Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Usos de la Ciencia de datos aplicados al sector Agrícola

Trabajo Fin de Grado

Grado en Ingeniería Informática

Autor: Alfredo Mompó Serrano

Tutor: María José Ramírez Quintana

Cotutor: Fernando Martínez Plumed

Curso: 2021-2022

Resumen

La ciencia de datos es una disciplina que intenta obtener conocimiento nuevo a partir de grandes cantidades de datos. Esta disciplina cuenta con innumerables aplicaciones en los sectores económicos secundario y terciario, siendo especialmente significativo su papel como herramienta en procesos de producción o ciencias de la salud. También puede ejercer un papel importante en el sector primario, donde puede participar de forma activa en el proceso de toma de decisiones. En este trabajo se estudia la viabilidad del uso de la ciencia de datos en el sector agrícola describiendo varias aplicaciones y realizando algunos experimentos para ejemplificar esta relación, haciendo hincapié en sus usos en tareas como: el análisis de viabilidad y productividad de un producto, control de salud de cultivos, control de plagas y control de calidad.

Palabras clave: ciencia de datos, agrícola, predicción de producción de cultivos, control de plagas, visión artificial, minería de datos

Abstract

Data science is a discipline based in obtaining new knowledge from large amounts of data. Data science has plenty of uses in secondary and tertiary economic sectors being specially important its role as a tool in manufacturing or health sciences. But data science can also play an important role in the primary sector participating actively in decision-making processes. In this paper the usability of data science in the agricultural sector is studied, describing some applications and making experiments to illustrate this relationship, emphasizing its uses in: analysis of viability and productivity of a product, crop health control, pest control and quality control.

Keywords : data science, agricultural, crop yield prediction, pest control, artificial vision, data mining

Tabla de contenidos

1. Introducción

1.1. Motivación

La ciencia de datos lleva siendo una disciplina científica desde finales del siglo pasado, pero gracias a la computación avanzada sus métodos han sufrido una revolución en los últimos veinte años. Destacan sus usos en distintos ámbitos, entre ellos el sanitario donde se puede utilizar para proponer diagnósticos e incluso detectar células cancerígenas. O en el ámbito empresarial dónde forma un papel crucial a la hora de tomar decisiones estratégicas por parte de los directivos y también ofrece previsiones de ventas, recomendación de productos a clientes y otras muchas aplicaciones.

Por su parte el sector agrícola ha formado parte de la humanidad desde que el ser humano adaptó una forma de vida sedentaria con respecto al nomadismo y se trata de la base de la industria alimenticia. Desde la revolución agraria y posteriormente la revolución verde se ha ido incorporando cada vez más tecnología al sector, prescindiendo de la mano de obra manual y animal a favor de la utilización de maquinaria, también se han utilizado distintos métodos científicos como, por ejemplo: la utilización de abonos químicos, la ingeniería genética y los pesticidas. Todos estos cambios han propiciado el aumento masivo de la productividad agrícola.

Parece lógico que se haya desarrollado una relación entre la ciencia de datos y la agricultura en los últimos años (1). La mayoría de los fines para los que se utiliza la ciencia de datos en la agricultura puede resumirse en los siguientes puntos:

- Reducir el impacto ambiental de las actividades agrícolas. El excesivo uso de los recursos hídricos y de pesticidas pueden tener efectos negativos en el medio ambiente. Algunas de las soluciones de ciencias de datos expuestas pueden precisar las necesidades de los cultivos de manera que estos recursos no sean utilizados en exceso (2).
- Aumentar la producción y el rendimiento de los cultivos. Aumentando la producción de los cultivos se obtienen mejores beneficios para los productores y estos pueden hacer frente a demandas mayores de alimentos (3).
- Reducir costes de producción. Muchas de estas aplicaciones se centran en utilizar eficientemente los recursos ya disponibles, reduciendo los costes de producción.
- Ejercer un mayor control sobre la producción. Muchas de las soluciones a los problemas presentados ofrecen información directamente como, por ejemplo, mediante monitorización o indirectamente sobre el estado de la plantación, ofreciendo mayor control a los productores y en algunos casos también aumentando la seguridad alimentaria (4).

1.2. Objetivos

El objetivo de este trabajo de fin de grado es exponer y detallar algunas aplicaciones distintas de la ciencia de datos en el sector agrícola. También se realizarán algunos experimentos de minería de datos que ejemplifiquen estas aplicaciones y se analizarán sus resultados. Por último, se realizará un estudio sobre las posibles dificultades a las que se enfrenta la aplicación de estas soluciones en el sector.

Las aplicaciones detalladas en este trabajo se encuentran en distintas fases de desarrollo, algunas son conocidas y aplicadas comercialmente desde hace varios años, otras en cambio se encuentran en estudio. Las aplicaciones comentadas no son las únicas existentes en la actualidad y probablemente surjan más en los próximos años.

Todas las aplicaciones tienen como fin solucionar problemas relacionados directamente con los puntos ya mencionados: reducir el impacto ambiental, aumentar el rendimiento de cultivos, reducir costes de producción o ejercer un mayor control sobre la producción.

La parte experimental consiste en la construcción de dos modelos: el primero para la estimación de producción de mandarinas y el segundo centrado en clasificación de cuatro especies distintas de plagas de insectos por imagen. Estos modelos sirven de ejemplo práctico para dos de las aplicaciones descritas en el trabajo. Estos modelos serán estudiados y detallados en el apartado de Experimentos y también se hace referencia a los procesos realizados para la generación de los modelos en los comentarios dentro del código. Las herramientas utilizadas serán detalladas en el Anexo Herramientas.

2. Estado del arte de la materia

La ciencia de datos es la disciplina por la que se extrae información nueva a partir de datos ya disponibles. Esta disciplina se basa normalmente en modelos de carácter estadístico o matemático formados computacionalmente a partir de los datos ya obtenidos para descubrir relaciones o información nueva que sería imposible o menos eficiente sin dichos modelos. En la actualidad es una disciplina que cuenta con gran interés académico gracias a su potencial y a los avances en computación de los últimos años.

Las aplicaciones de ciencia de datos en el sector agrícola forman parte del marco de la agricultura de precisión. La agricultura de precisión se podría definir como un modo de gestión de los recursos agrícolas centrado en los datos. De manera que estos se recopilan, se analizan y se extrae información de ellos. Normalmente estos datos son presentados en un Sistema de Información Geográfica que establece una relación visual entre los datos y su localización geográfica. La agricultura de precisión nació en la década de 1980 en EE. UU. y se ha desarrollado hasta ser un campo de interés propio.

Por el estado de la ciencia de datos como una herramienta dentro de la agricultura de precisión se podría establecer un paralelismo entre los problemas que intentan solucionar ambas. Aparte de la ciencia de datos, la agricultura de precisión establece relaciones con otras tecnologías:

- Geolocalización. La mayoría de los sistemas de agricultura de precisión se apoyan en los sistemas GPS y en algunos casos en los sistemas globales de navegación por satélite para tractores u otra maquinaria mejorando la precisión de los trabajos realizados por estos.

- Robótica. Los avances en robótica permiten un menor uso de mano de obra en funciones repetitivas ahorrando costes de producción, también se están empezando a poner a la venta drones pilotados para tareas de fumigación o riego. La mayoría de estos sistemas dependen de la geolocalización para funcionar.

- Sistemas de irrigación inteligentes. Estos sistemas son capaces de agilizar la tarea de riego gracias a la aplicación de la telemetría. También son importantes los sistemas móviles de riego preciso y algunos pueden contribuir a un menor uso de agua.

1. Sistema de irrigación inteligente

- Siembra de ratios variables. Conocida como VRS por sus siglas en inglés, esta técnica trata de mejorar la producción de cierta parcela aplicando cantidades distintas de semilla, de manera que las zonas con mayor productividad reciban más semillas.

- Integración móvil e internet de las cosas. Muchos de estos sistemas tienen la capacidad de enviar la información de las actividades que realizan directamente al productor, además algunas de ellas permiten su control directo o programación de actividades mediante dispositivos móviles (5).

3. Estudio sobre aplicaciones de la ciencia de datos en el sector agrícola

3.1. Visualización en tiempo real de datos de producción

La visualización y formateado de datos para extraer información e interpretarlos de manera rápida también forma parte de la ciencia de datos. Por esto podemos incluir en el estudio los métodos que recopilan información sobre las parcelas agrarias de cierto propietario o de una zona particular y que ofrecen dicha información en tiempo real. El primer paso del proceso se trata de instalar sensores en la parcela de la que se quiera extraer la información.

Los sensores se suelen instalar dividiendo la parcela a monitorizar por zonas, para poder ubicar los datos con exactitud. Además, la mayoría de las empresas de monitorización de cultivos suelen ofrecer una representación geográfica de estos datos mediante mapas usando GPS. Los sensores utilizados más comúnmente suelen medir: PH del suelo, humedad del suelo, conductividad en el suelo o en el agua, temperatura del suelo o del ambiente, nutrientes como nitratos, fosfatos o potasio. Además, existen otros sensores compuestos que permiten saber información compleja cómo el clima, el estado del grano almacenado o algunos parámetros fisiológicos de las plantas como nivel de humedad interno o flujo de savia.

Todos estos datos se recogen y se plasman de una manera fácilmente interpretable cómo tablas o gráficos e incluso algunas empresas permiten programar alertas para ciertos parámetros como por ejemplo avisar al usuario si el nivel de humedad en el suelo de la parcela es inferior a cierta cifra. Con esta información el agricultor puede tomar decisiones rápidas basadas en el estado actual de la parcela.

Estas decisiones suelen estar más enfocadas al mantenimiento y no a la estrategia de producción a largo plazo por lo que no suelen ser necesarios análisis de datos más complejos o la figura de un analista especializado para controlarlos. También cabe destacar que, gracias a la instalación de estos sensores, se puede recoger una gran cantidad de datos útiles para realizar otros tipos de análisis.

Algunas empresas que ofrecen ya este tipo de servicio en España son Climate Fieldview (6) y Agrodato (7).

3.2. Análisis de viabilidad y productividad

I. Análisis técnico de viabilidad de cultivos y análisis derivados

El análisis de viabilidad técnico es un método por el cual se intenta determinar si la producción de cierto tipo de cultivo se puede realizar satisfactoriamente en una parcela dadas las características de ambos.

Es importante realizar este tipo de análisis para planificar la producción correctamente y ajustar las características que gracias al análisis se sabe que podrían afectar negativamente a la producción, como por ejemplo rectificar el pH del suelo de cultivo. El análisis también ayuda a precisar la inversión económica necesaria a lo largo de la producción con lo que ayuda en gran medida a la realización de estudios para determinar la rentabilidad del proyecto.

Este tipo de análisis es muy complejo y costoso económicamente pues precisa de personal cualificado para su realización. Además, se requiere de la toma de gran cantidad de datos y muchos de ellos requieren de material específico para su recogida.

Los datos más comunes a tener en cuenta en un análisis técnico son:

- Características propias del cultivo
- Localización de la parcela a analizar
- Características climáticas de la zona
    - Temperaturas
    - Radiación solar
    - Precipitaciones
    - Humedad
    - Vientos
    - Otras características meteorológicas
- Características edáficas
    - Tipo de suelo
    - Textura
    - pH
    - Conductividad eléctrica
    - Cantidad de materia orgánica
    - Salinidad
    - Concentración en nutrientes
- Características del agua de riego
    - pH
    - conductividad
    - compuestos disueltos
    - salinidad
- Rotación de cultivos.
    - Cultivos complementarios.
- Necesidades y régimen de riego.
- Necesidades de nutrición del cultivo.

- Plagas que afectan comúnmente al cultivo.

Como hemos mencionado anteriormente se trata de gran cantidad de datos y muchos de estos se deben recoger manualmente, pues no hay sensores automatizados para ciertos parámetros o no son lo suficientemente precisos para este tipo de análisis.

También hay que destacar que la mayoría de estos análisis precisan de equipamiento y software específico, cómo por ejemplo calculadores de régimen de riego.

Después de obtener los datos se suele realizar una comparación con los rangos de tolerancia del producto que se intenta cultivar. La salida del análisis incluye un informe que contiene:

- La viabilidad general del cultivo en la parcela.
- Ajustes a características edáficas (ver glosario) y mantenimiento.
- Régimen de abonado.
- Régimen de riego.
- Plagas que suponen una amenaza y estrategia de control de plagas.
- Diseño de rotación de cultivos.
- Material y herramientas necesarias para el cultivo.
- También pueden contener información económica como inversión necesaria, costes de mantenimiento o análisis de rentabilidad del proyecto.

Desde el punto de vista de la ciencia de datos este tipo de análisis es interesante no solo a la hora de utilizar técnicas de minería de datos para agilizar el proceso del análisis sino por la oportunidad que presenta la gran cantidad de datos utilizados.

Uso de minería de datos en el análisis
Los procesos de minería de datos podrían ayudar a reducir los recursos necesarios para realizar un análisis técnico de viabilidad. Tanto en el análisis y comparación de parámetros con las necesidades del cultivo, como en la identificación de los posibles parámetros que influyen críticamente en la viabilidad.

Oportunidades de uso de datos del análisis
Para el aprovechamiento de estos datos se requiere la construcción de un almacén de datos en la que se incluirían los datos del análisis para una gran cantidad de explotaciones, así como su productividad por años y los trabajos y problemas que se hayan encontrado en la producción.

Con esta base de datos se propone usar minería de datos en dos procesos distintos:

Recomendación de cultivo basado en resultados anteriores.
Este sería un uso parecido al proceso del informe técnico de viabilidad, pero de manera inversa. Dadas unas características técnicas, el sistema recomendaría el cultivo que mejor producción obtendría dadas estas características. Este proceso es especialmente interesante por su contribución a la sostenibilidad pues permitiría conocer si hay algún cultivo que se pueda producir sin realizar cambios en los parámetros o realizando los mínimos posibles. Por ejemplo, si se dispone de una tierra ácida con poca cantidad de materia orgánica el sistema nos recomendaría un cultivo que en estas condiciones

obtenga una buena producción sin necesidad de corregir estos valores, ahorrando en productos utilizados y minimizando el impacto medioambiental de la producción.

Descubrimiento de nueva información o relaciones desconocidas.

Este tipo de estudio sobre los datos se está usando para descubrir relaciones entre tratamientos o técnicas distintas y un aumento o disminución de la productividad de los cultivos. También está siendo utilizado para realizar estudios de adaptabilidad a condiciones desfavorables para un mismo cultivo. Con el objetivo de descubrir qué tipo de variedad sería más resistente a condiciones como un ataque bacteriano o un periodo de sequía.

II. Estimaciones de producción agrícola

La predicción de producción agrícola, nombrado a veces como CPY (Crop Yield Prediction) por sus siglas en inglés es un problema que trata de predecir numéricamente la cantidad de productos agrícolas obtenidos en el futuro para una zona geográfica predeterminada.

Esta información es de extrema utilidad y podemos clasificar dos ámbitos de uso distintos dependiendo de la extensión geográfica a analizar y de sus usuarios principales.

Extensión grande - Usuarios: Organizaciones y colectivos agrícolas y organizaciones de carácter estatal. En los usos de esta información para este ámbito destacan la prevención de hambrunas o de escasez de productos concretos y el control sobre los productos.

Extensión reducida – Usuarios: Productor y clientes. En este ámbito la información se puede usar para desarrollar la estrategia de producción, gestionar eficientemente recursos o informar a los clientes o intermediarios para realizar previsiones de venta.

La solución a este problema lleva planteándose desde el siglo pasado y las primeras aproximaciones dependían mayoritariamente de la propia experiencia de los propietarios obteniendo resultados poco fiables. Posteriormente, en la década de los noventa, con el aumento de la capacidad de computación se fueron desarrollando modelos matemáticos basándose en datos meteorológicos, y del propio desarrollo de la planta. Estos modelos son específicos para cada producto agrícola y son construidos tras un intenso trabajo de investigación sobre la influencia de varios factores sobre la cosecha, por lo que su coste es muy elevado.

Posteriormente empezaron a utilizarse modelos basados en datos que resultaron ser más baratos y fáciles de utilizar y modificar, por lo que adaptar un modelo a diversos cultivos requería una inversión menor que generar un modelo nuevo.

En estos modelos podemos diferenciar dos subgrupos: los modelos basados en estadística y modelos basados en aprendizaje automático, siendo estos últimos los que

han obtenido mejores resultados en los estudios comparativos para el problema de predicción agrícola (8,9).

Las técnicas más utilizadas en estos modelos son árboles de regresión, redes neuronales artificiales y regresión con vectores de soporte (10). Algo a considerar es que cada cultivo obtiene resultados más precisos con un método distinto. Así que siempre se recomienda realizar pruebas comparativas para la obtención de estos resultados (11).

Estos modelos utilizan datos tales como:

- Tamaño de la parcela.
- Datos sobre trabajos realizados en el campo.
- Datos sobre productos fitosanitarios aplicados.
- Datos sobre plagas o enfermedades.
- Datos meteorológicos
    - Precipitaciones.
    - Heladas.
    - Radiación solar.
- Otros datos.

También se han probado otras técnicas como el reconocimiento de patrones desde satélite o a nivel de campo, analizando los productos automáticamente para predecir la cantidad de producto listo para ser cosechado en cierto instante.

3.3. Control de calidad

I. Control de calidad de productos

La mayor parte del control de calidad utilizando ciencia de datos se basa en visión artificial. Estos sistemas se integran en sistemas mecanizados que capturan en tiempo real el producto valorando características como color, forma y tamaño a la hora de tomar las decisiones.

Este problema originalmente solía tratarse de una clasificación binaria en la que se clasifican los productos en válidos o no válidos para ser comercializados, aunque la mayor exigencia de los clientes ha forzado a las empresas a buscar nuevas distinciones en calidad de un mismo producto pudiendo obtener distintas calidades para satisfacer las preferencias y demandas de distintos clientes (12). O incluso se puede dividir los productos no válidos en categorías que se destinan a otro uso comercial como la alimentación de ganado.

La visión artificial supone un avance considerable en el proceso de control de calidad ya que tradicionalmente se ha hecho de manera manual con gran coste económico y temporal. En primera instancia, la clasificación automática solo podía valorar pesos o tamaños, pero con la implementación de visión artificial y tecnologías como rayos

ultravioletas en los sensores ópticos se pueden clasificar frutas incluso por su dulzura (13).

El proceso de visión artificial comprende una serie de pasos:

1. Adquisición de imágenes: en este paso se captura una imagen de la pieza a analizar, en el caso del control de calidad de productos agrícolas lo común sería tomar imágenes de frutas o verduras.

2. Preprocesado: En este paso se utilizan algoritmos que separan el objeto a analizar del fondo de la imagen y se mejora la calidad general de ésta. También se aplican otras técnicas a los datos que mejoran la calidad del modelo.

3. Segmentación: En este paso se divide la imagen en segmentos o píxeles con características similares como zonas con color o textura similares, dependiendo del análisis a realizar.

4. Evaluación: En el último paso se comparan los datos de los segmentos con los parámetros establecidos en la clasificación y se obtiene un resultado.

Este proceso se hace mediante redes neuronales convolucionales (ver glosario) donde se analizan las imágenes y la propia red genera las capas o filtros que se aplican a la imagen para obtener las características necesarias para su posterior clasificación(14).

En Andalucía una empresa es pionera en la aplicación de esta tecnología al cultivo de la oliva(15).

3.4. Pronósticos meteorológicos

Saber qué tiempo va a hacer es una cuestión de preocupación para la humanidad desde su origen, donde ya se estimaba el tiempo que haría mirando los colores del amanecer o la forma de las nubes. Aunque se trate de una ciencia anterior a la agricultura de precisión se incluye la meteorología en este proyecto por su implicación fundamental en el cultivo de cualquier producto. Además, el papel de la ciencia de datos en la previsión meteorológica ha ido tomando un protagonismo cada vez más importante en los últimos años complementando a los modelos numéricos.

La meteorología ha cambiado mucho en los últimos años, gracias a la enorme capacidad de computación de los sistemas informáticos actuales, que permiten realizar los millones de cálculos necesarios para las simulaciones. Los avances en ciencia de datos aplicados a esta materia han conseguido predecir algunos desastres naturales o la cantidad de contaminación y cómo afecta ésta al clima. Pero dado que el clima es un sistema caótico (ver glosario) su predicción es muy compleja.

Los pronósticos meteorológicos tienen especial importancia en el sector agrícola, más allá de la predicción de lluvias para la planificación de riegos, también involucra varios

otros procesos productivos como la siembra, la recolección, la aplicación de productos fitosanitarios, etc. Y su estudio en plazos más extensos de tiempo puede aportar información sobre producción, plagas y viabilidad económica de cultivos como se describe en este trabajo.

El proceso actual de pronóstico meteorológico utiliza el llamado modelo numérico de predicción meteorológica, que se trata de un conjunto de modelos de ecuaciones que toman gran cantidad de datos obtenidos por la red global de sensores meteorológicos. Estos sensores registran los datos que más tarde se procesan en los modelos generando datos de salida. Con estos datos de salida resultantes se construyen simulaciones y más tarde estas se adaptan según las características requeridas en la predicción.

Estos modelos son generados y mantenidos de manera manual por los meteorólogos, es decir, el aprendizaje de los algoritmos usados en los modelos no es automático. Los modelos pertenecen a organizaciones privadas o públicas y son distintos entre sí, además se suelen utilizar modelos distintos adaptados para zonas específicas.

El uso de inteligencia artificial y redes neuronales en los modelos es tema de estudio los últimos años y por ahora no se ha logrado el uso de aprendizaje automático para realizar un algoritmo completo de pronóstico meteorológico tan preciso como los modelos numéricos de predicción meteorológica (16). Como ejemplo, los modelos actuales basados en inteligencia artificial sólo consiguen predecir hasta 24 horas mientras que los modelos numéricos consiguen predicciones para 15 días.

3.5. Salud de cultivos

I. Detección de enfermedades y aplicación de fitosanitarios

Otro uso de especial interés de la visión artificial y la ciencia de datos es el posible diagnóstico de enfermedades en los cultivos mediante imágenes de hojas o tallos (17,18).

Esta tecnología se ha popularizado en los últimos años y como resultado han aparecido diversas aplicaciones tanto de uso comercial como personal basadas en ésta.

La mayoría de las aplicaciones se despliegan en dispositivos móviles con cámara para realizar la captura de la planta a analizar y ofrecen en pocos segundos la posible causa de la enfermedad, así como consejos para paliar sus efectos o para evitarla la próxima vez.

Estas aplicaciones suponen una ventaja pues pueden sugerir un diagnóstico rápido casi de inmediato sin necesidad de traer a un experto a la explotación ni de tomar muestras y desplazarlas, prescindiendo de la inversión económica y temporal que esto supone. De esta manera también se agiliza el tiempo de respuesta de los agricultores para tomar medidas.

Otro ámbito dentro de la agricultura de precisión en el que se puede usar esta tecnología es a la hora de aplicar los productos fitosanitarios en los cultivos como por ejemplo herbicidas, fungicidas o insecticidas.

Muchas veces se aplica la dosis errónea de producto ya sea por exceso o por falta. Por exceso se utiliza mucho más producto necesario pudiendo llegar a causar daños a los cultivos o al medio ambiente, además de suponer un gasto excesivo. Y por falta se utiliza menos producto del necesario pudiendo no resolver el problema e incluso si se tratase de una plaga poder hacer a esta resistente al tratamiento.

Por eso es necesario aplicar la dosis correcta en las zonas afectadas. Sobre este problema se ha investigado la ciencia de datos para mejorar la aplicación de dos técnicas distintas.

La primera técnica es la aplicación a nivel de campo del producto, normalmente mediante esparcimiento de producto granulado o pulverizado. En esta técnica la ciencia de datos se utiliza para reconocer las plantas afectadas que necesitan de tratamiento (19), pudiéndose incluir esta tecnología en sistemas de pulverizado automático montados en tractores o incluso en robots diseñados para realizar esta función (20).

También se ha utilizado la ciencia de datos en pulverizaciones aéreas con drones. En este caso los drones son capaces de reconocer la zona a pulverizar y diferenciarla de las zonas en las que no tienen que aplicar el producto. Este es el primer paso hacia el desarrollo de drones de pulverización completamente autónomos (21).

3.6. Control de plagas

I. Reconocimiento de insectos

Algunos insectos pueden causar graves daños a las plantaciones, se trata de una de las causas más importantes en la pérdida de calidad de los productos agrícolas. En este problema es de vital importancia el diagnóstico precoz para evitar la propagación y reproducción del insecto en cuestión.

Tradicionalmente, en pequeñas explotaciones agrícolas la identificación se ha basado en el propio conocimiento del agricultor o cuando este no es suficiente se suele complementar con la consulta de un experto.

Como en el caso de detección de enfermedades y de control de calidad, para resolver este problema se puede utilizar visión artificial. Su aplicación es muy parecida a la ya mencionada en el apartado de detección de enfermedades: es necesario tomar una imagen de la plaga a procesar o a veces incluso de los daños causados a la planta, acto seguido se procesa la imagen mediante aprendizaje automático y obtendremos un resultado de clasificación sobre la especie de la plaga (22).

La toma de imágenes se puede realizar de manera manual realizando una fotografía con un dispositivo móvil e incluso algunas empresas automatizan este paso ofreciendo

trampas inteligentes que identifican a los insectos que caen en ellas una vez instaladas en las parcelas. Estas empresas consiguen monitorizar casi en tiempo real qué especies de insectos hay en el campo y si su cantidad supone una amenaza al cultivo o no.

Una vez realizada la toma de imágenes, estas se procesan mediante ciencia de datos, de una manera parecida a la utilizada en el reconocimiento de enfermedades o control de calidad. En este proceso es muy importante la extracción de características de la forma del insecto tales como área, perímetro o longitud de ambos ejes.

II. Uso de ciencia de datos y minería de datos en el control de enjambres de langostas

Los enjambres de langostas, insectos de la familia *Acrididae,* son una de las mayores amenazas para la agricultura en zonas de todo el mundo, mayoritariamente en el este de África y Asia meridional, siendo muy preocupante en países como Pakistán, Yemen, Etiopía o India. En este último, el año pasado hubo diez enjambres activos simultáneamente y los insectos destruyeron casi cincuenta mil hectáreas de terreno de cultivo, en el que se ha estimado como uno de los peores años en pérdidas debido a los enjambres (23).

Estos enjambres se originan debido a un cambio en los ciclos de reproducción de los insectos en situaciones climáticas específicas, normalmente en la estación húmeda tras un período prolongado o especialmente duro de sequía.

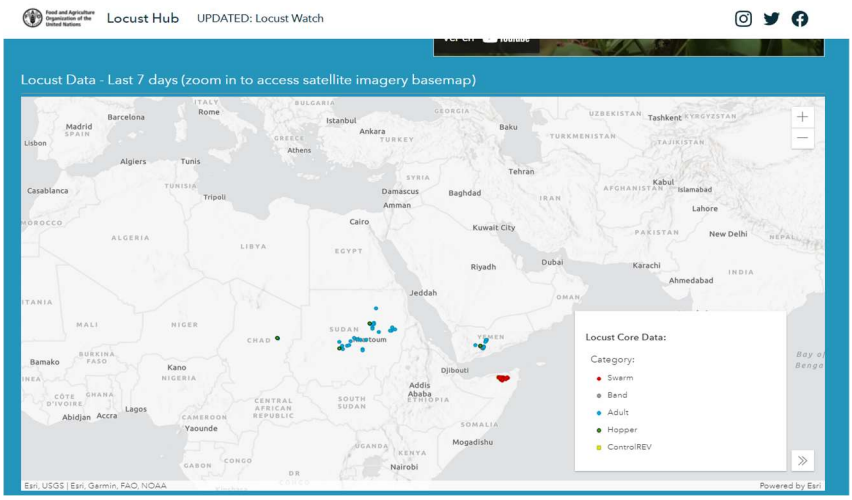
Estos sucesos meteorológicos y otros sucesos relacionados con variables biológicas características de las diferentes especies o incluso la densidad de población, hace que las langostas normalmente sedentarias sufran cambios en su comportamiento. Estas langostas se vuelven nómadas, reuniéndose en grupos cada vez más grandes y acelerando su consumo de alimento y ciclo de reproducción.

Frente a esta amenaza, en los últimos años se han desarrollado numerosas técnicas para reducir su impacto en los cultivos. Una vez iniciado un enjambre es muy difícil de detener o controlar, por eso estas técnicas se basan mayoritariamente en sistemas preventivos, como la fumigación precoz de grupos de langostas que en un futuro pueden crear un enjambre.

En este contexto es muy importante la recopilación y análisis de grandes cantidades de datos para obtener un tiempo de reacción de las autoridades competentes lo más rápido posible. Entre estas medidas destacan las siguientes:

La red de monitorización Locust-hub desarrollada por la Organización de las Naciones Unidas para la Alimentación y la Agricultura.

Esta página nos permite ver en tiempo real las observaciones de grupos de langostas en forma de mapa junto con los datos de cada observación como: fecha de inicio, hectáreas que ocupa o país donde surgió el enjambre.



2.La página Locust Hub y su mapa en tiempo real

Desde esta página también existe la capacidad de descargar conjuntos de datos sobre las plagas, sobre la ecología de los insectos o sobre las operaciones de este grupo relacionadas con la crisis de las langostas, pudiendo hacer el usuario sus propios análisis si lo desea.

Otras funciones de la página incluyen una tabla configurable para ver los datos según los parámetros escogidos y un mapa en tiempo real del nivel de humedad del suelo, dato importante en el estudio de la formación de enjambres.

El modelo de predicción de movimiento de plagas desarrollado por WISER y ICPAC

WISER (Weather and Climate Information Services for Africa) es una subdivisión de la agencia estatal de meteorología de Reino Unido que presta ayuda y servicios en esta materia a diferentes organizaciones africanas.

El ICPAC (IGAD climate prediction and applications centre) es una rama de la Autoridad intergubernamental para el desarrollo. Esta autoridad en la que participan seis miembros de la zona oriental de África tiene como objetivo conseguir la paz, la prosperidad y la integración regional de sus países miembros.

El modelo, alojado en un superordenador en el centro de la ICPAC en Kenya, toma datos de las predicciones meteorológicas como velocidad del viento y dirección, temperatura y humedad del aire para predecir con exactitud el movimiento de los enjambres.

También se está estudiando integrar en el modelo parámetros como la humedad del suelo y densidad de vegetación con el objetivo de establecer zonas de puesta de huevos, complementando la vigilancia terrestre. Fumigando estas zonas de puesta se consigue controlar las poblaciones de ejemplares jóvenes antes de que formen un enjambre (24).

4. Dificultades en la aplicación de la ciencia de datos en el sector agrícola

Pese a sus aplicaciones la ciencia de datos sigue siendo poco utilizada en el sector agrícola. En este apartado se describen algunos de los posibles motivos.

Alta inversión necesaria

La necesidad de contratar personal cualificado y equipos de alta tecnología suponen que para implementar ciencia de datos se necesita una alta inversión económica.

También hay que añadir que muchos de los datos necesarios para estos análisis se basan en una temporalidad extensa como años o meses por lo que el horizonte temporal de recuperación de la inversión es de medio a largo plazo.

Estos dos factores suponen que la inversión necesaria para desplegar tecnologías basadas en ciencia de datos solo sea asumible por explotaciones agrícolas de gran tamaño y con gran capacidad financiera. En países o zonas en las que la mayoría de parcelas pertenece a pequeños propietarios con pocas extensiones de terreno y poco margen para invertir en nuevas tecnologías nos encontramos frente a una dificultad de aplicación.

Tecnología en desarrollo

En todo este trabajo se ha visto que la aplicación de la ciencia de datos al sector agrícola está aún en proceso de desarrollo, comparada a la aplicación en otros sectores como las finanzas o las ventas. Esto presenta una serie de problemas frente a otro tipo de sectores en los que ya está establecida, entre estos destacan el mayor coste de implementación y mantenimiento y la desconfianza por parte del consumidor al no tener un mercado establecido.

Difícil acceso a los datos

Muchos de las aplicaciones anteriormente descritas necesitan de datos recogidos expresamente para este fin. Es decir, la mayoría de los datos no estaban registrados antes de plantearse la aplicación de estas tecnologías por lo que estos datos se convierten de difícil acceso para fines experimentales. También es interesante mencionar la falta de registros sobre operaciones agrícolas o de análisis de viabilidad en la mayoría de parcelas.

5. Experimentos

A continuación, se describen los experimentos relacionados con dos de las aplicaciones mencionadas anteriormente. Todos los datos junto con el código de los dos experimentos se pueden encontrar en el siguiente repositorio: https://github.com/Alfred-MS/experimentosTFG

Experimento 1: Estimación de producción de diversas plantaciones de mandarinas

Para ejemplificar el problema de la predicción de producción se va a realizar un análisis usando técnicas de minería de datos para tratar de entrenar y construir un modelo que pueda aproximar la producción agraria de cierta parcela dados los datos de esta en el año deseado.

La primera tarea a realizar en este problema es conocer el producto a analizar. Todas las parcelas se encuentran en la comarca de La Ribera Alta, provincia de Valencia y producen mandarinas de la variedad “satsuma okitsu”. Se trata de una mandarina de origen japonés de tamaño pequeño, su color es naranja amarillento, su sabor es más ácido que dulce y su piel suele ser fina, rugosa y fácil de pelar.

3.Satsuma Okitsu

En cuanto al árbol, es de tamaño mediano muy resistente a las heladas y moderadamente resistente a las sequías. La “satsuma okitsu” es de los últimos cítricos en florecer, pero de los primeros en ser recogidos. La recolección se realiza desde finales de septiembre a octubre, cuando el cítrico aún presenta un color verdoso. También es importante tener en cuenta el proceso de cuidado de los árboles que necesitan de podas regulares y de aclareo (ver glosario).

Los datos de las explotaciones son datos reales que han sido anonimizados y cedidos por una cooperativa local, que se entregaron en formato .xlsx. En lo referente a la semántica de los datos obtenemos tres tipos:

- Identificador de parcela de tipo int64: Identificador para cada parcela su cifra varía del uno al sesenta y tres.
- Hanegadas de tipo float64: Tamaño de la parcela en hanegadas, más tarde será reconvertido a metros cuadrados
- Kilogramos producidos por año de tipo float64: Producción de cada parcela de determinado año en kilogramos.

Este es un extracto que corresponde a las primeras cuatro filas de los datos:

PARCELA	HGS	KG 2017	KG 2018	KG 2019	KG 2020	KG 2021
1	4,448	11.130	13.332	14.807	12.628	14.679
2	1,768	3.216	1.848	3.248	0	0
3	1,841	0	0	0	0	0
4	7,518	24.366	21.749	1.727	18.803	20.615

4.Extracto de las primeras 4 filas de los datos proporcionados

Los datos completos se estructuran en 63 filas de las que se puede observar que cada parcela tiene detallado su tamaño en hanegadas y la producción que se obtuvo de cada parcela para los años 2017, 2018, 2019, 2020 y 2021.

Integración y Limpieza de los Datos

Tras un análisis inicial, podemos observar que los datos están incompletos, ya que faltan valores de distintas columnas, estos datos serán omitidos en la etapa de preprocesado y no serán utilizados en el análisis. Por su parte, la tarea de minería de datos se ha identificado como un problema de regresión. Aunque se podría intentar entrenar un modelo usando únicamente los datos de las áreas de las parcelas, dado el número limitado de instancias disponibles, se ha intentado buscar más datos para la creación del modelo.

Como se ha estudiado anteriormente, para solventar la anterior limitación se podría añadir al modelo una mayor cantidad de datos complementarios sobre las explotaciones como por ejemplo: datos sobre operaciones agrícolas realizadas sobre el cultivo o datos sobre la salud de los árboles. Sin embargo, estos no suelen ser registrados y por lo tanto no son fácilmente accesibles. Por estos motivos se ha decidido complementar los datos de área cultivada con información meteorológica.

Los datos climáticos más importantes como hemos visto anteriormente son: [1] la cantidad de precipitaciones, y [2] las temperaturas extremas que puedan afectar a los cultivos. En el caso de las temperaturas extremas, el estudio anterior ha demostrado que la variedad “satsuma okitsu” es resistente a estas. Además, una exploración de los datos climáticos muestra que las heladas en la zona se dan de diciembre a febrero cuando la fruta ya se ha recogido, por lo tanto, no puede ser dañada por las heladas, que además en los últimos años carecen de la fuerza necesaria para dañar cultivos similares.

Después de descartar las heladas como datos relevantes en el modelo se ha procedido a obtener las cifras correspondientes de las precipitaciones. Para ello se ha recurrido a la página oficial de la AEMET y a su proyecto de open data (25). Para este modelo se

han seleccionado los datos de la estación 8414A por su relativa cercanía a la zona de los cultivos y por su completitud en los datos.

Preprocesamiento de datos

Una primera parte del preprocesamiento de datos se trata de agregar los datos de precipitaciones, realizar la conversión de hanegadas a metros cuadrados y desagrupar cada fila para obtener una tripla que contenga información acerca de: área de la parcela en metros cuadrados, datos de pluviosidad de la temporada en mm, y producción en kg.

Cada fila del conjunto de datos cuenta además de un identificador para poder asociarla a cada uno de los registros originales y otro para cada parcela. El resto de preprocesamiento de datos se desarrolla en la libreta del análisis, se trata de eliminar los valores de producción nulos y de estandarizar las características para los métodos en los que ha sido necesario.

La descripción de las herramientas utilizadas en esta tarea se encuentra descrita en el Anexo. Hay que destacar que como herramientas de desarrollo se ha utilizado Anaconda (Python) y, Jupyter Notebooks, usando como librerías principales Numpy y sklearn.

	ID_REGISTRO	ID_PARCELA	Area	Lluvia_Temporada	KG
0	1	1	3696.29	324.0	11130.0
1	2	2	1469.21	324.0	3216.0
3	4	4	6247.46	324.0	24366.0
4	5	5	2763.91	324.0	3031.0
5	6	6	2715.71	324.0	9512.0
...	...	...	...	...	...

5.Vista minable de los datos

Análisis

Para la estrategia de partición se han dividido los datos en dos partes: la parte de entrenamiento y la parte de evaluación. La primera se utiliza para entrenar el modelo y la segunda para evaluarlo con datos con los que no haya trabajado antes, probando su capacidad de predicción a partir de información nueva.

Partición	Nombre de variable	Porcentaje de muestras deseado	Tamaño final de partición
Entrenamiento	x_train	80%	213
Evaluación	x_test	20%	54

6.Tabla explicativa de la estrategia de partición

Se ha realizado un análisis con tres modelos distintos de regresión: regresión por mínimos cuadrados ordinarios, regresión con vectores de soporte y regresión por bosques aleatorios.

Resultados del análisis y conclusiones

Para evaluar la calidad del análisis obtenemos la métrica de calificación R^2, también conocida como el coeficiente de determinación. En el caso de un modelo con capacidad predictiva R^2 tomará valores de 0 a 1 siendo este último el mayor valor posible, capaz de estimar con la máxima precisión el resultado. En el caso de un modelo sin capacidad predictiva, la métrica será negativa.

Las valoraciones para los modelos resultantes que se utilizan para comprobar la validez y la calidad de los modelos generados han sido bajas, como puede verse en la tabla a continuación:

Modelos	R^2 (Porcentaje)
Regresión por mínimos cuadrados ordinarios	47%
Regresión con vectores de Soporte	Negativo (Sin capacidad predictiva)
Regresión con bosques aleatorios	25%

7.Tabla de métricas de los modelos construidos

Como podemos observar el modelo que menor valoración ha obtenido ha sido el modelo de regresión con vectores de soporte, siendo esta incluso negativa, lo que indica que no podemos construir un modelo predictivo con esta técnica y estos datos. El modelo de regresión lineal por mínimos cuadrados ordinarios, en cambio sí es usable para predecir la producción, dados el tamaño de la parcela y las precipitaciones del año, aunque la precisión de la previsión no sea muy elevada. Finalmente, el modelo de regresión con bosques aleatorios, aunque obtiene una valoración en R^2 positiva, ésta es menor a la obtenida por la regresión por mínimos cuadrados ordinarios, con lo que se trata de un modelo de menor precisión.

Como conclusión, la construcción de un modelo de predicción de cultivos dados datos climáticos y de tamaño de explotación para la variedad de mandarinas “satsuma okitsu” es factible, sin embargo, estos datos no son suficientes para realizar un modelo preciso. Esto se debe a que es un cultivo de regadío y la cantidad de precipitaciones, aunque sí que es una característica influyente, no es especialmente importante. El tamaño en contrapunto sí es una característica altamente relevante para el resultado de la predicción.

Si se continuara el desarrollo del análisis se sugiere incluir otro tipo de datos que puedan mejorar la calidad de la predicción como por ejemplo los trabajos realizados sobre las explotaciones, en especial el ya comentado aclareo o si los árboles han sufrido algún tipo de problemas como plagas o enfermedades.

Experimento 2: Modelo de clasificación de plagas de insectos

Para el segundo experimento se construye un modelo con el fin de clasificar automáticamente insectos basándose en imágenes de estos. Este experimento se trata de un ejemplo de la aplicación de control de plagas: reconocimiento de insectos. Se trata de un problema de clasificación que resolveremos utilizando una red neuronal artificial.

Como en el experimento anterior, la descripción de las herramientas utilizadas en esta tarea se encuentra en el Anexo.

Origen de datos

Los datos se tratan de imágenes de cuatro insectos distintos originarios de Asia: Eudocima phalonia (Polilla perforadora de frutos), insectos de la familia Cecidomyiidae (Mosca de las agallas), Acrididae(Langosta) y larvas de las órdenes coleóptera y lepidóptera (Larva barrenadora).

8.Mosca de las agallas

9.Polilla perforadora de frutos

Las imágenes se han obtenido a partir de búsquedas en la web y su conjunto se encuentra publicado gratuitamente bajo licencia de dominio público (26).

Preprocesado

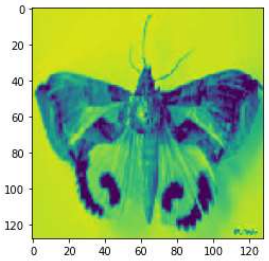
Algunas imágenes se descartan del conjunto original por su poca claridad o relevancia.También se importa para trabajar un conjunto de 295 imágenes de cada uno de los insectos ya que es el mínimo de imágenes de los cuatro grupos, perteneciendo este mínimo a las langostas. Una vez importada la cantidad de imágenes necesarias estas se redimensionan a 128x128 y se transforma su color en una escala de grises.

Además, las imágenes se transforman a formato matricial para poder ser procesadas y se les aplica un algoritmo de normalización.

En esta etapa también se construye el array en el que se codifican de cero a tres las etiquetas de clase, asignando un valor por cada una de las imágenes según la especie. Este array de clases se denomina “array_species” y según la cifra almacenada se puede obtener la equivalencia de la especie de la imagen.

Valor en array de resultados (Y)	Especie
0	Polilla perforadora de frutos
1	Mosca de las agallas
2	Langosta
3	Larva barrenadora

10.Tabla de equivalencia entre los valores de "array_species" y la especie a la que pertenece la imagen



11.Ejemplo de imagen tras el preprocesado

Construcción del modelo

Antes de construir el modelo se debe plantear una estrategia de partición de datos, así pues, se dividen las imágenes en entrenamiento, evaluación y validación. La parte de entrenamiento será la más numerosa y será usada para el entrenamiento del propio modelo. Tendremos dos partes de tamaño similar para la validación y la evaluación. La validación se usará para evitar el sobre entrenamiento del modelo y la parte de evaluación se usará para probar el modelo con datos que no hayan pasado previamente por este.

Partición	Nombre de variable	Porcentaje de imágenes deseado	Tamaño final de partición
Entrenamiento	X_fit	75%	885
Validación	X_val	12.5%	147
Evaluación	X_test	12.5%	148

12.Tabla explicativa de la estrategia de partición

Una vez definidas las particiones se procede a la construcción del modelo donde se utilizarán dos capas convolucionales: ambas con un tamaño de kernel de 3x3, la primera con un tamaño de entrada de 128x128 pues es el número total de píxeles de la imagen y un tamaño de canal de 128 también. Después de cada capa convolucional se han añadido capas de pooling (ver glosario) para facilitar la extracción de características. Para finalizar el modelo se han añadido dos capas densas para obtener la clasificación.

En este modelo la precisión se interpreta como la media de la discrepancia entre los resultados de clasificación predichos por el modelo para cada imagen y su clasificación real contenida en el array “array_species”. También definimos una parada precoz

utilizando la partición de validación. Si en las iteraciones de entrenamiento, aunque se mejore la precisión del modelo para el conjunto de entrenamiento no lo hace para la partición de validación, el entrenamiento se detiene, esto se realiza para evitar el sobreajuste (ver glosario).

13.Esquema de capas del modelo

Resultados del modelo y análisis

En las evaluaciones del modelo obtenemos los siguientes resultados en términos de la tasa de acierto del modelo (accuracy) (que es la medida de evaluación más utilizada en clasificación) para los siguientes conjuntos de imágenes:

Partición	Precisión
Entrenamiento	92.88%
Validación	57.82%
Evaluación	54.05%

14.Tasa de acierto del modelo para cada partición

Como se puede observar en el modelo obtiene predicciones bajas para imágenes nuevas, la mayoría de los errores en estas clasificaciones vienen dadas por la clase 0, es decir, el modelo confunde las polillas más que otros insectos como se puede ver en la imagen inferior. En comparación con otro tipo de experimentos similares (27) se deduce que la cantidad y la calidad de las imágenes son un factor que considerar para este tipo de análisis, sin embargo, como ya se ha comentado anteriormente recopilar este tipo de datos es muy costoso.

	precision
0	0.37
1	0.64
2	0.62
3	0.48

15. Precisión obtenida por clase

En conclusión, se ha realizado un modelo para verificar la viabilidad de la aplicación de control de plagas: reconocimiento de insectos con resultados satisfactorios dados los datos utilizados. Para un modelo más potente se deberían usar imágenes de manera uniforme, es decir tomadas a los insectos desde el mismo ángulo y distancia a la cámara. Esto ayudaría al modelo a extraer características tales como tamaño y morfología de manera más precisa.

6. Conclusiones

Como conclusiones de este trabajo, podemos afirmar que la ciencia de datos es aplicable al proceso agrícola en muchos ámbitos distintos y este podría verse ampliamente beneficiado por su uso. Esta relación junto a la variedad de usos posibles presenta una oportunidad en la aplicación de la ciencia de datos, que puede resultar interesante desarrollar en un futuro mejorando las aplicaciones ya presentes y desarrollando nuevas.

Es, sin embargo, necesario destacar que esta relación entre el mundo agrícola y la ciencia de datos no es una relación tan consolidada como en otros sectores como podrían ser el marketing o la salud. Esto ocasiona que muchas veces estas aplicaciones están actualmente en desarrollo o en fase experimental y no haya muchas empresas que ofrezcan soluciones basadas en ciencia de datos de manera comercial.

Como contrapunto positivo, en los últimos años se ha incrementado el interés académico en este tema y se están investigando nuevas aplicaciones en diversas universidades destacando países como España, Estados Unidos o India. Un problema al que se enfrentan estas investigaciones es la dificultad de obtener datos usables y fiables, pues la mayoría de los datos presentados por organizaciones o recogidos por terceros suelen ser de baja calidad. Esto supone que muchos de los estudios se basan en datos recogidos por los mismos investigadores o por las universidades. Y dado el tiempo necesario para extraer estos datos como producción anual durante varios años, sigue siendo un proceso costoso y lento.

También es importante mencionar que el elevado coste del despliegue de estas tecnologías en producciones agrícolas comprende una dificultad adicional para pequeños productores o productores con pocos recursos económicos.

Pese a todos estos factores la relación entre ciencia de datos y el sector agrícola puede ser beneficiosa para ambos campos y con el suficiente tiempo y desarrollo podría llegar a convertirse en una de las herramientas más importantes para el mantenimiento, prevención, planificación y toma de decisiones en las parcelas.

Glosario

Redes neuronales convolucionales: Red neuronal artificial basada en un perceptrón. Su capacidad para extraer características de matrices bidimensionales las convierte en el método más utilizado para la visión artificial.

Sistema caótico: Sistema altamente complejo y con multitud de variables que además puede mutar en el tiempo. Los sistemas caóticos pueden cambiar su resultado drásticamente si alguna de sus variables cambia mínimamente lo que los convierte en sistemas difíciles de predecir o estudiar.

Edáfico: Relativo al suelo, suele usarse para referirse a las cualidades de este cuando se habla de plantas.

Aclareo: Proceso agrícola en el que se descarta el fruto aun en el árbol que no alcanza algún estándar de calidad, normalmente su tamaño. Este proceso potencia el desarrollo de los demás frutos del árbol y también se puede aplicar cuando estos frutos están demasiado apelotonados o dañados.

Pooling: Proceso por el que se reduce el tamaño de una matriz representativa de datos como por ejemplo una matriz obtenida de una imagen sin reducir su información.

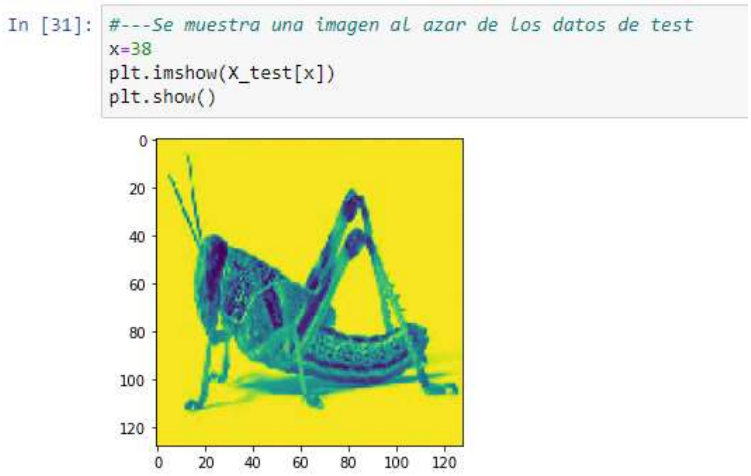
Sobreajuste: Situación dada cuando una red neuronal se entrena más de la cuenta, ocasionando que esta obtenga buenos resultados para los valores de entrenamiento, pero sea incapaz de obtener resultados para valores nuevos.

7. Herramientas

Anaconda: Es una distribución de Python y R centrada en la gestión de módulos, paquetes de módulos y entornos independientes centrada en computación (28). Ha sido elegido por su capacidad de gestión de módulos y entornos de manera separada, por lo que se puede disponer de distintos entornos con diferentes propósitos y librerías en un mismo equipo. Esta característica evita algunos errores y permite trabajar con las librerías necesarias para cada trabajo sin que estas se mezclen.

Python: Lenguaje de programación muy popular por su legibilidad, es un lenguaje interpretado, dinámico y multiplataforma creado en 1991. Muy popular en ciencia de datos por su habilidad para importar módulos ampliando su funcionalidad (29). Se ha elegido esta herramienta por su versatilidad y facilidad de lectura de código, todos los experimentos están realizados con este lenguaje como base.

Jupyter Notebook: Formato de archivo que contiene código ejecutable junto con sus resultados y otro tipo de datos como comentarios y títulos. Popular en aprendizaje por su facilidad de envío y lectura (30). Se ha elegido esta herramienta por su capacidad de integrar otros tipos de datos y añadidos al código, en particular la capacidad de mantener los resultados de una ejecución de parte de código, aunque se haya terminado su ejecución y se haya cerrado el programa. También es interesante su uso en el segundo experimento por su capacidad de integrar algunas imágenes en el propio código con lo que se puede seguir más fácilmente las transformaciones que sufren las imágenes.



16.Ejemplo de utilización de jupyter notebook en el experimento 2: la salida de la ejecución con comentarios es una imagen

Numpy: Módulo de Python especializado en computación y álgebra. Utilizado por su capacidad de procesar operaciones con vectores y matrices de manera rápida e intuitiva (31). Sobre esta librería se construyeron las dos librerías siguientes. Su uso es necesario en los dos experimentos por sus tipos de datos y sus operaciones sobre estos.

Sklearn: Módulo de Pyton para aprendizaje automático (32). Utilizado en el primer experimento, dispone de una gran variedad de modelos de minería de datos para todas sus aplicaciones como por ejemplo clasificación, regresión y agrupación. Esta librería también incorpora otras funcionalidades que complementan la minería. Algunas de estas funcionalidades son: métricas para selección de modelo y herramientas para el preprocesado. Estas también han sido utilizadas en el experimento de estimación de producción en la evaluación de modelos o el descarte de valores nulos.

```
# Partición de los datos
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

```
# Construimos el modelo
model=LinearRegression().fit(x_train,y_train)
```

17.Ejemplo de utilización del módulo sklearn en el experimento 1 para la partición de datos y construcción del modelo

Keras: Módulo de Pyton para Redes Neuronales (33). Utilizada en el segundo experimento, esta librería basada en la popular librería de redes neuronales Tensorflow, añadiendo funcionalidades. Se caracteriza por facilitar la construcción y entendimiento de los modelos desarrollados con esta herramienta gracias a su capacidad de construcción de redes neuronales con capas.

```
#---Early stopping con datos de validación para evitar el sobreajuste
callbacks = [EarlyStopping(monitor='val_accuracy',patience=3,restore_best_weights=True)]
```

18.Utilización de un método de tensorflow a través de keras

Bibliografía

1. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine Learning in Agriculture: A Review. Sensors 2018, Vol 18, Page 2674 [Internet]. 2018 Aug 14 [cited 2021 Nov 29];18(8):2674. Available from: https://www.mdpi.com/1424-8220/18/8/2674/htm

2. Cómo influye el cambio climático en la agricultura [Internet]. [cited 2021 Nov 29]. Available from: https://www.suez-agriculture.com/es/blog/como-influye-el-cambio-climatico-en-la-agricultura

3. Agricultura de precisión para una producción más sostenible frente al cambio climático y la creciente demanda de alimentos | Cátedra bp de Medio Ambiente Industrial [Internet]. [cited 2021 Nov 24]. Available from: https://www.catedrabpmedioambiente.es/agricultura-de-precision-para-una-produccion-mas-sostenible-frente-al-cambio-climatico-y-la-creciente-demanda-de-alimentos/

4. Agricultura de precisión: una posible respuesta al cambio climático y a la seguridad alimentaria. - Sostenibilidad [Internet]. [cited 2021 Nov 29]. Available from: https://blogs.iadb.org/sostenibilidad/es/agricultura-de-precision-una-posible-respuesta-al-cambio-climatico-y-a-la-seguridad-alimentaria-pero-es-asequible-para-todos-2/

5. IoT in Agriculture | Oxagile [Internet]. [cited 2021 Nov 29]. Available from: https://www.oxagile.com/competence/internet-of-things/agriculture/

6. Data-driven agricultural decisions and insights to maximize every acre [Internet]. [cited 2021 Nov 29]. Available from: https://climatefieldview.es/

7. AGRODATO, agricultura de precisión en España [Internet]. [cited 2021 Nov 29]. Available from: https://www.agrodato.com/

8. Kamath P, Patil P, S S, Sushma, S S. Crop yield forecasting using data mining. Global Transitions Proceedings. 2021 Nov 1;2(2):402–7.

9. Estimation of Major Agricultural Crop with Effective Yield Prediction using Data Mining. [cited 2021 Nov 30]; Available from: www.data.gov.in].For

10. Gonzalez-Sanchez A, Frausto-Solis J, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research. 2014 Apr 29;12(2).

11. Sahu S, Chawla M, Khare N. Viable Crop Prediction Scenario in BigData Using a Novel Approach. In 2019.

12. Kaur S, Girdhar A, Gill J. Computer Vision-Based Tomato Grading and Sorting. Lecture Notes in Networks and Systems [Internet]. 2018 [cited 2021 Oct 29];38:75–84. Available from: https://link.springer.com/chapter/10.1007/978-981-10-8360-0_7

13. Nazulan WNSW, Asnawi AL, Ramli HAM, Jusoh AZ, Ibrahim SN, Azmin NFM. Detection of Sweetness Level for Fruits (Watermelon) with Machine Learning. 2020 IEEE Conference on Big Data and Analytics, ICBDA 2020. 2020 Nov 17;79–83.

14. Magomadov VS. Deep learning and its role in smart agriculture. Journal of Physics: Conference Series [Internet]. 2019 Dec 5 [cited 2021 Nov 29];1399(4). Available from: https://www.researchgate.net/publication/337764156_Deep_learning_and_its_role_in_smart_agriculture

15. Monitorización en tiempo real del rendimiento y calidad en olivar superintensivo basado en Deep Learning | ANDALUCIATECH US [Internet]. [cited 2021 Nov 25]. Available from: https://andaluciatech.org/ecosistema-innovador/proyectos-singulares/agroindustria-y-alimentacion-saludable/monitorizacion-en

16. Schultz MG, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen LH, et al. Can deep learning beat numerical weather prediction? Philosophical Transactions of the Royal Society A [Internet]. 2021 Apr 5 [cited 2021 Oct 28];379(2194). Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0097

17. Corrales DC. Toward detecting crop diseases and pest by supervised learning. Ingenieria y Universidad. 2015 Jul 15;19(1).

18. Sushma B , Suraksha I S,. Disease Prediction of Paddy Crops Using Data Mining and Image Processing Techniques. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. 2016 May 30;5(5).

19. Alam M, Alam MS, Roman M, Tufail M, Khan MU, Khan MT. Real-Time Machine-Learning Based Crop/Weed Detection and Classification for Variable-Rate Spraying in Precision Agriculture. In: 2020 7th International Conference on Electrical and Electronics Engineering (ICEEE). IEEE; 2020.

20. Rincón VJ, Grella M, Marucco P, Alcatrão LE, Sanchez-Hermosilla J, Balsari P. Spray performance assessment of a remote-controlled vehicle prototype for pesticide application in greenhouse tomato crops. Science of The Total Environment. 2020 Jul;726.

21. Gao P, Zhang Y, Zhang L, Noguchi R, Ahamed T. Development of a Recognition System for Spraying Areas from Unmanned Aerial Vehicles Using a Machine Learning Approach. Sensors. 2019 Jan 14;19(2).

22. Kasinathan T, Singaraju D, Uyyala SR. Insect classification and detection in field crops using modern machine learning techniques. Information Processing in Agriculture. 2021 Sep;8(3).

23. India faces its worst locust swarm in nearly 30 years | Asia | An in-depth look at news from across the continent | DW | 27.05.2020 [Internet]. [cited 2021 Oct 28]. Available from: https://www.dw.com/en/india-locusts/a-53579409

24. Scientists turn to tech to prevent second wave of locusts in east Africa | Food security | The Guardian [Internet]. [cited 2021 Oct 28]. Available from: https://www.theguardian.com/global-development/2020/mar/04/scientists-turn-to-tech-to-prevent-second-wave-of-locusts-in-east-africa

25. AEMET OpenData [Internet]. [cited 2021 Nov 4]. Available from: https://opendata.aemet.es/centrodedescargas/inicio

26. Pests Identification | Kaggle [Internet]. [cited 2021 Nov 25]. Available from: https://www.kaggle.com/abhinandanroul/pest-normalized

27. Wu X, Zhan C, Lai Y-K, Cheng M-M, Yang J. IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition. [cited 2021 Nov 30]; Available from: https://github.com/

28. Anaconda Documentation — Anaconda documentation [Internet]. [cited 2021 Nov 30]. Available from: https://docs.anaconda.com/

29. van Rossum G, Drake FL. Python 3 Reference Manual; CreateSpace. Scotts Valley, CA [Internet]. 2009 [cited 2021 Nov 30];242. Available from: https://www.python.org/

30. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016 [Internet]. 2016 [cited 2021 Nov 30];87–90. Available from: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-649-1-87

31. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 585:7825 [Internet]. 2020 Sep 16 [cited 2021 Nov 30];585(7825):357–62. Available from: https://www.nature.com/articles/s41586-020-2649-2

32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research [Internet]. 2011 [cited 2021 Nov 5];12(85):2825–30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html

33. Keras: the Python deep learning API [Internet]. [cited 2021 Nov 30]. Available from: https://keras.io/