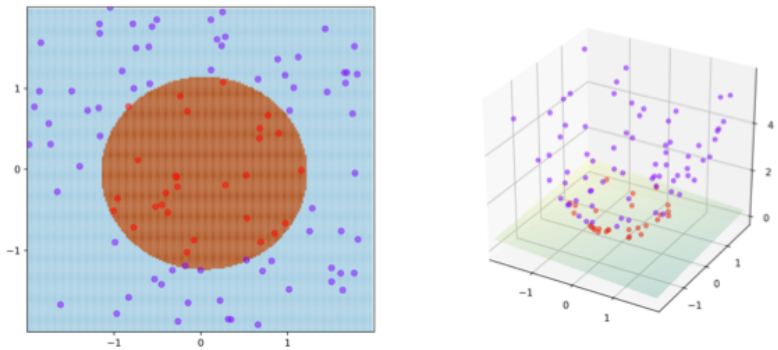


En el Gráfico 8.3 se ilustra los conceptos de hiperplano y margen óptimo del modelo. Como se puede observar en el espacio del margen, pueden existir un sinnúmero de hiperplanos alternativos, en el Gráfico se ilustran dos hiperplanos alternativos posibles (a y b). El hiperplano óptimo usado para separar las dos clases se define a partir de una pequeña cantidad de datos del conjunto de entrenamiento llamados vectores de soporte, que en el gráfico se encuentran sombreados. Estos vectores de soporte son los que determinan el margen del modelo. La elección del mejor hiperplano fue resuelta por Vapnik y Kotz (1982) con el planteamiento de que el hiperplano óptimo es definido como la función de decisión lineal con el máximo margen entre los vectores de soporte de las dos clases.

Sin embargo, en la mayoría de problemas del mundo real, los datos no son linealmente separables y por este motivo es necesario recurrir a estrategias como la identificación de otras dimensiones de separación. Las funciones kernel, son utilizadas para transformar el espacio original multidimensional, en otro espacio en el que las clases sean linealmente separables. En la práctica, las máquinas de soporte vectorial son entrenadas usando distintos kernels para seleccionar

aquel que tenga el mejor desempeño para el problema planteado. Entre los kernel más utilizados están el polinomial y el gaussiano (función de base radial), éste último cuenta con un parámetro sigma (σ) que ajusta el tamaño del kernel.

Gráfico 8.4 SVM con un kernel gaussiano φ ((a, b)) = (a, b, a2 + b2) (Shiyu, Nov, 13, 2016)



En el Gráfico 8.4 se observa cómo los datos de entrenamiento se trasladan a un nuevo espacio de 3 dimensiones en el que un hiperplano es capaz de separarlos linealmente con mayor facilidad.

La búsqueda de parámetros óptimos de una SVM es fundamental en la construcción de un modelo de predicción para que sea preciso y estable. Los parámetros del kernel son ajustables en las SVM para controlar la complejidad de la hipótesis resultante y evitar el sobreajuste del modelo.

Las SVM también pueden ser utilizadas en problemas regresión, esta versión fue propuesta por Vapnik, Golowich y Smola (1997). El método se llama Support Vector Regression (SVR). En este caso, el modelo depende únicamente de los vectores de soporte, ya que la función de pérdida para la construcción del modelo no considera los puntos que se encuentren fuera del margen, asimismo la función de pérdida ignora cualquier los datos que estén cerca al modelo de predicción, dentro de un umbral ε.

Las SVM se han aplicado en varios campos como series temporales, finanzas, aproximaciones de ingeniería, programación cuadrática convexa, clasificación binaria, regresión multivariada, entre otros.

Redes de neuronas artificiales

Las Redes de Neuronas Artificiales (Artificial Neural Network - ANN) están inspiradas en el funcionamiento del sistema nervioso de los animales, cuyas redes de neuronas biológicas poseen bajas capacidades de procesamiento de forma individual, sin embargo su capacidad cognitiva se sustenta en la conectividad de éstas. De modo similar al mecanismo biológico, las ANN son capaces de realizar tareas complejas de clasificación, identificación, diagnóstico, optimización y predicción, gracias a la conectividad de unidades de procesamiento sencillas.

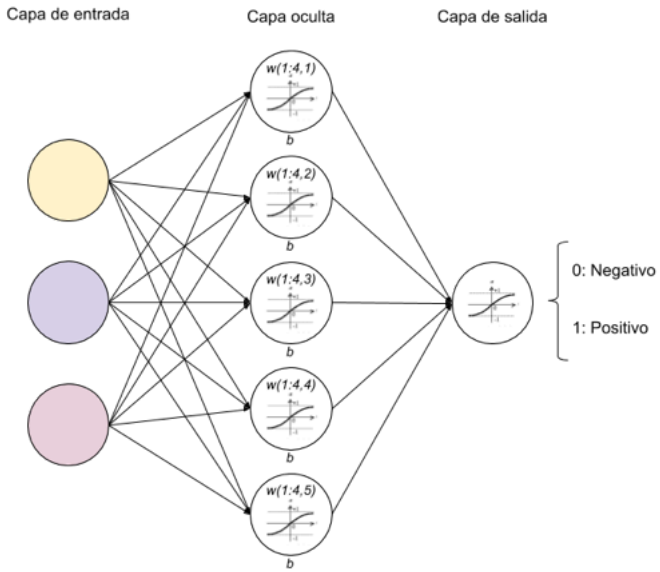
Las ANNs son algoritmos tanto de aprendizaje no supervisado, como de aprendizaje supervisado. Se pueden utilizar para agrupamiento, clasificación y regresión. La organización de las neuronas, permite aprender de los patrones y generalizar hacia nuevas entradas de datos.

Las redes de neuronas artificiales han atraído especial atención en los últimos años, sin embargo fueron McCulloch and Pitts, (1943) quienes presentaron el primer modelo de neurona artificial. Se plantea que las redes neuronales de múltiples capas ocultas, son capaces de aproximar cualquier función medible, por lo que se las considera aproximadores universales.

El perceptrón multicapa (Multilayer Perceptron - MLP) es un tipo de ANN cuyas neuronas están organizadas en capas. Las conexiones en esta red se realizan únicamente entre capas consecutivas. De manera general un MLP tiene una capa de entrada, una o múltiples capas ocultas y una capa de salida. La función de transferencia en las neuronas de la capa oculta y de la capa de salida usualmente es una sigmoidea, sin embargo, pueden estar presentes otras funciones como las lineales, las no lineales o las escalonadas.

En el Gráfico 8.5 se muestra una estructura característica del MLP, los patrones de entrada se proporcionan a la red a través de una capa que simplemente envía esta información a la siguiente capa. El procesamiento y la extracción de la información es realizado en las capas ocultas y en la capa de salida. Cada neurona recibe señales de salida de las neuronas en la capa anterior y envía su señal de salida a las neuronas de la capa siguiente. La capa de salida, recibe las entradas de las neuronas y de acuerdo a un cálculo probabilístico asigna la clase a la que pertenece el ejemplo desconocido. Los MLP pueden ser entrenados tanto para clasificación como para regresión.

Gráfico 8.5 Representación de un MLP con una capa oculta



Uno de los métodos más usados para optimizar el proceso de entrenamiento de un MLP busca localizar el error mínimo utilizando una técnica de gradiente descendiente. En primer lugar se inicializan con valores aleatorios los pesos y los bias de las neuronas, luego, se determina la dirección de la pendiente más pronunciada (gradiente descendiente), se

modifican los pesos, y se re-calcula el gradiente hasta llegar a un valor mínimo de la función.

Para mejorar el desempeño de una ANN es necesario seleccionar una arquitectura adecuada, esto consiste en determinar el número de capas ocultas, el número de neuronas y la forma como estarán interconectadas. La arquitectura de red va a depender del problema a resolver, y no existe una regla o método que permita decidir cuál es la mejor. Generalmente la selección de la mejor arquitectura, resulta de un proceso empírico, en el que es necesario probar distintas alternativas hasta que se encuentra una que proporcione buenos resultados.

El interés en el uso de las redes neuronales va en aumento gracias a su naturaleza paralela, lo que hace que puedan aumentar su velocidad de cálculo, por este motivo ha sido aplicada en una gran variedad de aplicaciones, entre las que destaca la predicción de precios futuros de productos exportables, estimación de la humedad del suelo, la predicción de los rendimientos de cultivos, la elaboración de mapas digitales de territorio, entre otras aplicaciones.

Redes neuronales profundas

También conocidas como Deep Neural Networks (DNN), se distinguen de las redes neuronales comunes por su mayor profundidad, es decir, el número de capas ocultas a través de las cuales pasan los datos en un proceso de múltiples etapas de reconocimiento de patrones.

Las redes neuronales tradicionales se tienen hasta dos capas ocultas. Cuando se tiene más de tres capas, se considera DNN. En las DNN, cada capa se entrena con un conjunto distinto de características generadas como salidas de la capa anterior. Cuanto más se avanza hacia la red neuronal, la más compleja de las características de sus nodos pueden reconocer, ya que se agregan y se recombinan características de la capa anterior.

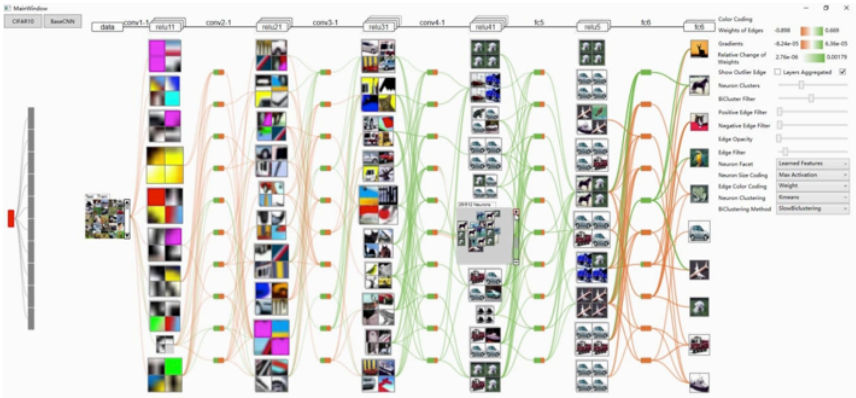
En el Gráfico 8.6 se puede observar una representación de una DNN, en la primera capa oculta se extraen característi-

cas básicas como bordes, en las siguientes se tiene niveles mayores de complejidad como formas, en la siguiente se cuenta con representaciones más precisas del objeto a clasificar o detectar. Este avance creciente en la complejidad y la abstracción se conoce como jerarquía de características.

Las DNN son capaces de manejar conjuntos de datos muy grandes, de muy alta dimensión con miles de millones de parámetros que pasan a través de funciones no lineales.

Algo interesante es que las DNN son capaces de descubrir las estructuras latentes en datos no etiquetados. Un aspecto importante dado que la gran mayoría de los datos en el mundo no tienen etiqueta. Es decir que mediante redes profundas, es posible agrupar de acuerdo a su similitud conjuntos de datos de millones de imágenes, y de esta manera por ejemplo, contar automáticamente con un sistema que agrupe imágenes de plantas sin enfermedades, y distintos grupos según el tipo de enfermedad.

Gráfico 8.6 Representación de una Red Neuronal Profunda (Liu et al, 2017)



Aplicaciones en el ámbito agropecuario

En el sector agropecuario, la inteligencia artificial tiene un gran potencial. Principalmente por su capacidad para el reconocimiento de patrones. Esta característica permite aplicaciones tales como la clasificación o estimación de paráme-

tros a partir de matrices numéricas, la detección temprana de problemas de producción utilizando series temporales, el análisis de imágenes para clasificación, el análisis de sonidos para detección de enfermedades o el análisis de videos para determinación de patrones de comportamiento.

La gama de posibles aplicaciones en el ámbito agropecuario es variada, en esta sección se realiza una descripción de algunas aplicaciones que han permitido optimizar algún proceso en el ámbito agropecuario con la consecuente mejora de los resultados económicos de las empresas.

Análisis de señales

En el sector agropecuario es cada vez más común la generación de datos a partir de sensores, estos equipos generan señales que en ocasiones son muy complejas para su análisis manual. Es por esto que varios investigadores han recurrido al uso de las técnicas de aprendizaje automático.

Una experiencia que se desarrolló en la Universidad Técnica de Machala consiste en el desarrollo de un nuevo método para el análisis de mastitis subclínica en el ganado bovino. Este método se basa en el uso de un espectrómetro de reflectancia en el infrarrojo cercano (Near Infrared Reflectance - NIR), aplicado sobre muestras de leche cruda que fueron previamente etiquetadas con la metodología estándar de California Mastitis Test.

Se recogieron un total de 210 muestras de leche en receptores estériles etiquetados individuales. Se obtuvieron muestras de 67 vacas lecheras de raza mixta con 4,3 ± 1,8 años de edad, seleccionadas al azar de cinco granjas de la zona.

En el Gráfico 8.7 se observa las características de los espectrogramas NIR y sus ligeras diferencias que deberán ser analizadas utilizando técnicas de ML. El conjunto de datos estará disponible al público para su análisis una vez que el manuscrito sea publicado.

En el trabajo presentado, los modelos fueron desarrollados utilizando una técnica k-NN cuyo objetivo era detectar

el grado de mastitis. Los resultados de este trabajo muestran una gran potencialidad de la combinación de sensores de bajo costo con técnicas de ML. Los resultados no se detallan debido a que a la fecha de cierre de este libro, el manuscrito está en revisión por una importante revista científica del área.

Gráfico 8.7 Espectrogramas NIR de muestras de leche cruda etiquetadas de acuerdo al grado de mastitis bovina que presentan.

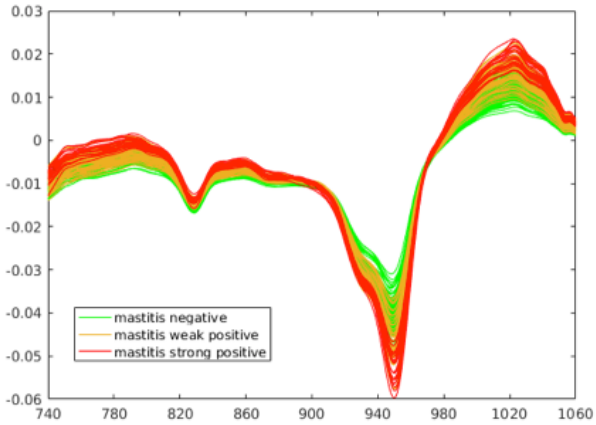
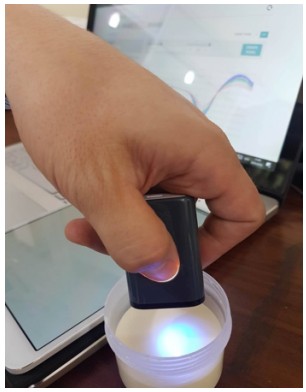


Imagen 8.1 Espectrómetro NIR portátil



El desarrollo de nuevos dispositivos portátiles abre un abanico de posibilidades para su aplicación en el campo agropecuario. En la Imagen 8.1 se puede observar la extracción de señales espectrales en una muestra de leche cruda uti-

lizando el dispositivo SCiO^{z} desarrollado por la compañía Israelita Consumer Physics. Estos nuevos dispositivos son esencialmente una nueva fuente de señales que requieren ser analizadas utilizando técnicas precisas para la obtención de información relevante.

Predicción en series temporales

En relación con el análisis de series temporales, su utilización en aplicaciones del sector agropecuario tiene que ver con la predicción de valores futuros y la alerta temprana de problemas.

Las series temporales, se analizan a partir de la reestructura de los patrones de entrada previo al entrenamiento de algoritmos de aprendizaje supervisado. Esto se hace, mediante la utilización de los datos previos como variables de entrada y utilizar un dato del siguiente día como la variable de salida (Kapoor & Bedi, 2013).

Este método se conoce como método de ventana deslizante, consiste en la creación de diferentes secuencias de puntos de datos consecutivos de la serie temporal. Existen dos parámetros en este método: tamaño de ventana y tamaño de paso en la ventana. El parámetro más importante es el tamaño de ventana, normalmente se experimenta con distintos valores hasta encontrar el valor óptimo, mientras que el tamaño de paso en la ventana se mantiene típicamente igual a 1.

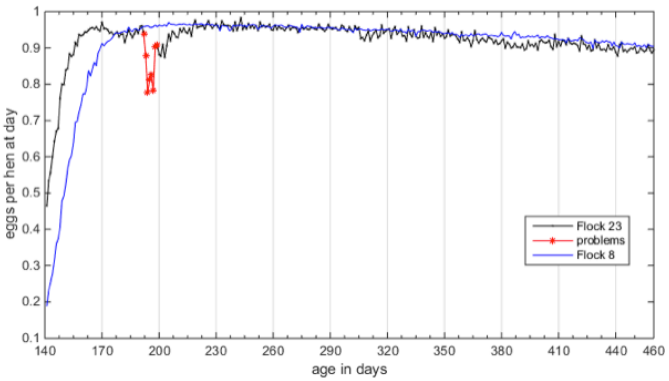
Dos trabajos publicados por Ramírez et al. (2016) y Ramírez et al. (2017) demuestra el uso de máquinas de soporte vectorial y de redes neuronales con una ventana deslizante para obtener un modelo de alerta temprana.

En este estudio, fueron registrados datos de campo de una granja de gallinas ponedoras alojadas en un sistema productivo de reemplazo denominado “todo dentro - todo fuera”, es decir que todas las aves de un mismo lote tienen la misma edad y son alojados juntos por grupos durante todo el tiempo de producción.

Los datos fueron recopilados diariamente desde enero 2008 a diciembre 2015. Debido a la organización y logística interna de la granja, los huevos fueron recogidos a distintas horas, por lo que el espaciado temporal de los datos no es de 24 horas. En algunos días el intervalo entre registros es de 20 horas y en otros de 28 horas. Los datos con variaciones en el espaciado temporal, representan un desafío para cualquier modelo (Jones, 1984), ya que debe ser capaz de de discriminar entre una anomalía en la curva producto de un problema real y las alteraciones relacionadas con el momento de la recolección.

En el gráfico 8.8, se muestra dos lotes representativos de la base de datos, el lote 8, que tiene una curva de producción característica, sin que se presenten problemas durante todo el tiempo de producción, y el lote 23, que a pesar de que inicia su producción con menos edad, muestra una fuerte caída entre los 191 días y los 199 días, este intervalo de tiempo fue etiquetado como anomalías en la curva, ya que a partir del día 199 las aves empiezan a recuperarse.

Gráfico 8.8 Producción por ave / día de dos lotes representativos de la base de datos



Los resultados de estos trabajos indican que es posible realizar un pronóstico automático de caídas de producción con una precisión, sensibilidad y especificidad superiores a 0.95. A nivel de finca, una pronostico con un día de antelación, podría resultar útil para la inspección diagnóstica en finca

en busca de síntomas clínicos, u otros hallazgos para la toma de medidas tendientes a la solución inmediata del problema. Esto mejora la capacidad preventiva en el sistema de producción avícola, brindando monitorización asistida de manera automática como complemento a la observación humana, lo que resulta especialmente útil, al manejar altas poblaciones de animales.

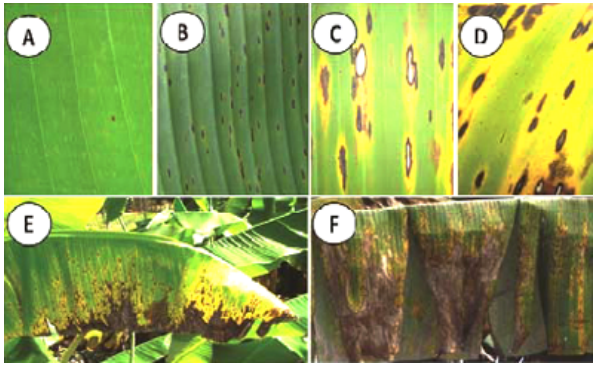
Análisis de imágenes

Entre la comunidad del sector agropecuario es sabido que las enfermedades de las plantas y de los animales amenazan a la seguridad alimentaria, en esta área en particular, el uso de técnicas de inteligencia artificial tiene un papel fundamental para la identificación precisa y oportuna de enfermedades en los cultivos. Sin embargo esta tarea no es para nada trivial, y requiere de una gran cantidad de recursos para el entrenamiento y desarrollo de los algoritmos.

Actualmente se utilizan imágenes multiespectrales e hiperespectrales para el cálculo de índices de salud de la vegetación, sin embargo su utilización está muy limitada debido al alto costo de los equipos. Por otra parte, en los últimos 10 años se ha dado un fenómeno de universalización de la posesión de smartphones, al punto de que prácticamente en todas las unidades de producción agropecuaria hay al menos un dispositivo.

Esta particularidad ha hecho que el diagnóstico de enfermedades mediante smartphone sea una realidad cada vez más cercana. Existen bases de datos tanto públicas como privadas que han recopilado y etiquetado decenas de miles de imágenes de plantas enfermas y sanas. En algunos casos estas imágenes han sido recolectadas en condiciones controladas, por lo que se infiere que su veracidad es alta. En el caso de bases de datos de animales sanos y enfermos, a criterio de los autores, no existen muchas fuentes de información, por lo que se recomienda iniciar una investigación en este sentido.

Imagen 8.2 Estadíos de afectación por Sigatoka Negra en hojas de banano (Vézina 2017).



En el Imagen 8.2 se puede observar los estadíos de afectación por el hongo que produce la enfermedad en el banano denominada Sigatoka Negra. En la Universidad Técnica de Machala, se ha propuesto para este año un proyecto que sea capaz de brindar una asistencia al diagnóstico en la evaluación del estado de afectación de las plantaciones de banano.

En la literatura científica se describen decenas de artículo científicos basados en ensayos experimentales que prueban la precisión de algoritmos de clasificación de imágenes, con resultados de más del 99% de exactitud, lo que pone en evidencia la viabilidad de este enfoque que más adelante será capaz de generar recomendaciones inteligentes asistidas por un smartphone a escala global.

Análisis de sonidos

Uno de los signos para el diagnóstico de enfermedades en los animales de granja, está relacionado con el sonido que emiten los animales. Particularmente en las enfermedades respiratorias. Los médicos veterinarios consideran a la tos, como un mecanismo de defensa del cuerpo, contra la posible entrada de agentes extraños en el sistema respiratorio.

Las características de la tos son indicativos de posibles enfermedades respiratorias. Partiendo de esta premisa, varios investigadores han estudiado los sonidos durante un

cuadro de tos en los animales para monitorizar posibles problemas de salud con la ayuda de un sistema experto.

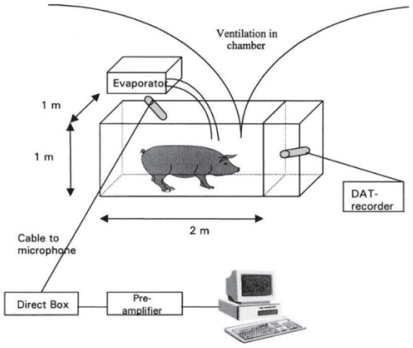
El uso de sistemas de soporte a la toma de decisiones en los sistemas agropecuarios tienen un alto potencial, debido a que en los sistemas de producción intensivos se manejan grandes cantidades de animales por lote, de tal manera que resulta un costo elevado tener sistemas de monitorización basados en observación humana.

El uso correcto de este tipo de sistemas es capaz de prevenir una zoonosis, o una epizootia, por este motivo, el desarrollo y aplicación de sensores y técnicas de detección para el diagnóstico automático es hoy un "hot topic" en la investigación y en la industria pecuaria.

Para el desarrollo de un sistema automático, se requiere que un experto etiquete una base de datos de sonidos de tos presencia de una enfermedad potencial, es decir, se utiliza técnicas de aprendizaje supervisado (Gráfico 8.9).

En el estudio de Chedad y colaboradores (2001) se utilizó redes de neuronas artificiales para predecir enfermedades respiratorias en cerdos. Para esto los autores construyeron una cámara de metal en la que cada cerdo es expuesto a variaciones de las condiciones ambientales tales como temperatura, el polvo, concentración de NH3, y otras variables. El sistema logró un reconocimiento correcto de los sonidos con más del 90% de precisión.

Gráfico 8.9 Esquema del ensayo para la grabación de los sonidos emitidos por los cerdos en el estudio de Chedad y colaboradores (2001)



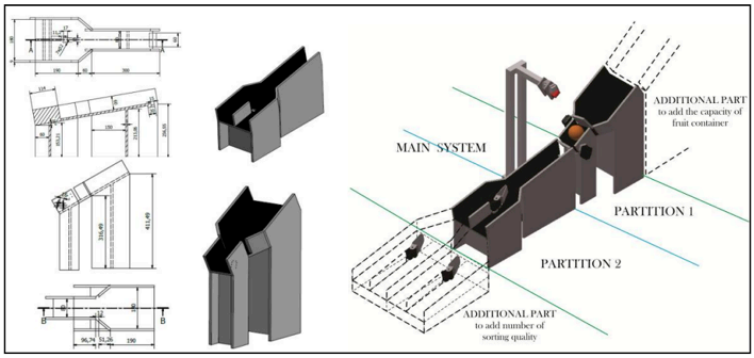
Análisis de videos

La supervisión por video se utiliza comúnmente en aplicaciones de detección y clasificación en la industria agropecuaria, principalmente en las cadenas agroindustriales y procesos de postcosecha.

Una aplicación que resulta interesante debido a su potencialidad para automatizar las medianas y pequeñas fábricas agroindustriales tiene que ver con la utilización conjunta de técnicas de visión por computadora, técnicas de deep learning y algunos servo motores. En el trabajo de Afrisal et al (2013) se utilizó una webcam para obtener vídeos en una planta de procesamiento de frutas.

El algoritmo de visión por computadora transforma el RGB (rojo, verde y azul) en el espacio de color HSV (tono, saturación y valor) para facilitar los procesos de segmentación de color. Luego un algoritmo de agrupamiento separa las frutas de acuerdo con el nivel de madurez y tamaño. Finalmente, los servo motores se activan para mover la fruta a una bandeja de acuerdo con su grado de calidad.

Gráfico 8.10 Diseño del clasificador portátil desarrollado por Afrisal et al (2013)



En el Gráfico 8.10, se puede apreciar de mejor manera el diseño del clasificador. El sistema es capaz de realizar la tarea de forma precisa en menos de un segundo, y el operador tiene la posibilidad de ver en tiempo real los resultados y los datos en su computador.

Referencia Bibliográfica

Afrisal, H., Faris, M., P., G. U., Grezelda, L., Soesanti, I., & F., M. A. (2013). Portable smart sorting and grading machine for fruits using computer vision. *In 2013 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)* (pp. 71–75). ieeexplore.ieee.org.

Chedad, A., Moshou, D., Aerts, J. M., Van Hirtum, A., Ramon, H., & Berckmans, D. (2001). AP—Animal Production Technology: Recognition System for Pig Cough based on Probabilistic Neural Networks. *Journal of Agricultural Engineering Research*, 79(4), 449–457.

Jones, R. H. (1984). Fitting Multivariate Models to Unequally Spaced Data. In E. Parzen (Ed.), *Time Series Analysis of Irregularly Observed Data* (pp. 158–188). Springer New York.

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

Liu M, Shi J, Li Z, Li C, Zhu J, Liu S. Towards Better Analysis of Deep Convolutional Neural Networks. IEEE Trans Vis Comput Graph 2017;23:91–100.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5(*4), 115–133.

Ramírez, I., Rivero Cebrián, D., Fernández Blanco, E., & Pazos Sierra, A. (2016). Early warning in egg production curves from commercial hens: A SVM approach. *Computers and Electronics in Agriculture, 121*, 169–179.

Ramírez-Morales, I., Fernández-Blanco, E., Rivero, D., & Pazos, A. (2017). Automated early detection of drops in commercial egg production using neural networks. *British Poultry Science*. https://doi.org/10.1080/00071668.2017.1379051

Shannon, C. E. (1950). XXII. Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 41*(314), 256–275.

Shiyu, J. (Nov, 13, 2016). Kernel method in SVM. Retrieved from https://commons.wikimedia.org/w/index.php?curid=60458994

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

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

Vézina A. Sigatoka leaf spot | The knowledge platform on the banana. The knowledge platform on the banana 2017. http://www.promusa.org/Sigatoka+leaf+spot (consultado el 11 de mayo de 2018).

Análisis de Datos Agropecuarios
Edición digital 2017- 2018.
www.utmachala.edu.ec

Redes

Redes es la materialización del diálogo académico y propositivo entre investigadores de la UTMACH y de otras universidades iberoamericanas, que busca ofrecer respuestas glocalizadas a los requerimientos sociales y científicos. Los diversos textos de esta colección, tienen un espíritu crítico, constructivo y colaborativo. Ellos plasman alternativas novedosas para resignificar la pertinencia de nuestra investigación. Desde las ciencias experimentales hasta las artes y humanidades, Redes sintetiza policromías conceptuales que nos recuerdan, de forma empeñosa, la complejidad de los objetos construidos y la creatividad de sus autores para tratar temas de acalorada actualidad y de demanda creciente; por ello, cada interrogante y respuesta que se encierra en estas líneas, forman una trama que, sin lugar a dudas, inervará su sistema cognitivo, convirtiéndolo en un nodo de esta urdimbre de saberes.

UNIVERSIDADE DA CORUÑA

UNIVERSIDAD TÉCNICA DE MACHALA
Editorial UTMACH
Km. 5 1/2 Vía Machala Pasaje
www.investigacion.utmachala.edu.ec / www.utmachala.edu.ec

ISBN: 978-9942-24-120-7
9 789942 241207