DIGITALAGRI

MASTER EN TRANSFORMACIÓN DIGITAL EN EL SECTOR AGROALIMENTARIO Y FORESTAL

TRABAJO FIN DE MÁSTER

MODELOS PREDICTIVOS DE PRODUCCIÓN AGROINDUSTRIAL CON MACHINE LEARNING A PARTIR DE FUENTES DE INFORMACIÓN PÚBLICA

Alumno: **Miguel Ángel Marqués Gozalbo**

Directores: **Adolfo Peña Acevedo**

José Manuel García Nieto

Fecha: Córdoba, 9 de noviembre de 2020

Agradecimientos

En primer lugar, agradezco a mis tutores Adolfo Peña Acevedo y Jose Manuel García Nieto por haberme dado los mejores consejos para llevar a cabo este trabajo, y por estar hasta el último momento aportando valor.

Quiero agradecer a mis compañeros de trabajo en ec2ce, empresa especializada en el desarrollo de soluciones para el sector agroalimentario basadas en Inteligencia Artificial. De ellos, y especialmente de Juan, Jose David y Jose María he aprendido lo poco que se de este extensa y amplia área de conocimiento que es el Machine Learning.

En especial quiero agradecer a mi familia por todo su apoyo, y en especial a mi madre por estar siempre disponible para escuchar.

Resumen

Conocer cuál va a ser el resultado de una campaña de producción agrícola siempre ha sido del interés de los actores del sector agroindustrial. A lo largo de los años, se han desarrollado diferentes diversos métodos para la estimación de cosecha en diferentes cultivos.

Hoy día, gracias a las nuevas tecnologías y a la disponibilidad de datos abiertos sobre el desarrollo de cultivos es posible desarrollar modelos matemáticos que tienen la capacidad de predecir la producción del cultivo meses antes de la cosecha. Países como Estados Unidos (en adelante EEUU) han establecido como política nacional la puesta a disposición de datos sobre el desarrollo de cultivos para que diferentes actores tengan la oportunidad de utilizarlos para desarrollar soluciones que aporten valor a la cadena agroalimentaria.

El principal cultivo en EEUU es el maíz, por lo que existe una extensa variedad de datos abiertos sobre el desarrollo de este cultivo. Variables como índices satelitales, variables climáticas, productividades históricas y muchas más están disponibles para poder desarrollar modelos matemáticos mediante técnicas de Machine Learning.

Este trabajo tiene como objetivos la realización de diferentes casos de uso ligados a la predicción de productividad a partir de datos abiertos. Estos datos son obtenidos de diferentes fuentes de información para posteriormente ser tratados y utilizados como base ajustar modelos con algoritmos de Inteligencia Artificial.

Una vez los modelos están desarrollados se comparan los sistemas actuales de predicción de productividad que ofrece el USDA los meses previos a la cosecha de maíz en EEUU.

Palabras clave: Maíz, Open Data, Machine Learning, Artificial Inteligence, Agricultura de Precisión

Abstract

Knowing what the result of an agricultural production campaign will be has always been of interest to actors in the agro-industrial sector. Over the years, different methods have been developed for yield estimation in different crops.

Today, thanks to new technologies and the availability of Open Data about crop development, it is possible to develop mathematical models that have the capacity to predict crop yield months before the harvest. Countries such as the United States have established as national policy the nececity to share crop development data to allow diferente actor the opportunity to use it to develop solutions that add value to the agrifood chain.

The main crop in the US is maize, so there are a wide variety of open data about the crop evolution. Variables such as satellite indices, climate variables, historical productivity and many more are available to develop mathematical models using Machine Learning techniques.

This work has as objectives the realization of different cases linked to the prediction of productivity from open data. These data are obtained from different sources to be treated and used for adjusting models with Artificial Intelligence algorithms.

Once the models are developed, it is compared with the current productivity prediction systems offered by the USDA in the months before to the corn harvest in the United States.

KeyWords: Maize, Open Data, Machine Learning, Articificial Inteligence, Precision Agriculture

Índice

AGRADECIMIENTOS 3

RESUMEN 5

ABSTRACT 6

ÍNDICE 7

ÍNDICE DE FIGURAS 9

ÍNDICE DE TABLAS 11

INTRODUCCIÓN 12

1.1. Contexto 12

1.2. Motivación 15

1.3. Objetivo 15

MARCO TEÓRICO 17

2.1. Producción de maíz en Iowa 17

2.2. Algoritmos matemáticos para Machine Learning 27

DEFINICIÓN Y METODOLOGÍA DEL TRABAJO 37

3.1. Descripción del problema 37

3.2. Definición del alcance del trabajo 38

3.3. Metodología 39

3.4. Fases del trabajo 40

EXTRACCIÓN, TRANSFORMACIÓN Y ESTRUCTURACIÓN DE LOS DATOS 41

4.1. Identificación de potenciales fuentes de información 41

4.2. Fuentes de información seleccionadas 42

4.3. Tratamiento de los datos 51

4.4. Estructura jerárquica de los datos 53

4.5. Tratamientos de limpieza y reducción dimensional del dataset 55

MODELADO PREDICTIVO Y EXPERIMENTACIÓN 57

5.1. Etapas iterativas del proceso de modelado matemático 57
5.2. Caso de uso 1: Modelos predictivos de maiz 61
5.3. Caso de uso 2: Número mínimo de años 63
5.4. Caso de uso 3: Mes para iniciar las predicciones 64
RESULTADOS Y DISCUSIÓN 66
6.1. Caso de uso 1 66
6.2. Caso de uso 2 79
6.3. Caso de uso 3 80
CONCLUSIONES Y TRABAJOS FUTUROS 87
BIBLIOGRAFÍA 89
ANEXOS 92

Índice de figuras

Figura 1: Producción de maíz en EEUU en 2019. Fuente: USDA....18

Figura 2: Condados del estado de Iowa con código FIPS. Fuente: Universidad de Iowa....20

Figura 3: Productividad (bu/acre) de maíz en grano por condado en el año 2019. Fuente: USDA 20

Figura 4: Productividad media de maíz en grano (bu/acre) del estado de Iowa. Fuente: USDA...21

Figura 5: Sistema radicular del maíz....23

Figura 6: Conjunto de plantas de maíz en desarrollo....24

Figura 7: Flor masculina (derecha) y flor femenina del maíz (izquierda)....24

Figura 8: Mazorca de maíz....25

Figura 9: Calendario de cultivo de maíz en Iowa. Fuente: Universidad de Iowa....25

Figura 10: Representación funcionamiento SVM....30

Figura 11: Algoritmo SVR con diferentes kernels....31

Figura 12: Representación Red Neuronal Artificial....31

Figura 13: Representación funcionamiento algoritmo GPR....32

Figura 14: Representación del funcionamiento del algoritmo RFR....33

Figura 15: Índice NDVI correspondiente al día 15-10-2019 por el satélite MODIS.....45

Figura 16: Posición de las estaciones meteorológicas correspondientes al estado de Iowa y utilizadas en este trabajo....47

Figura 17: Bucle iterativo de modelado matemático....58

Figura 18: Representación gráfica de la división del dataset para el entrenamiento y validación del modelo....60

Figura 19: Importancia de cada atributo seleccionado por el método LassoCV....67

Figura 20: Matriz de correlación del conjunto de datos....68

Figura 21: Representación gráfica de la validación en el año 2019....74

Figura 22: Análisis de los resultados del modelo para los años 2000-2018....75

Figura 23: Diagrama box and whisker de las productividades de los condados para los años 2000-2018.....76

Figura 24: Atributos del modelo distribuidos a lo largo del ciclo de desarrollo fenológico del maíz. 77

Figura 25: Rango del error cometido por el modelo en cada condado para el año 2019.....78

Figura 26: Distritos del estado de Iowa. Fuente: U.S. Department of Agriculture Crop Reporting Districts 82

Figura 27: Comparación entre el error del NASS y del modelo para cada distrito en el 2019.....86

Índice de tablas

Tabla 1: Resumen de las características de las fuentes de información.....51

Tabla 2: Resumen de los indicadores de error de los modelos ajustados.....72

Tabla 3: Evolución de los indicadores del error en el caso de uso 2.....79

Tabla 4: Evolución de los indicadores del error en el caso de uso 3.....80

Tabla 5: Productividad real y estimaciones del NASS para los distritos de Iowa para el año 201983

Tabla 6: Productividad real y predicciones del modelo para los distritos de Iowa en el año 2019..83

Tabla 7: Comparación de los errores del NASS y del modelo respecto a la producción real 85

Capítulo 1
Introducción

1.1. Contexto

Disponer de previsiones de producción de cultivos durante la campaña agrícola es cada vez más importante para que los actores de la cadena agroalimentaria puedan optimizar la toma de decisión en las diferentes operaciones que realizan (Hansen et al., 2004; Hansen and Indeje, 2004; IPCC, 2013; Newlands et al., 2014).

Por eso, que desde hace años se ha intentado desarrollar soluciones que permitan anticiparse al resultado de la cosecha y predecir el resultado final de la producción de los cultivos. Los tres principales acercamientos han sido usados para estudiar esta relación: biofísicos o modelos de simulación de cultivo (Hoogenboom, 2000; Jones et al., 2003; Kotlowski, 2007), modelos de regresión empírica (Kandiannan et al., 2002; Thompson, 1988; Tannura et al., 2008; Lobell et al., 2007) y modelos funcionales (Basso et al., 2013; Chipanshi et al., 2015) que son simplificaciones y/o combinación de los otros dos tipos.

Actualmente, y gracias al uso de nuevas tecnologías y a la gran disponibilidad de datos, se han empezado a desarrollar modelos predictivos de producción en muchos cultivos. Así se citan algunos trabajos a modo de ejemplo donde se han desarrollado modelos predictivos de arroz (Tian et al., 2020), soja (Stepanov et al., 2020; Schwalbert et al., 2020), trigo (Bhojani et al., 2020; Guo et al., 2020), maíz (Jiang et al., 2020; Shahhosseini et al., 2020; Tack et al., 2019; Pede at al., 2019; Ameline et al., 2019).

A parte de los cultivos mencionados, se han desarrollado trabajos similares en muchos otros cultivos, pero es en este último, el maíz, donde existe un mayor

número de trabajos académicos y esfuerzos destinados al desarrollo de soluciones predictivas de producción.

Dado el interés que existe en estas predicciones de producción, están apareciendo empresas que proveen de pronósticos de producción para ayudar a la toma de decisiones en la cadena agroalimentaria. En 2015 y 2016, Descartes Labs 1proporcionó predicciones de rendimiento de maíz disponibles públicamente con unos niveles de error muy bajos. Ese mismo año, el laboratorio se asoció con Cargill2, un gran comerciante de granos, y Descartes dejó de publicar sus predicciones de producción. La asociación se hizo pública en 2018.

En 2016 y 2017, TellusLabs brindó excelentes predicciones del rendimiento del maíz en EE.UU. pero esta dejó de publicar sus predicciones de rendimiento después de fusionarse con IndigoAg3, una importante comercializadora de cereales.

En 2019, el mercado se movió antes de que el USDA4 (United States Department of Agriculture) publicara sus pronósticos. Esto sugiere que algunos actores del sector disponían de predicciones de rendimiento y sabían cuál sería la estimación de rendimiento de NASS5 (National Agricultural Statistics Service).

Estos hechos ponen de manifiesto tres cosas de forma clara: hoy día es posible predecir la producción o rendimiento de los cultivos mejorando las herramientas existentes, el coste de la herramienta es menor al valor que aporta y que esto solo puede realizarse si se disponen de datos sobre el desarrollo de los cultivos y su medio.

Los casos de éxito de las empresas mencionadas en la predicción del rendimiento del cultivo de maíz en EEUU, se debe a que es un país que a través de

[1] Descartes Labs
[2] Cargill
[3] IndigoAg
[4] USDA
[5] NASS

diferentes organismos recopila y facilita el acceso de forma gratuita a gran cantidad de datos históricos sobre el desarrollo de cultivos.

Existen muchas fuentes de información sobre el desarrollo del cultivo que pueden proporcionar información útil para relacionarla con el rendimiento final de la cosecha. La productividad de los cultivos está fuertemente influenciada por diversos factores (genéticos, propiedades del suelo, riego, etc) pero el clima es el factor que no se puede controlar que mayor influencia tiene en el desarrollo de los cultivos (Taylor and Carlson, 1997).

Hay muchas fuentes de información que tienen relación con el desarrollo del cultivo, pero no todas están disponibles en abierto para el público. Si se realiza una recopilación de las fuentes de información utilizadas en trabajos similares (Jiang, Z et al, 2020; Ceglar, A et al, 2018) para el cultivo de maíz se encuentran fuentes de datos satelitales, datos climáticos, datos sobre propiedades del suelo, ubicación de cultivos, índices climáticos, y estadísticas de producción.

Respecto a los datos anteriormente mencionados como base para el desarrollo de modelos predictivos de producción, son los datos sobre estadísticas de producción los que más escasean a nivel global. En EEUU desde distintos organismos se hace un esfuerzo importante en recopilar esos datos y ponerlos a disposición del público para, a partir de ellos, generar valor para el sector agroalimentario.

Gracias a que EEUU realiza una importante labor de recolección de datos sobre distintos estadísticos de producción, muchos de los trabajos sobre el desarrollo de modelos predictivos se realizan ahí, ya que es donde existen datos de calidad para desarrollar e investigar modelos predictivos de producción.

En este trabajo se realizará un acercamiento de modelo empírico de regresión mediante modelado matemático a partir de fuentes de información abiertas para el cultivo de maíz en EEUU. Estos modelos tienen una menor orientación a los procesos físicos de desarrollo de los cultivos y una mayor orientación al uso de

fuentes de datos. Estos modelos son ajustados con datos históricos de desarrollo del cultivo con la menor cantidad de información necesaria para alcanzar los niveles de error necesarios para que aporten valor.

1.2. Motivación

El desarrollo de este trabajo pretender constatar fehacientemente que, a partir de unos conocimientos básicos y con una dotación de medios técnicos a nivel medio, es posible para cualquier persona o empresa el desarrollo de herramientas predictivas de producción que le permitan optimizar sus operaciones en el ámbito de la agricultura de precisión.

Por otro lado, se pretende poner de manifiesto que la inversión pública en generar y poner a disposición del público fuentes de información sobre el desarrollo de cultivos es positiva para la generación de valor y la optimización de la cadena agroalimentaria de cualquier país.

Además, el desarrollo de este trabajo pretense ser un ejercicio completo donde se utilicen diversas tecnologías utilizadas durante la ejecución de máster del que es punto final este trabajo.

1.3. Objetivo

El objetivo general del trabajo es generar una serie de casos prácticos de tratamiento de datos y desarrollo de modelos predictivos a partir de fuentes de información pública, para mostrar su utilidad en la digitalización de procesos productivos en agricultura, sirviendo así de semilla dinamizadora para futuros desarrollos en este sector. Ello conllevaría la necesidad de construir bases de datos robustas para su uso por parte de los actores de la cadena agroalimentaria.

A continuación, se definen los objetivos específicos del TFM:

(1) Generar caso de uso de modelos predictivos de producción a partir de los conjuntos de Open Data para el cultivo de maíz en EEUU;

(2) determinar el número de años mínimo para obtener un buen ajuste de los modelos y;

(3) determinar cuánto tiempo antes de la cosecha es posible predecir la producción con un nivel de certeza adecuado.

Capítulo 2

Marco teórico

En este capítulo se realiza una exposición de las áreas de conocimiento que abarca este trabajo. Por un lado, se contextualiza y caracteriza el cultivo de maíz en EEUU, en concreto para el estado de Iowa, y por otro lado se exponen los algoritmos matemáticos de Machine Learning que serán utilizados para el desarrollo de los modelos de predicción.

2.1. Producción de maíz en Iowa

En esta sección se presenta en primer lugar es estado del cultivo de maíz en EEUU, poniendo especial atención al estado de Iowa. Una vez contextualizado el cultivo en su zona productiva, se expone el ciclo de cultivo del cultivo en las condiciones de producción del estado de Iowa.

2.1.1. El maíz en EEUU y Iowa

El cultivo de maíz en EEUU es el que mayor volumen de superficie ocupa, ya que según el informe del USDA[6] en el año 2019, de un total de 302,6 millones de acres[7] sembrados de los principales cultivos (lo que equivale a 122,4 millones de ha), 89,7 millones de estos fueron dedicados al cultivo de maíz. Esto quiere decir que 3 de cada 10 acres de los principales cultivos en EEUU está ocupado por cultivo de maíz. Esta producción posiciona a EEUU como el primer productor mundial, con un 31,3% de la producción mundial.

[6] Crop Production 2019 Summary

[7] 1 acre = 0,404686 ha

El maíz puede ser sembrado para dos aprovechamientos, para la producción de grano de maíz o para producir maíz para ensilar. En el caso de EEUU la gran mayoría del cultivo va destinado a maíz en grano, y solamente una pequeña parte se siembre para destinarlo a la producción de ensilado de maíz. En el año 2019, de los 89,7 millones de acres sembrados de maíz, 81,4 millones fueron destinados a un aprovechamiento en grano, lo que representa un 90%.

La superficie productiva de maíz se concentra principalmente en el llamado cinturón del maíz o Corn Belt. Típicamente, esta región se ha descrito geográficamente incluyendo a los estados de Iowa, Illinois, Indiana, Nebraska oriental, Kansas oriental, Minnesota meridional y partes de Misuri. En ocasiones también se incluyen partes de Dakota del Sur, Dakota del Norte, Ohio, Wisconsin, Michigan y Kentucky. En cualquier caso, independientemente de cuantas regiones se incluyan, la realidad es que son cuatro estados los que producen más del 50% del total: Iowa, Nebraska, Illinois y Minnesota. En la Figura 1 se puede observar gráficamente la distribución del cultivo de maíz en EEUU viéndose perfectamente el Corn Belt.

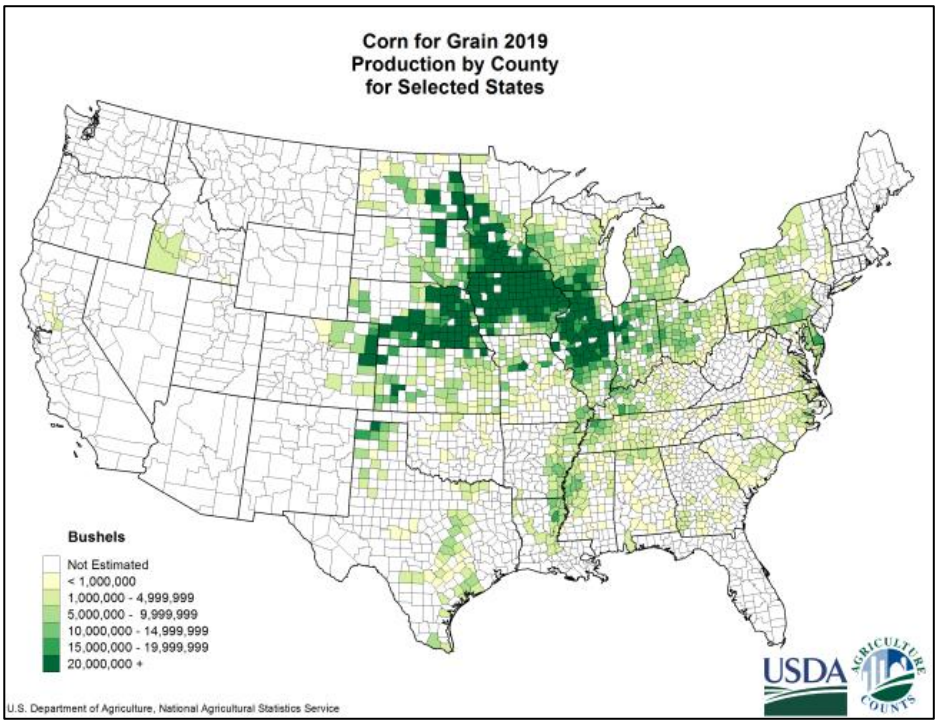
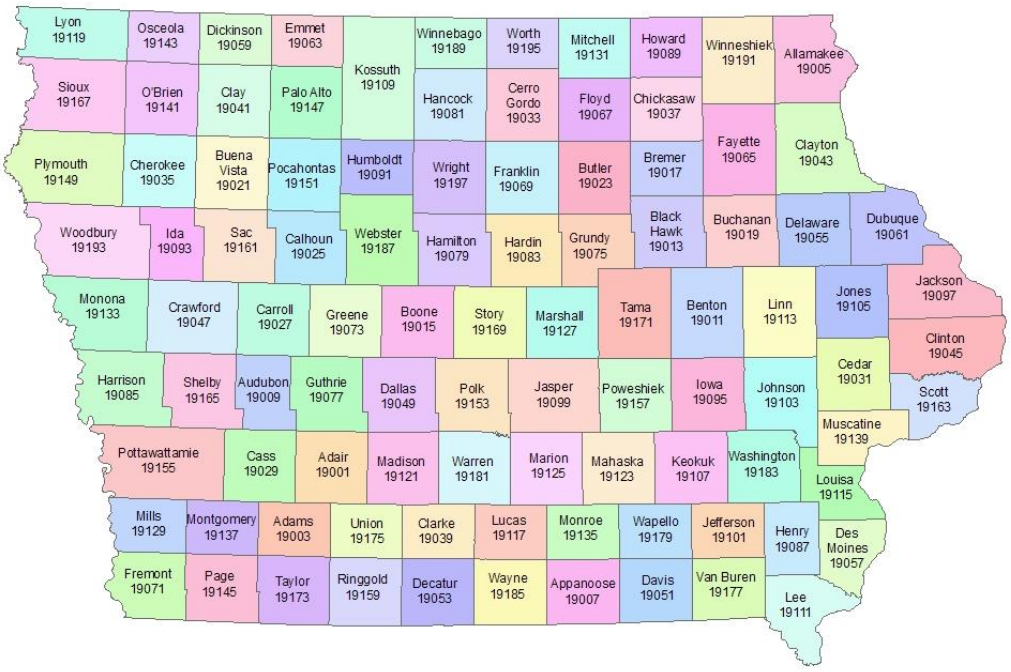


Figura 1: Producción de maíz en EEUU en 2019. Fuente: USDA

El estado de Iowa, con una superficie de 13,5 millones de acres y una producción de 2.583 millones de *bushels*[8] es el máximo exponente de la producción de maíz en EEUU.

El estado de Iowa está dividido administrativamente en 99 condados. Cada uno de estos condados se identifica por un código único, el County FIPS Code, determinado por el ANSI[9] (*American National Standards Institute*). Los códigos FIPS son números que identifican áreas geográficas de forma única. El número de dígitos en los códigos FIPS varía según a nivel de geografía. Los códigos a nivel estatal tienen dos dígitos, los códigos FIPS a nivel de condado tienen cinco dígitos de los cuales los dos primeros son el código FIPS del estado al que pertenece el condado. El código FIPS que corresponde al estado de Iowa es el 19, por lo que todos los códigos FIPS de los condados de Iowa empiezan por 19.

En la Figura 2 puede verse la distribución geográfica de cada uno de los condados dentro del estado de Iowa junto con su código FIPS que lo identifica.



Iowa Counties and Federal Information Processing Standards (FIPS) Codes

[8] 1 bushel de maíz = 25,40 kg de maíz

[9] ANSI

Figura 2: Condados del estado de Iowa con código FIPS. Fuente: Universidad de Iowa

En todos los condados que conforman el estado de Iowa se produce maíz, y existe una variabilidad importante entre cada uno de ellos. Si se analiza el informe[10] que emite el USDA sobre el cultivo de maíz en Iowa correspondiente al año 2019 se puede ver la productividad o rendimiento de cada uno de los condados que conforman el estado. Se puede ver que para el año 2019 la productividad media de maíz en grano del estado de 198,0 bu/acre[11].

Como se puede apreciar en la Figura 3, en el estado de Iowa el rendimiento medio en el año 2019 fue de 198,0 bu/acre. Si se ven las productividades para el mismo año por cada condado es se aprecia bastante dispersión, presentando para una productividad máxima de 234,7 bu/acre en el condado de Crawford y una mínima de 151.5 en el condado de Wayne.

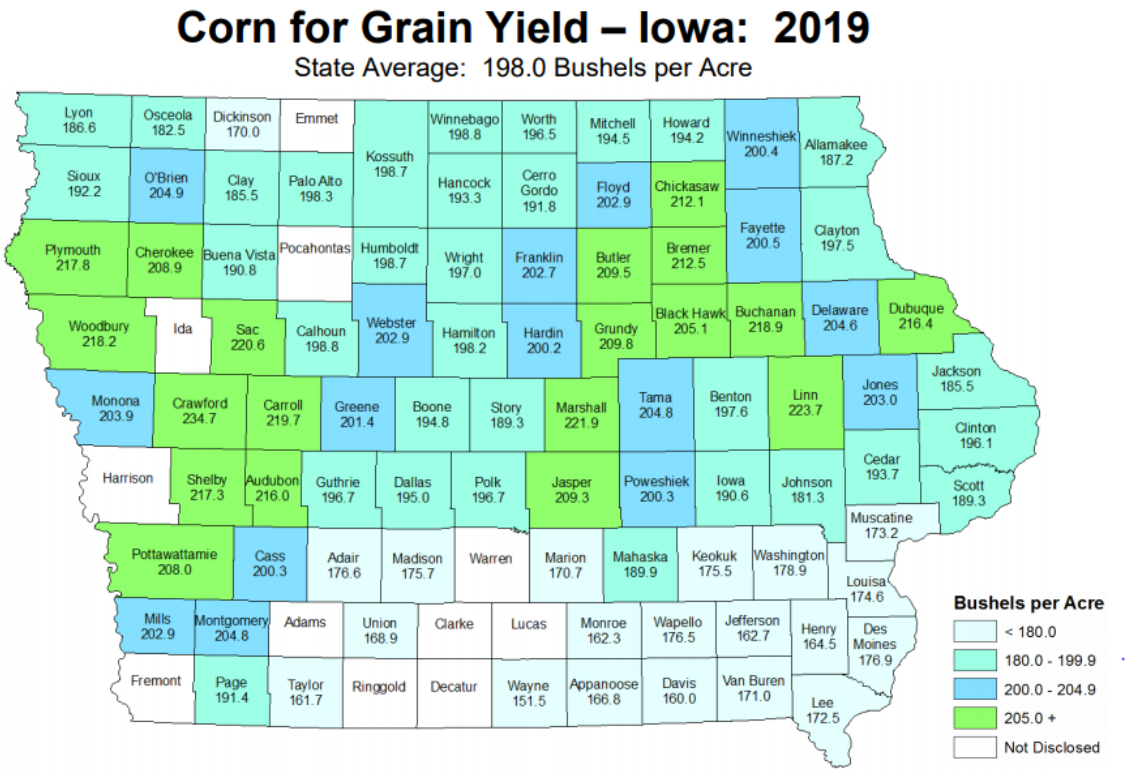


Figura 3: Productividad (bu/acre) de maíz en grano por condado en el año 2019. Fuente: USDA

[10] Iowa Ag News – 2019 Corn County Estimates

[11] 1 bu/acre = 0,06276 ton/ha

Se puede observar que la productividad por condado presenta un rango amplio de valores para el mismo año, y que la productividad media del estado para los últimos 20 años también presenta una variación importante (Figura 4).

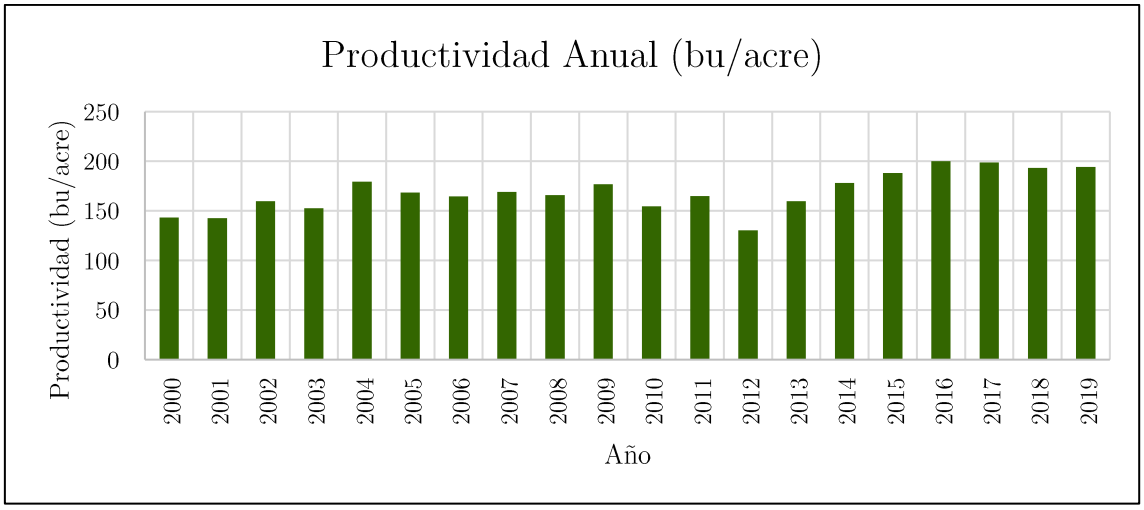


Figura 4: Productividad media de maíz en grano (bu/acre) del estado de Iowa. Fuente: USDA

Como se puede apreciar en la Figura 4, la productividad media máxima del estado de Iowa se alcanzó en el año 2016 con 200,1 bu/acre y la mínima fue en el año 2012 con un valor de 130,3 bu/acre. Queda por tanto patente que existe una importante variabilidad en las productividades anuales en el estado.

2.1.2. El cultivo de maíz

El maíz que se cultiva actualmente poco tiene que ver con su ancestro más lejano, el teocinte. El origen del cultivo del maíz se inició con la domesticación del cultivo en los valles del centro de México (Hufford et al, 2012), de donde más tarde, se extendería por Sudamérica.

Hasta la década de 1920, los productores realizaban una selección de las mazorcas más grandes y con mayor calidad para conservar sus granos, y que sirvieran como semillas para sembrar en la siguiente campaña. De este modo, se fueron creando variedades adaptadas a las condiciones de cada zona productiva. A partir de esos años, se empiezan a desarrollar las primeras semillas híbridas de maíz,

procedentes de cruces genéticos controlados para generar semillas con un potencial productivo más elevado.

Actualmente, la gran mayoría de las semillas de maíz sembradas en el mundo son híbridos desarrollados por importantes empresas productoras de semillas, que han hecho que año tras año, los rendimientos se incrementen debido, en parte, a la mejora genética.

El maíz (*Zea mays*) es una planta anual de gran desarrollo vegetativo, pudiendo alcanzar entre 2 y 2,5 metros de altura. La planta completa se puede dividir en las siguientes partes: raíz, tallo, hojas, flores y frutos.

La raíz posee un sistema radicular fasciculado muy extenso, que está formado por tres tipos diferentes de raíces:

- Raíces primarias: emitidas por la semilla. Comprenden la radícula y las raíces seminales.
- Raíces principales: comienzan a formarse por encima de las raíces primarias y a partir de la corona. Estas constituyen casi la gran parte del sistema radicular.
- Raíces aéreas o adventicias: nacen en último lugar, en los nudos de la base del tallo, por encima de la corona, y sirven como sistema de anclaje de la planta.

El sistema radicular del maíz se desarrolla a partir de la radícula de la semilla, que ha de ser sembrada a una profundidad adecuada para lograr un buen desarrollo posterior.

Figura 5: Sistema radicular del maíz

El tallo del maíz es aproximadamente cilíndrico y está formado por nudos y entrenudos. Los entrenudos son más cortos en la base, y más largos en la parte superior, siendo el más largo el último entrenudo donde se encuentra la base de la espiga.

Las hojas se desarrollan a partir de las yemas foliares. Al inicio, el crecimiento es principalmente apical. Más adelante se van diferenciando los tejidos mediante crecimiento en todos los sentidos hasta adquirir la forma característica de la hoja del maíz. La hoja es larga, angosta, con venación paralelinervia y constituida por la vaina, la lígula y el limbo. Se desarrollan entre 15 y 30 hojas, alargadas y abrazadoras de 4 a 10 cm de ancho por 35 a 50 cm de longitud. Tienen un borde áspero, finamente ciliado y algo ondulado.

Figura 6: Conjunto de plantas de maíz en desarrollo

El maíz es una planta hermafrodita, por lo que produce flores masculinas y femeninas separadas en la misma planta. La flor masculina, llamada panoja, produce polen, mientras que la flor femenina, la mazorca, produce óvulos que se convierten en la semilla.

Figura 7: Flor masculina (derecha) y flor femenina del maíz (izquierda)

El fruto es una mazorca formada por una parte central llamada zuro, corazón o pirulo. Esta parte represente entre el 15% y el 30% del peso del fruto. El fruto está clasificado como cariópside, ya que no se cae de su soporte una vez seco. El grano se adhiere fuertemente al pericarpio. Estos se disponen en hileras

longitudinales, llegando a alcanzar un número de varios centenares en cada mazorca.

Figura 8: Mazorca de maíz

Una vez descrita la fisiología del maíz, se analizan las fases del desarrollo del cultivo donde se realiza este trabajo, en el estado de Iowa.

La Universidad de Iowa ha elaborado una figura que resume las fechas más importantes para el cultivo de maíz en Iowa. En la Figura 9 se pueden apreciar los diferentes estados fenológicos mediante sus curvas de distribución en el tiempo.

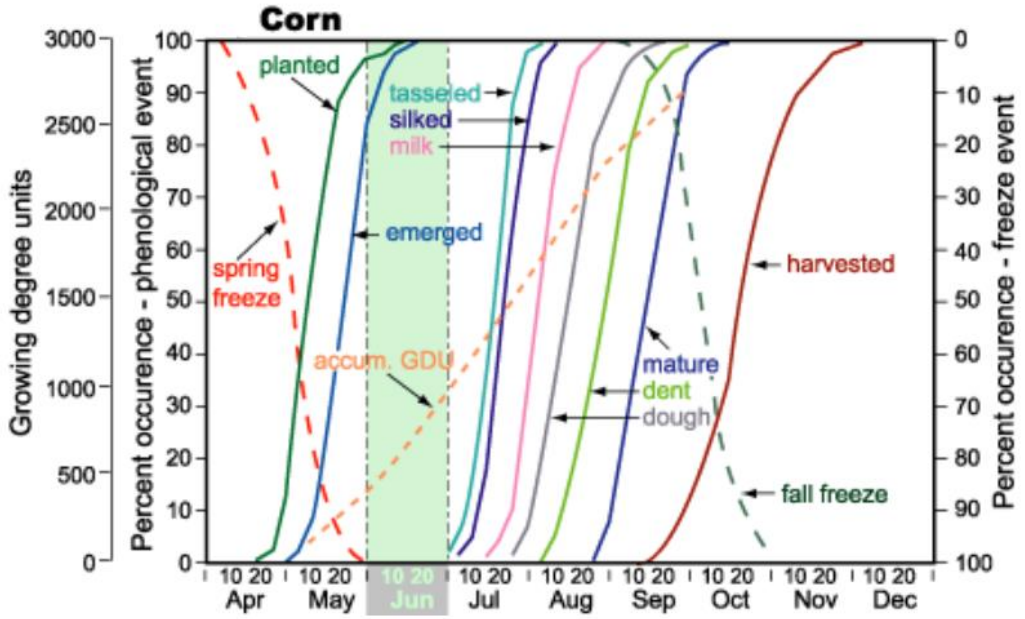


Figura 9: Calendario de cultivo de maíz en Iowa. Fuente: Universidad de Iowa

Este calendario se basa en 25 años de datos climáticos. En el eje horizontal se representan los diferentes meses de desarrollo del cultivo, indicándose con una línea el inicio de cada mes y los días 10 y 20 de cada mes. En el eje vertical, los valores oscilan entre 0 y 100%, lo que indica qué porcentaje de la cosecha ha alcanzado esa etapa. Por ejemplo, la primera línea se corresponde con la siembra. El valor del 50% indica que normalmente la mitad de la cosecha de maíz se ha plantado el 10 de mayo. Por lo general, el 20 de mayo, la cosecha está plantada en más del 90%. Realizando la misma lectura de la curva de cosecha, se observa que el 50% de la cosecha se alcanza a mediados de octubre y a principios de noviembre ya debería estar cosechada más del 90% de la superficie sembrada.

En cada una de las fases del desarrollo del cultivo, este tiene unas necesidades climáticas y nutricionales para tener un correcto desarrollo. Del mismo modo, si el cultivo no se desarrolla dentro de estos rangos adecuados, el desarrollo se verá afectado negativamente, y con él la productividad futura.

El maíz necesita unas temperaturas medias del suelo cercanas a los 10ºC en el momento de la siembra para una correcta nascencia. Para una correcta floración, se recomienda disponer de temperaturas como mínimo de 18ºC. Para el desarrollo, la franja de los 24 a los 30ºC es la más adecuada. Por encima de 30ºC empiezan a producirse problemas en la actividad celular, disminuyéndose la capacidad de absorción de agua. La diferencia de temperatura entre día y la noche favorece el crecimiento rápido del cultivo. Las altas temperaturas durante la noche hacen que la planta consuma demasiada energía realizando la respiración celular, restando así energía para el llenado de los granos durante el día. Las heladas antes de la maduración, cuando todavía no se han trasformado los azúcares del grano en almidón, hacen que se interrumpa el proceso de forma irremediable, produciendo una considerable pérdida de productividad.

El maíz se adapta bien a diferentes tipologías de suelos, siendo poco recomendables suelos arcillosos por su facilidad de inundación, o suelos muy arenosos al ser propensos a secarse rápidamente y almacenar poca cantidad de agua. Por lo general, los mejores suelos para el cultivo de maíz son los de textura franca, profundos, fértiles y con elevada capacidad de almacenamiento de agua. Una profundidad de 60 cm es adecuada para su cultivo, pero si los suelos son más profundos mejoran las productividades. El cultivo tolera suelos con pH entre 5.5 y 8, aunque las mayores productividades se dan con suelos ligeramente ácidos. Respecto a la salinidad del suelo, el maíz es medianamente tolerante, siendo el umbral de 153 g/l en la solución del suelo el límite para no provocar perdidas de producción.

2.2. Algoritmos matemáticos para Machine Learning

El objetivo de este trabajo es desarrollar modelos para obtener predicciones de la productividad del cultivo de maíz mediante un enfoque Machine Learning. Este enfoque requiere la utilización de algoritmos matemáticos para realizar tareas como selección de atributos, entrenamiento de los modelos y evaluación del error.

Todos los algoritmos utilizados en este trabajo han sido extraídos de la biblioteca especializada en Machine Learning llamada Scikit-Learn[12]. A continuación, se hace una descripción de los algoritmos utilizados para cada una de las tareas.

2.2.1. Selección de parámetros o atributos

Para realizar la tarea de selección de atributos se ha optado por el método LASSO.

[12] Scikit-Learn

LASSO (Least Absolute Shrinkage and Selection Operator) fue formulado por Robert Tibshirani en 1996. LASSO basa su funcionamiento a partir de modelos de regresión lineal. A continuación, se describe en funcionamiento de LASSO.

En un modelo de regresión lineal, para representar la dependencia de una variable y (variable dependiente o variable respuesta) con respecto a otras variables X' (variables independientes o variables explicativas) se puede expresar como:

$$y = \alpha + X'\beta' + \epsilon$$

Donde α es el término independiente, β' es el parámetro que acompaña a cada variable explicativa y que se quiere estimar y ϵ es el error.

La estimación de los parámetros β' se realiza mediante el método de mínimos cuadrados. Así se escoge como parámetro β' aquel donde se alcanza:

$$min_{\beta'} \sum_{i=1}^{n} (y_i - X'_i\beta')^2$$

Siendo y_i la observación de la variable respuesta e X'_iel valor de cada variable explicativa en un conjunto de n observaciones. A partir de este concepto, LASSO mejora el método aplicando una penalización para obligar a que algunos componentes β' sean 0. De este modo, los β' que no resulten 0 serán los correspondientes a las variables que se deben mantener, y el resto deben no ser consideradas en el conjunto de datos de entrenamiento de los modelos, logrando de este modo la selección de atributos.

Así, LASSO es un método que permite de forma simultánea la selección de variables y la estimación de los parámetros de las variables seleccionadas. El algoritmo de la librería Scikit-Learn utilizado es LassoCV[13].

[13] LassoCV

2.2.2. Entrenamiento de los modelos

Existen multitud de modelos matemáticos que se pueden utilizar a la hora de entrenar el modelo matemático predictivo. Cada modelo matemático tiene una naturaleza diferente, con una serie de particularidades, que impiden que a priori se sepa que modelo será el más adecuado para el problema a resolver. Es por esto, que en este trabajo se entrenarán diferentes modelos con el mismo conjunto de datos para ver cuál de ellos presenta los menores errores.

A continuación, se hace una pequeña mención a cada uno de los algoritmos de entrenamiento utilizados:

Dummy Regressor - DR

DummyRegressor (DR)[14] es un algoritmo de regresión que hace predicciones usando reglas simples. Este algoritmo es útil como una línea de base simple para comparar con otros regresores (reales). Para considerar que otro algoritmo tiene un buen nivel de ajuste, debe tener mejores resultados que el DummyRegresor.

Este algoritmo permite diferentes configuraciones, por lo que se le puede configurar para que la predicción que realice sea la media, la mediana o un percentil concreto del conjunto de datos de entrenamiento.

Linear Regression – LR

En estadística la regresión lineal (LR)[15] es un modelo matemático utilizado para aproximar la relación entre la variable dependiente (y) y las variables independientes (X') y un término asociado al error. El modelo se puede expresar como:

$$y = \alpha + \sum X'\beta' + \epsilon$$

[14] DummyRegressor
[15] LinearRegression

El algoritmo busca los parámetros β' para cada variable independiente que forme parte del proceso de modelado de manera que se minimice el error del modelo.

Support Vector Regression - SVR

Las máquinas de vectores soporte (SVM, del inglés Support Vector Machines) tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los años 90 por Vapnik y sus colaboradores. Aunque originariamente las SVMs fueron pensadas para resolver problemas de clasificación binaria, actualmente se utilizan para resolver diversos tipos de problemas, por ejemplo, la regresión, donde se utiliza el algoritmo SVR[16].

La idea es seleccionar un hiperplano de separación que equidiste de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento que distan del hiperplano la distancia margen. Estos ejemplos reciben el nombre de vectores soporte. En la Figura 10 puede verse como funciona gráficamente un algoritmo SVM.

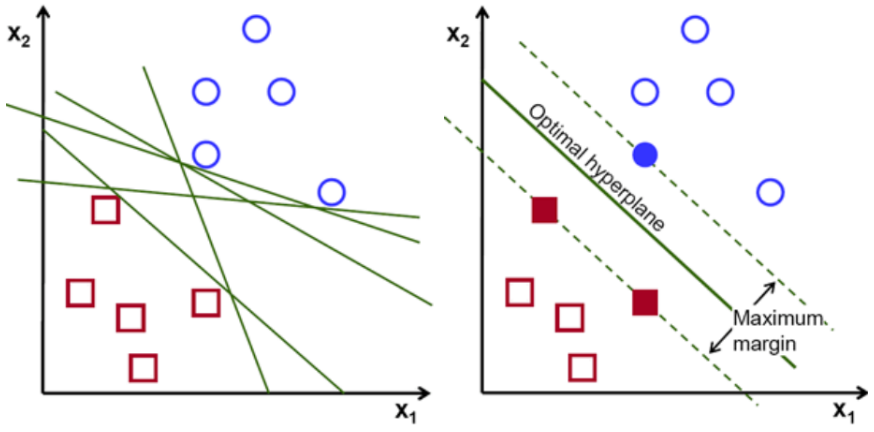


Figura 10: Representación funcionamiento SVM

[16] Support Vector Regression

En el caso los algoritmos SVM tengan la misión de diferenciar conjuntos de datos no linealmente separables, modificando el hiper-parámetro kernel es posible que realicen la tarea. En la Figura 11 puede verse como se modifica al comportamiento del modelo ajustado con el algoritmo SVR en función del kernel que se utilice.

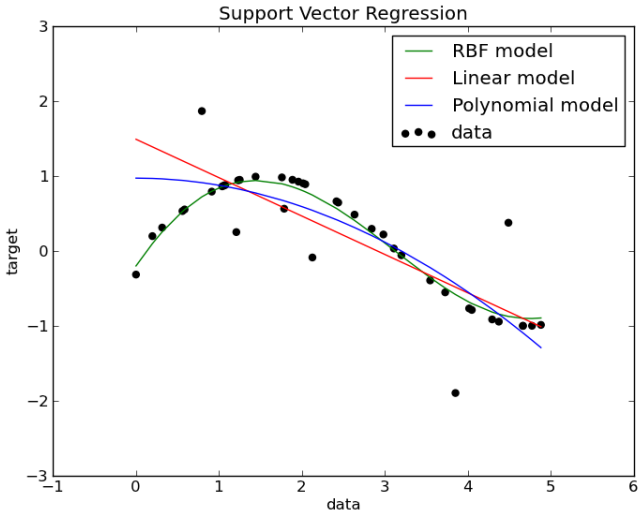


Figura 11: Algoritmo SVR con diferentes kernels

Multi-layer Perceptron Regresor - MLPR

El Multi-layer Perceptron es una red neuronal artificial (RNA) formada por múltiples capas, de tal manera que tiene capacidad para resolver problemas no linealmente separables. El perceptrón puede estar conectado con todas las siguientes capas o solamente con algunas.

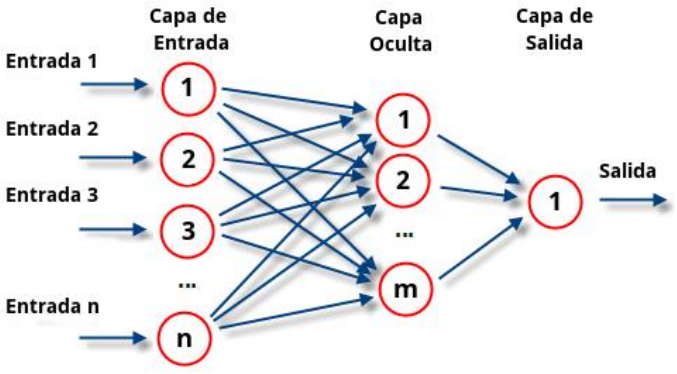


Figura 12: Representación Red Neuronal Artificial

Para la resolución de problemas de regresión se utiliza el algoritmo MLPR[17]. Este debe constar de al menos tres capas, una de entrada, una oculta y una de salida. El algoritmo utiliza la técnica de backpropagation para conseguir realizar el aprendizaje.

Gaussian Process Regressor - GPR[18]

En teoría de probabilidad y estadística, un proceso gaussiano es un proceso estocástico (una colección de variables aleatorias indexadas por tiempo o espacio), de modo que cada colección finita de esas variables aleatorias tiene una distribución normal multivariante. Es decir, cada combinación lineal finita de ellas es normalmente repartida. La distribución de un proceso gaussiano es la distribución conjunta de todas esas (infinitas) variables aleatorias, y como tal, es una distribución sobre funciones con un dominio continuo, como por ejemplo tiempo o espacio.

El concepto de procesos gaussianos lleva el nombre de Carl Friedrich Gauss porque se basa en la noción de distribución gaussiana (distribución normal). Los procesos gaussianos pueden verse como una generalización de dimensiones infinitas de distribuciones normales multivariadas. En la Figura 13 puede verse como funciona un algoritmo GPR para un conjunto de datos dado.

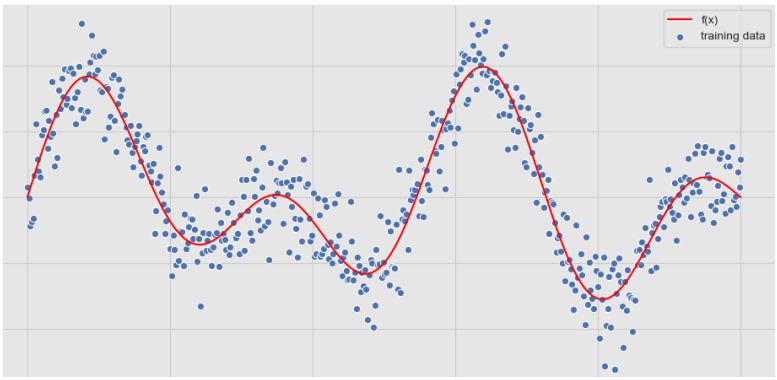


Figura 13: Representación funcionamiento algoritmo GPR

[17] MLPR
[18] GPR

Random Forest Regressor - RFR

Un Random Forest (Bosque Aleatorio), es una técnica de aprendizaje automático muy popular. Los Random Forests tienen la capacidad de resolver problemas de clasificación o regresión. Para las tareas de regresión se utiliza el algoritmo Random Forest Regresor (RFR) [19].

Los RandomForest constan de una gran cantidad árboles de decisión, cada uno de ellos construido sobre una extracción aleatoria de las observaciones del conjunto de datos y una extracción aleatoria de las características. No todos los árboles ven todas las características o todas las observaciones, y esto garantiza que los árboles estén descorrelacionados y, por lo tanto, sean menos propensos a un ajuste excesivo. La salida para cada observación será el promedio de los árboles individuales. Tienen tanto éxito porque proporcionan en general un buen rendimiento predictivo, un bajo sobreajuste y una fácil interpretación. En la Figura 14 puede verse una representación del funcionamiento de un RFR.

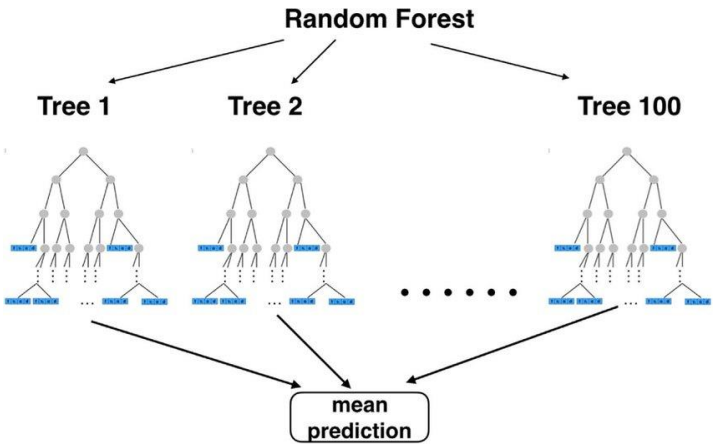


Figura 14: Representación del funcionamiento del algoritmo RFR

Extremely Randomized Tree Regressor - ETR

Los árboles extremadamente aleatorios (Extra-Trees o Extremely Randomized Trees o ETR) [20] llevan la aleatoriedad de Random Forest un paso más allá. Además

[19] RFR
[20] ETR

de considerar un subconjunto de las características predictivas para cada uno de los árboles a crear, a la hora de escoger una característica y un valor de corte (threshold) para dividir cada nodo, en lugar de escoger el threshold que mejor divida cada característica (y escoger la característica y valor de corte que minimice el criterio de impureza), se genera un valor de corte aleatorio para cada característica planteada, escogiéndose como regla de división el mejor de ellos. Otra diferencia con Random Forest es que las muestras con las que se entrena cada aprendiz se escogen sin reemplazo (o, como suele decirse, no son observaciones "bootstrap").

GridSearchCV

Todos los modelos que se han mencionado anteriormente tienen la posibilidad de modificar sus propiedades mediante diferentes hiper-parámetros que son propios de cada modelo. A priori, es imposible saber que combinación de hiper-parámetros ofrecerá el mejor ajuste de ese modelo.

Para realizar un ajuste automático de los hiper-parámetros específicos de cada modelo, se utiliza el algoritmo GridSearchCV[21]. Este algoritmo permite probar de forma iterativa todas las combinaciones de hiper-parámetros para cada algoritmo de manera que se obtiene la mejor combinación de estos para minimizar el error del modelo.

El algoritmo realiza una división del conjunto de entrenamiento en cinco partes, para mediante validación cruzada determinar el valor del ajuste global para esa combinación de parámetros.

2.2.3. Evaluación de los modelos

Para realizar la evaluación del ajuste de los modelos, es necesario calcular diferentes parámetros que permitirán conocer la bondad de cada uno realizando la tarea de predicción. En la librería que se está utilizando en este trabajo, Scikit-

[21] GridSearchCV

Learn existe una colección de algoritmos que permiten calcular los diferentes parámetros de una forma muy sencilla para modelos de regresión. A continuación, se describen los algoritmos utilizados para calcular los diferentes parámetros del ajuste de los modelos predictivos.

Coeficiente de varianza explicada

Siendo y los valores correctos de las observaciones, e y' los valores predichos para dichas observaciones, el coeficiente de varianza se calcula del siguiente modo. El mejor resultado es 1, y cuanto más pequeño peor.

$$varianza\ explicada\ (y, y') = 1 - \frac{Var\ (y - y')}{Var(y)}$$

Error máximo

La función calcula el error residual máximo, una métrica que captura el error del peor de los casos entre el valor predicho y el valor verdadero.

$$MaxError\ (y, y') = \max\left(\left|y_i - y'_i\right|\right)$$

MAE - Error Absoluto Medio

El error absoluto medio mide cual es la diferencia promedio para todas las observaciones entre el valor predicho y el valor verdadero.

$$MAE\ (y, y') = \frac{\sum_{i=0}^{n} |y_i - y'_i|}{n_{observaciones}}$$

R^2

El coeficiente R^2 determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo.

$$R^2 = \frac{{\sigma^2}_{XY}}{{\sigma^2}_X\ {\sigma^2}_Y}$$

Donde:

- ${\sigma^2}_{XY}$ es la covarianza de (X,Y)

- σ^2_X es la varianza de la variable X
- σ^2_Y es la varianza de la variable Y

MAPE – Error Porcentual Absoluto Medio

El MAPE pese a no disponer de un algoritmo para su cálculo directo, es un indicador muy utilizado en la evaluación de modelos predictivos. El MAPE es el MAE dividido por la media de los valores reales de las observaciones. De este modo, se puede contextualizar el MAE dentro de su entorno de valores.

Capítulo 3
Definición y metodología del trabajo

En este capítulo se realiza una definición del problema que se va a abordar en este trabajo. Una vez definido el problema, se enmarca el trabajo definiendo el alcance del mismo. Finalmente, se realiza una descripción de la metodología seguida para llevar a cabo el trabajo.

3.1. Descripción del problema

El tratar de conocer cuál será la evolución de los cultivos ha sido un anhelo de los productores agrarios desde tiempos inmemoriales. Para afrontar esta necesidad, los actores del sector tradicionalmente han desarrollado soluciones basadas en su experiencia, modelos empíricos, consultores expertos en la materia o han tenido acceso a estimaciones de producción generadas por organismos públicos competentes en la materia.

Hoy día, gracias a que se dispone de multitud de datos sobre la evolución del cultivo para los últimos años, se abre la oportunidad de utilizarlos y mediante técnicas de análisis de datos, ser capaces de desarrollar modelos predictivos de producción. Esto se puede hacer a nivel particular para las explotaciones productivas propias o a nivel regional para conocer de antemano la producción de un cultivo en una zona determinada.

Como se ha mencionado anteriormente, EEUU es un país que a través de diferentes organismos públicos recoge y pone a disposición del público distintos tipos de datos sobre el desarrollo de los cultivos del país.

Por ello, actores del sector tienen la oportunidad de desarrollar soluciones predictivas de producción a partir de fuentes de información públicas que les permitan disponer de una ventaja competitiva frente al resto de actores de la cadena agroalimentaria. Este hecho permite optimizar la toma de decisiones en el negocio y maximizar el resultado económico de cada operación.

3.2. Definición del alcance del trabajo

Como se ha mencionado, EEUU pone a disposición del público importantes bases de datos que permiten desarrollar modelos predictivos de producción de cultivos, por lo que se opta por desarrollar este trabajo sobre cultivos de dicho país.

El principal cultivo del país es el maíz, por lo que se determina, que al ser este el principal cultivo de interés en EEUU, los datos disponibles sobre sus producciones históricas u otras variables de su desarrollo serán de mayor calidad comparadas con otros cultivos.

Dado el tamaño de EEUU, se ha estimado oportuno limitar el trabajo al estado de Iowa, que se corresponde con el estado americano líder en cultivo y producción de maíz. Como ya se ha mencionado, el estado de Iowa está formado por 99 condados, siendo este el nivel de granularidad espacial menor al que se trabajará, dado que los datos de producción histórica están recogidos a esta escala.

Así, una vez enmarcado el trabajo, la definición del alcance del trabajo se puede resumir en los siguientes puntos:

- País: Estados Unidos
- Estado: Iowa
- Cultivo: Maíz Grano
- Granularidad espacial: Condado

3.3. Metodología

La metodología usada en este proyecto ha sido KDD, conocida como proceso de descubrimiento. Esta metodología tiene unas etapas establecidas que no siempre siguen el mismo patrón, pero que si tienen la misma filosofía.

A continuación, se describe cada una de las etapas:

1. **Abstracción del escenario**

En esta etapa inicial se tiene que dejar de lado las matemáticas y la estadística y entender el problema que hay que resolver, para contextualizarlo bien y proponer soluciones viables y reales. Es importante conocer las propiedades, limitaciones y reglas del escenario en estudio, para posteriormente definir las metas a alcanzar.

2. **Análisis de las fuentes**

La etapa de análisis consta de dos etapas: descubrimiento de las fuentes de datos y evaluación de las mismas. Se realiza un análisis de las fuentes a las que se puede tener acceso y se las caracterizan, teniendo en cuenta su naturaleza.

3. **Selección de datos**

En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Una vez determinados los datos que serán utilizados, se realiza la extracción de los mismos desde la fuente de origen.

4. **Limpieza y pre-procesamiento**

En esta etapa se analiza la calidad de la información para garantizar que será de utilidad en el proceso de modelado. En el caso que sea necesario, se realizará un tratamiento a los datos para garantizar que cumplen con los requisitos y formato que se establezca.

5. **Transformación de los datos**

En esta etapa se mejora la calidad de los datos con transformaciones que pueden ser de reducción de la dimensionalidad del conjunto de datos o transformación de algunas variables en categorías.

6. **Modelado matemático**

En esta etapa se selecciona el algoritmo que mejores prestaciones ofrece para la tarea que se le asigna. Mediante diferentes pruebas de validación, se determina el algoritmo, junto con una determinada configuración de sus hiper-parámetros, ofrece mejores los indicadores de error.

7. **Interpretación y evaluación**

En esta etapa, se analizan los resultados que ofrece el modelo seleccionado, interpretando sus resultados y realizando una evaluación de los mismos.

8. **Presentación del conocimiento**

En esta etapa y una vez se dispone de los resultados, se busca la forma óptima de presentarlos para que el conocimiento se transmita de una forma sencilla y efectiva.

3.4. Fases del trabajo

Las etapas de la metodología que se ha seguido en el trabajo se pueden agrupar en dos fases.

1. Fase 1: Abarca desde la concepción inicial del trabajo hasta que se alcanza a configurar el dataset que contiene toda la información que será utilizada posteriormente como el elemento base de la siguiente fase.
2. Fase 2: Empieza a partir de los trabajos iniciales de limpieza y modelado matemático a partir del dataset estructurado en la anterior fase y termina en la presentación de los resultados de los modelos predictivos desarrollados para dar cumplimiento a los objetivos de este trabajo.

Capítulo 4
Extracción, transformación y estructuración de los datos

En este capítulo se realiza una exposición de los trabajos correspondientes a la identificación y caracterización de fuentes de información, obtención de los datos y las tareas necesarias para estructurar la información en un conjunto de datos válido para ser utilizado en un proceso de modelado matemático.

4.1. Identificación de potenciales fuentes de información

Cualquier trabajo relacionado con el modelado matemático parte de un mismo punto, que es, identificar y caracterizar las fuentes de información que potencialmente pueden ser utilizadas. A la hora de realizar la inversión en obtención de la información es muy importante tener presente la relación entre el coste de obtención de la información, en tiempo y recursos económicos, y el valor que aporten al modelo para minimizar su error.

Si se piensa en que fuentes de información se pueden utilizar para que formen parte del conjunto de datos de entrenamiento del modelo matemático se pueden dividir en dos clases: variables de causa y variables de efecto. Las variables de causa son aquellas que producen una reacción en el desarrollo del cultivo, y las variables de efecto, son aquellas que son producto de la evolución del cultivo. Además de variables sobre el desarrollo del cultivo también es necesario disponer de información geográfica para identificar la geometría y ubicación de los condados del estado de Iowa.

Para la ejecución de este trabajo se ha tomado como premisa limitar el coste de adquisición de la información, manteniendo un nivel aceptable de calidad, pero sin llegar a tener la máxima calidad de información de cada variable ni la totalidad de las variables que se podrían utilizar. En todo proyecto de modelado, es recomendable realizar un primer modelo básico, y una vez desarrollado el trabajo de modelado y analizada la bondad del modelo, realizar un trabajo de mejora haciendo un esfuerzo en mejorar la calidad de las fuentes de información hasta alcanzar el nivel de error que sea necesario para la toma de decisiones del usuario que utilice el modelo.

Es por esto, que de todas las potenciales fuentes de información se han seleccionado aquellas que ofrecen una información valiosa para la ejecución del trabajo pero que son de fácil acceso y manejo mediante medios tecnológicos convencionales.

4.2. Fuentes de información seleccionadas

A continuación, se describen y caracterizan las fuentes de información utilizadas para constituir el conjunto de datos sobre el que será ajustado el modelado matemático.

4.2.1. Productividad del cultivo

El Departamento de Agricultura de los Estados Unidos (USDA), a través de su servicio nacional de estadísticas agrícolas (NASS) pone a disposición del público mediante una plataforma online un conjunto de estadísticas de producción de distintos cultivos.

Mediante la plataforma QuickStats[22] habilitada para la selección y acceso a los datos, de una forma sencilla e intuitiva, se puede determinar el conjunto de

[22] QuickStats

datos que se quiere descargar. Para determinar el conjunto de datos a los que se quiere acceder hay que seleccionar entre las diferentes opciones que se ofrecen en función de tres tipos de opciones. La descarga de los datos seleccionados se realiza mediante un archivo en formato CSV.

En el caso de este trabajo la selección del conjunto de datos ha sido la siguiente:

- **Select Commodity**
  - Program: Survey
  - Sector: Crops
  - Group: Field Crops
  - Commodity: Corn
  - Catagory: Yield
  - Data Item: Corn, Grain – Yield, Measured in Bu/Acre
  - Domain: Total
- **Select Location**
  - Geographic Level: County
  - State: Iowa
  - Ag District: All
  - County: All (99)
- **Select Time**
  - **Year: 2000-2019**
  - **Period Type: Annual**
  - **Period: Year**

4.2.2. Información SIG

Para poder ubicar en el espacio cada uno de los 99 condados que forman el estado de Iowa es necesario disponer de una cartografía. Para ello, se puede acceder a la web gubernamental de Iowa GeoData[23], donde se puede descargar a un archivo con la geometría de los condados del estado.

4.2.3. Índices satelitales

Los índices satelitales del desarrollo del cultivo se han obtenido a partir del conjunto datasets de datos satelitales que proporciona el LP DAAC[24] (Land Processes Distributed Active Archive Center). El LP DAAC opera como una asociación entre el Servicio Geológico de los Estados Unidos (USGS) y la Administración Nacional de Aeronáutica y del Espacio (NASA) y es un componente del Sistema de Información y Datos del Sistema de Observación de la Tierra (EOSDIS) de la NASA.

Dentro de los múltiples datasets que están disponibles, se ha elegido MOD13A2 V6. Este producto proporciona con una resolución de 1 km, dos índices de vegetación: el índice de vegetación de diferencia normalizada (NDVI) y el índice de vegetación mejorado (EVI). El algoritmo de este producto elige el mejor valor de píxel disponible de todas las adquisiciones del período de 16 días. Los criterios utilizados son nubes bajas, ángulo de visión bajo y el valor más alto de NDVI / EVI.

Para tener acceso y obtener los datos satelitales se ha utilizado la herramienta Google Earth Engine[25], que facilita el acceso de una forma sencilla un conjunto de datasets sobre distintas variables obtenidas a partir de distintas constelaciones de satélites, como MODIS, Landsat o Sentinel, y entre ellos, el que se ha seleccionado

[23] Iowa GeoData
[24] LP DACC
[25] Google Earth Engine

para este trabajo. Utilizando la librería expresamente desarrollada llamada EE (Earth Engine) para el acceso y manipulación de estos datasets se puede acceder a los datos necesarios.

El código Python[26] para tener acceso a los datos satelitales se ha sido escrito en Google Colab, que es un servicio cloud, basado en los Notebooks de Jupyter, que permite el uso gratuito de las GPUs y TPUs de Google. El código Python utilizado puede verse en el Anexo 1. En la Figura 15 se puede ver una composición de las imágenes satelitales obtenidas por el MODIS donde se representa el índice NDVI en cada condado del estado de Iowa.

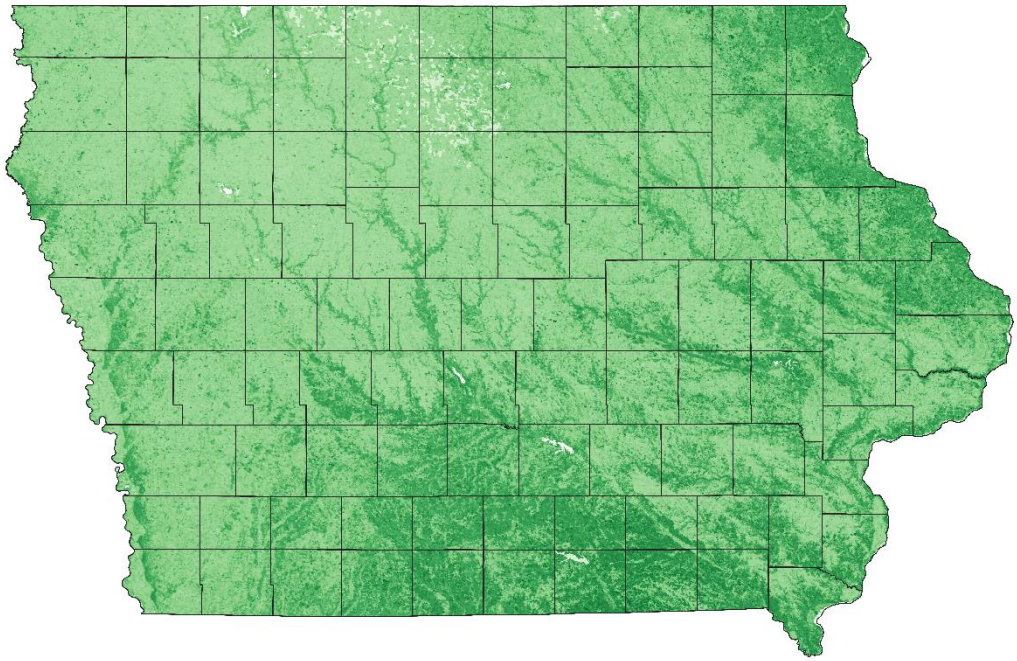


Figura 15: Índice NDVI correspondiente al día 15-10-2019 por el satélite MODIS.

4.2.4. Estaciones meteorológicas

Para acceder a información relativa a los factores climáticos a los que están sometidos los cultivos a lo largo de su desarrollo se utilizará una fuente de datos proporcionada por la Administración Nacional del Océano y la Atmósfera

[26] Python

(NOAA)[27]. En su página web se puede localizar el servicio de Datos Climáticos Online (CDO)[28] que proporciona acceso gratuito a los datos meteorológicos históricos de las estaciones meteorológicas que existen en la red. Estos datos incluyen mediciones diarias, mensuales, estacionales y anuales de temperatura, precipitación, viento y grados día, así como datos de radar y normales climáticas de 30 años.

Existe una opción para acceder a un resumen mensual de las variables analizadas en cada estación meteorológica. Los datos mensuales se calculan a partir de los datos de entrada utilizando el conjunto de datos de diario de la Red de Climatología Histórica Global (GHCN)[29]. Para este trabajo, se ha utilizado el resumen de los datos diarios que están ya calculados, ya que facilita el tratamiento de los mismos. Así, se evita tener que realizar un resumen de los registros diarios, ya que no es recomendable introducir los registros diarios en el modelo, sino un resumen mensual o semanal de los mismos.

Dependiendo del condado seleccionado y las características de las estaciones meteorológicas instaladas, las variables son distintas. Es por esto, que a la hora de descargar la información se optará por descargar siempre toda la información disponible. Más adelante se tratará el asunto de no poder disponer de la totalidad de variables para todos los condados. En la Figura 16 se pueden ver las ubicaciones de las estaciones meteorológicas utilizadas en este trabajo.

[27] NOAA
[28] CDO
[29] GHCN

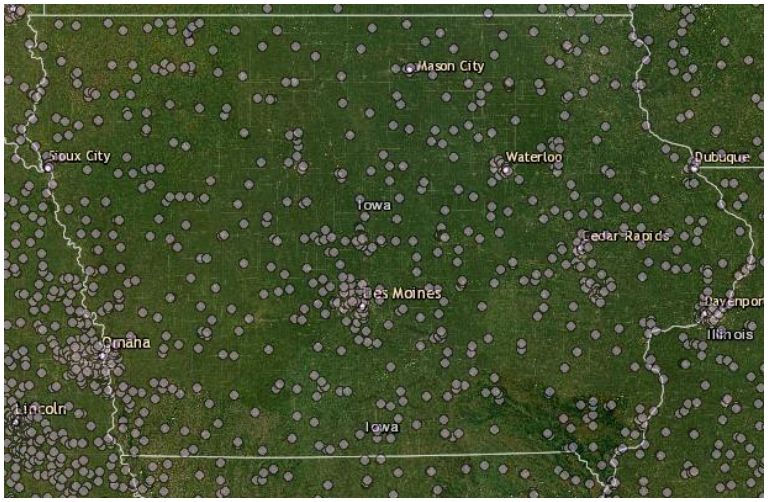


Figura 16: Posición de las estaciones meteorológicas correspondientes al estado de Iowa y utilizadas en este trabajo

A continuación, de describen las variables que están disponibles para todos los condados y que han sido utilizadas en este trabajo:

- Cooling Degree Days (CLDD) – Sumatorio del número de días en que la temperatura media del día ha sido inferior a 18,3ºC
- Heating degree days (HTDD) – Sumatorio del número de días en que la temperatura media del día ha sido superior a 18,3ºC
- Number days with maximum temperature < 32 F. (DX32) – Sumatorio del número de días en que la temperatura máxima ha sido superior a 0ºC
- Number days with maximum temperature > 70 F (21.1C) (DX70) – Sumatorio del número de días en que la temperatura máxima ha sido superior a 21,1ºC
- Number days with maximum temperature > 90 F (32.2C) (DX90) – Sumatorio del número de días en que la temperatura máxima ha sido superior a 32,2ºC

- Number days with minimum temperature less than or equal to 0.0 F (DT00) – Sumatorio del número de días en que la temperatura mínima ha sido inferior a -17,8ºC
- Number days with minimum temperature less than or equal to 32.0 F (DT32) – Sumatorio del número de días en que la temperatura mínima ha sido inferior a 0ºC
- Number of days with greater than or equal to 0.1 inch of precipitation (DP01) – Sumatorio del número de días con precipitación superior a 0,254 mm
- Number of days with greater than or equal to 1.0 inch of precipitation (DP10) – Sumatorio del número de días con precipitación superior a 2,54 mm
- Extreme maximum precipitation for the period. (EMXP) – Precipitación máxima en mm registrada en el periodo
- Extreme maximum snow depth for the period. (EMSD) – Acumulación de nieve máxima en mm registrada en el periodo
- Extreme maximum snowfall for the period. (EMSN) – Nevada máxima en mm registrada en el periodo
- Number days with snow depth > 1 inch(25.4mm) for the period. (DSND) – Sumatorio del número de días con una capa de nieve superior a los 25,4 mm
- Number days with snowfall > 1 inch. (DSNW) – Sumatorio del número de días con nevadas superiores a los 25,4 mm
- Precipitation (PRCP) – Precipitación total acumulada en mm en el periodo

- Snowfall (SNOW) – Nevadas totales acumuladas en mm en el periodo
- Average Temperature. (TAVG) – Temperatura media en ºC en el periodo. Calculada como la media entre la temperatura máxima media y la temperatura mínima media del perido.
- Cooling Degree Days Season to Date (CDSD) – Sumatorio de grados de fríos acumulados desde inicio de año hasta el periodo en ºC.
- Extreme maximum temperature for the period. (EMXT) – Temperatura máxima diária registrada en ºC en el periodo.
- Extreme minimum temperature for the period. (EMNT) – Temperatura mínima diária registrada en ºC en el periodo
- Heating Degree Days Season to Date (HDSD) – Sumatorio de grados de calor acumulados desde inicio de año hasta el periodo en ºC.
- Maximum temperature (TMAX) – Temperatura máxima diaria media en ºC en el periodo
- Minimum temperature (TMIN) – Temperatura mínima diaria media en ºC en el periodo

4.2.5.Índices climáticos

Existen diferentes índices climáticos que miden la fluctuación de fenómenos climáticos concretos. Consultada la bibliografía (Malone et al., 2009) se ha determinado que existen ciertos índices climáticos que guardan una relación con las producciones anuales de maíz en Iowa, por lo que se han considerado como una potencial fuente de información para el modelo.

En la página web del NOAA Physical Sciences Laboratory (PSL)[30] se puede acceder a distintos índices climáticos. A continuación, se describe los índices climáticos utilizados en este trabajo:

- NAO – North Atlantic Oscillation: es un fenómeno climático en el norte del océano Atlántico, de fluctuaciones en la diferencia de presión atmosférica entre la baja islandesa y la alta de Azores o anticiclón de las Azores. Moviéndose de este a oeste entre la baja de Islandia y la alta de Azores, va controlando la fuerza y dirección de los vientos del oeste y las formaciones tormentosas a través del Atlántico Norte.
- ONI – Oceanic Niño Index: es una medida de la condición de El Niño-Oscilación del Sur (ENOS) y sus fases cálida (El Niño) y fría (La Niña) en el Pacífico ecuatorial central. Es el promedio móvil de tres meses de las anomalías de la temperatura superficial del mar estimadas a partir del producto ERSST.v5 SST en la región Niño 3.4 (5°N-5°S, 120°-170°W), basado en periodos base de 30 años y que se actualizan cada 5 años.
- QBO – Quasi-Biennial Oscillation: es una oscilacióncuasiperiódica de vientos ecuatoriales zonales, en la estratósfera tropical, con un periodo medio entre 28 a 29 meses. Los regímenes de vientos alternos se desarrollan en la parte superior de la estratosfera inferior y se propagan hacia abajo, cerca de 1 km por mes hasta que se disipan en la tropopausa tropical.
- SOI – Southern Oscillation Index: es un índice que mide la Oscilación del Sur al correlacionar valores de presión atmosférica obtenidos en el

[30] PSL

Pacífico occidental con los del Pacífico central. Estos valores se han asociado a los fenómenos climáticos de El Niño y La Niña.

4.2.6. Resumen y caracterización de las fuentes de información

Una vez descritas y caracterizadas todas las fuentes de información utilizadas en este trabajo, se elabora un resumen (Tabla 1) de las principales características de cada una de ellas que después servirán para entender los trabajos a realizar en el tratamiento de datos.

Tabla 1: Resumen de las características de las fuentes de información

Variables	Origen	Periodo de datos	Frecuencia de datos	Granularidad espacial
Productividad (bu/acre)	USDA	2000-2019	Anual	Condado
Fronteras	Iowa GeoData	Fijo	Fijo	Condado
Índices Satelitales	LP DAAC	02/2000 – 12/2019	Bimensual	Condado
Variables climáticas	NOAA	2000-2019	Mensual	Varias estaciones por condado
Índices Climático	PSL	2000-2019	Mensual	Mundial

4.3. Tratamiento de los datos

Como se ha visto en el punto 4.2.6 de este trabajo, cada fuente de información presenta unas características diferentes entre sí, por lo que se requiere realizar un tratamiento de los datos originales para adaptarlas a un contexto que permita estandarizar sus características.

Dado que varias de las fuentes de datos tienen una agregación mensual, se ha determinado que esta será la granularidad temporal del resumen de las variables utilizadas en este trabajo, salvo la productividad por condado que únicamente tiene un valor único anual.

A continuación, se exponen las transformaciones que hay que realizar a los datos originales extraídos de cada fuente de datos para adaptarlos a unas características comunes.

4.3.1. Producción de maíz

Los datos de producción extraídos del USDA contienen el valor por año y condado. Del conjunto de datos solo se necesita conservar tres columnas: el ID por el que se identificará el condado, la productividad medida en bu/acre y el año al que se corresponde.

4.3.2. Índices Satelitales

Como se ha mencionado, la disponibilidad de este conjunto de datos va desde 02/2000 a 12/2019, cuando el resto de variables incluyen datos para el mes 01/2000.

Dado que el mes de enero no se considera muy relevante en lo que respecta a al desarrollo del cultivo de maíz en Iowa, y dado que en esas fechas el cultivo no se encuentra presente en el campo, se ha considerado oportuno rellenar ese espacio asignándole el valor medio del índice para ese mes del resto de la serie temporal. De este modo el valor de los índices NDVI y EVI para 01/2000 será igual a la media de todos de los valores del mes de enero de periodo 2001-2019.

Además, como se expone en la caracterización de esta fuente de información, se dispone de dos registros del valor del índice por cada mes. Para realizar un resumen mensual de los datos, se realiza la media entre los valores para cada mes y año.

4.3.3. Estaciones meteorológicas

En el momento de acceder a los datos meteorológicos se aprecia que para cada condado del estado de Iowa existe un número diferente de estaciones disponibles sobre las que se pueden obtener datos. En ocasiones existe solamente una estación y en otras ocasiones existen más de quince. No todas ellas tienen disponibilidad de datos en el periodo de interés, y muchas de ellas, pese a indicar que contienen datos para todo el periodo, tras analizar al detalle alguna de ellas, se observa que contienen espacio en blanco para algunos periodos. Además, a priori y sin realizar

un estudio específico, frente a varias estaciones ubicadas en el condado, no se sabe cuál de ellas representa mejor el clima al que está expuesto el cultivo de maíz de forma global en todo el condado.

Por todo esto, se ha decidido descargar la información de las estaciones que contienen información para el periodo deseado para más tarde realizar una media ponderada entre todas las estaciones que existan para cada condado.

De este modo, se consigue recopilar y representar la variabilidad de clima que pueda existir en el condado. Además, como se está realizando el promedio de cada variable entre todas las estaciones, en el caso una de ellas no disponga de información en un espacio concreto, al realizar el ponderado con el resto de estaciones, se consigue que existan el menor número es espacios vacíos en el conjunto de datos.

4.3.4. Índices Climáticos

En el caso de los índices climáticos no se tiene que realizar ninguna acción ya que se descargan en el formato adecuado.

4.4. Estructura jerárquica de los datos

A la hora de manejar los datos es importante manejar una estructura jerárquica de niveles que vaya desde los archivos originales hasta un archivo que contenga las información ya tratada y estructurada. Además, se debe realizar una correcta identificación de los archivos para que a la hora de realizar cualquier operación con ellos no se pierda la trazabilidad. El código FIPS que identifica a cada condado será el identificador a la hora de referenciarlo.

Es por esto que se ha diseñado una estructura y se ha establecido una nomenclatura de los datos en sus diferentes niveles. A continuación, se realiza una explicación del contenido de cada nivel y su nomenclatura:

- Nivel 0: En este nivel se disponen los archivos CSV que contienen los datos obtenidos de cada fuente de información. Se determina las siguientes abreviaciones para cada fuente de información: WEA para datos climáticos, SAT para datos satelitales, YIE para datos de producción e IND para los índices climáticos.
- Nivel 1: En este nivel se realiza una separación de los datos a nivel de condado por lo que se tiene un archivo CSV por cada condado y fuente de información (IND es igual para todos los condados por lo que no se generan los archivos para cada condado). La nomenclatura es para el condado 1 sería: SAT_1, WEA_1 y YIE_1.
- Nivel 2: En este nivel se realiza una agregación de las fuentes de información a nivel de condado, por lo que se tiene un número de archivos CSV equivalente al número de condado. Cada archivo será nombrado mediante su identificador.
- Nivel 3: En este nivel se agrega la información de todos los condados en un único archivo CSV que se llamará SDS, o superdataset.

Así, la estructura del SDS, que está formado por filas y columnas tiene las dimensiones que se comentan a continuación. En cada fila el SDS tiene una observación de productividad por año y condado, lo que hace un total de 1944 filas. En cada columna se tiene una observación de cada variable y mes del año. Esto hace un total de 714 columnas.

Todas las operaciones realizadas a partir de los datos originales hasta llegar a la constitución del SDS han sido llevadas a cabo mediante el lenguaje de

programación R[31] en el entorno de trabajo de RStudio[32]. En el Anexo 5 puede verse el código utilizado.

4.5. Tratamientos de limpieza y reducción dimensional del dataset

Una vez que el dataset está completo hay que realizar una serie de trabajos para mejorar su calidad y dejarlo preparado para el inicio de los trabajos de modelado matemático. Estas operaciones de limpieza del conjunto de datos consisten en: eliminar columnas con valores únicos, eliminar columnas con más de un 10% de valores ausentes e imputar valores a los espacios sin valor de las columnas restantes. El código Python utilizado para los tratamientos de limpieza y reducción dimensional del dataset puede verse en el Anexo 2.

4.5.1. Eliminación de columnas con valores únicos

Un atributo con valores únicos nunca aportará información de valor para un proceso de aprendizaje automático ya que este atributo tiene una variación cero. Así, incluir columnas con valores iguales no aporta nada positivo y solo hace que se incrementen los tiempos de cómputo. Es por esto que se recomienda su eliminación antes de iniciar los trabajos de modelado matemático.

4.5.2. Eliminación de columnas con más de 10% de valores ausentes

Se determina que los atributos que contengan un número mayor de celdas vacías equivalente al 10% del total de observaciones serán eliminadas ya que al no contener información para un número demasiado alto de observaciones se considera

[31] Lenguaje de programación R
[32] RStudio

que no se pueden imputar valores de una forma correcta y que el aprendizaje sobre este atributo no sería correcto.

4.5.3. Imputación de valores ausentes

Como se ha mencionado antes, solamente se mantienen atributos con menos del 10% de valores ausentes, por lo que existen valores ausentes en el dataset. Para la mayoría de operaciones que se realizarán más adelante no es recomendable que existan valores ausentes dentro del dataset, por lo que hay que imputar valores a estos espacios vacíos.

Existen muchos métodos para la imputación de valores ausentes, pero siguiendo con las premisas de este trabajo de minimizar el costo de las operaciones a llevar a cabo, se ha optado por asignar a cada uno de estos valores ausentes el valor medio de todos los registros en la columna.

Una vez que se finaliza con las operaciones de limpieza, se consigue un dataset que contiene el mismo número de observaciones, 1944, pero que solo conserva 287 columnas.

Esta reducción tan importante se debe a que en el conjunto de datos climáticos existen algunas variables que solo son medidas en algunas estaciones, lo que hace que cuando se compara frente al resto de estaciones, muchas de esas variables no puedan ser utilizadas. Además, algunas otras variables contienen demasiados espacios vacíos dentro de sí mismo, y se considera oportuno eliminarlas. En total se han eliminado 427 columnas.

Capítulo 5
Modelado predictivo y experimentación

En este capítulo se expone como funciona el proceso de modelado matemático para el desarrollo de modelos predictivos basados en Machine Learning. Se hace especial atención en las técnicas de selección de atributos y validación de los modelos. Por otro lado, se expone el procedimiento de trabajo a seguir para dar cumplimiento a los objetivos de este trabajo.

5.1. Etapas iterativas del proceso de modelado matemático

El modelado predictivo consiste en establecer las relaciones entre un conjunto de datos y otro para tratar que estas se aproximen lo máximo posible a la realidad. A partir del conjunto de datos disponibles preparado al finalizar los trabajos descritos en el anterior capítulo, se han de desarrollar una serie de operaciones en ciclo hasta alcanzar un modelo matemático que presente unos niveles de error satisfactorios para el trabajo.

El proceso de modelado matemático no es lineal, si no que consiste en un conjunto de operaciones iterativas entre sí que forman un bucle sin fin. El bucle se detiene cuando la inversión de más horas de trabajos destinadas a la optimización del modelado no produce una reducción sustancial de los niveles de error ni una mejora de la calidad del resultado. En ese momento, hay que detener el bucle de trabajo y conservar el mejor modelo resultante hasta dicha iteración del bucle para utilizarlo como herramienta predictiva.

A continuación, se representa el buble iterativo del proceso de modelado para el desarrollo de modelos matemáticos predictivos mediante Machine Learning:

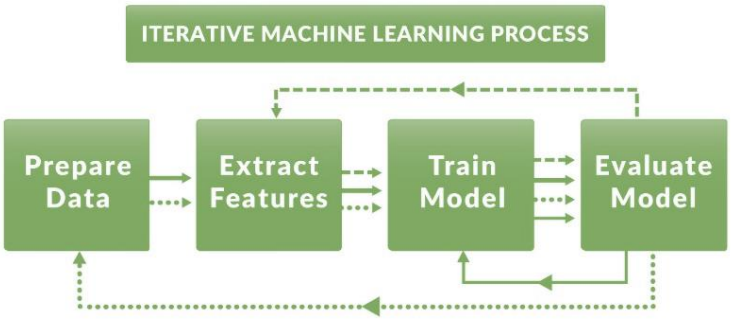


Figura 17: Bucle iterativo de modelado matemático

Como se puede ver en la Figura 17, existen cuatro etapas en el proceso, que tienen un orden de desarrollo, pero los que se vuelve después de cada siguiente etapa de forma iterativa para conseguir el desarrollo del mejor modelo.

5.1.1. Preparación de los datos

En esta etapa se prepara el conjunto de datos que se va a utilizar como base para el resto de etapas. A partir de un conjunto de datos dado, se realizan las operaciones necesarias para adaptarlos a las necesidades y objetivos del proceso de modelado matemático. Pueden realizarse operaciones de selección de una parte del conjunto de datos o transformación del tipo de variable de algunas columnas.

5.1.2. Selección de atributos

La selección de atributos es el proceso por el que se seleccionan un subconjunto de ellos para su uso en la construcción de modelos. Las técnicas de selección de características se utilizan por varias razones:

- simplificación de modelos para mejorar su interpretación
- reducir los tiempos de entrenamiento
- evitar el sobreajuste de los modelos y mejorar la generalización

La premisa central de la selección de atributos es que los datos que contienen información redundante o irrelevante sean excluidos del proceso de modelo. Redundante e irrelevante son dos conceptos distintos, ya que un atributo relevante puede ser redundante en presencia de otro atributo relevante con el que esté fuertemente correlacionado.

Para la selección de atributos primero se realizará una reducción de la dimensionalidad del dataset mediante la eliminación de atributos que tengan una fuerte correlación entre ellos. Se determina una correlación de 0,9 o superior para identificar y eliminar uno de los dos atributos que mantengan dicha correlación elevada.

La selección de atributos se realizará mediante el método LASSO descrito en el capítulo 2 de este trabajo.

5.1.3. Entrenamiento del modelo

El entrenamiento de los modelos se realizará mediante los algoritmos descritos en el Capítulo 2 de este trabajo.

Para realizar el modelado es necesario dividir el conjunto de datos en 4 subconjntos. Por un lado, hay que dividir el conjunto de datos para separar los datos a predecir (y) y el conjunto de datos que aportan conocimiento en el aprendizaje (X). Por otro lado, hay que separar ambos conjuntos en una parte de entrenamiento (Train) y otra de evaluación (Test). De esto modo se tienen los siguientes conjuntos de datos para realizar el entrenamiento de todos los modelos: X_train, X_test, y_train e y_test. A continuación, se realiza una representación gráfica (Figura 18) de la división del dataset para realizar el entrenamiento de los modelos.

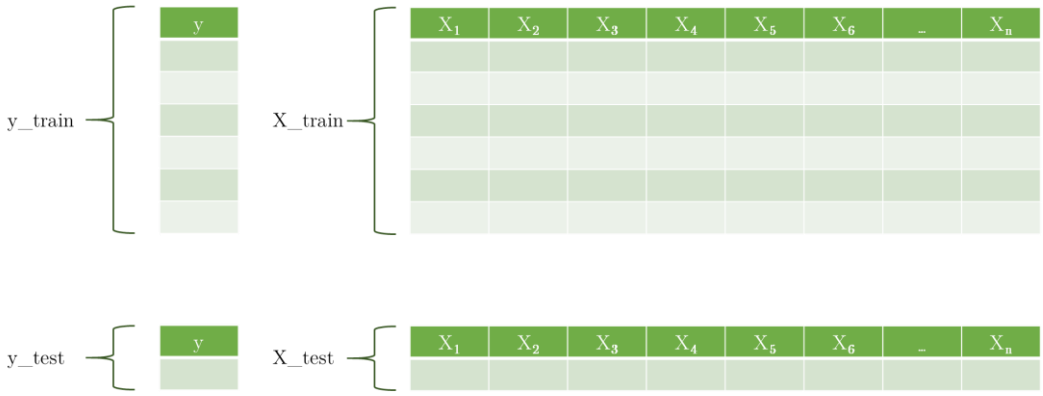


Figura 18: Representación gráfica de la división del dataset para el entrenamiento y validación del modelo.

No cabe duda que la división entre X e y es clara, al mantener en y la columna a predecir, y en X el resto. La división entre train y test no es baladí y depende de la metodología de validación utilizada para comprobar la bondad de los modelos.

En este trabajo, la validación de los modelos ajustados mediante cada algoritmo se realizará mediante la técnica de validación cruzada "Leave one year out". Esta técnica consiste en realizar una división entre train y test basada en el año en que se han realizado las observaciones. De una forma secuencial, se realiza el proceso de división, entrenamiento y cálculo del error en un número de veces igual al número de años que existan en el dataset.

Cada uno de los algoritmos utilizados tiene diferentes hiper-parámetros que pueden ser modificados para optimizar el proceso de modelado. Es por esto, que antes del modelado con cada algoritmo, se realizará la determinación de los hiper-parámetros que optimizan el resultado mediante el algoritmo GridSearchCV.

5.1.4. Evaluación del modelo

Debido a la forma de validar los modelos explicada en el punto anterior, para cada secuencia del proceso de división basada en el año de las observaciones, se dispondrá de unos parámetros del error cometido en dicho año, por lo que la evaluación completa de la bondad del modelo será un promedio de los valores de cada parámetro en todos los años.

Las métricas a analizar para determinar cuál es el modelo que presenta la mayor capacidad de predicción serán el MAE, MAPE y R^2. Todas ellas han sido descritas en el capítulo 2 de este trabajo.

Una vez seleccionado el modelo con mejor comportamiento, se realizará un test final de validación con el 2019 como año a predecir, que servirá para comprobar el nivel de acierto con un conjunto de datos que no ha formado parte en ningún momento del proceso de modelado y permitirá evaluar su capacidad real de predicción.

5.2. Caso de uso 1: Modelos predictivos de maiz

En este caso de uso se aborda el objetivo 1, que como ya se definió en el Capítulo 1 de este trabajo, pretende generar un caso de uso de modelos predictivos de producción de maíz. Para ello, a continuación, se describen las tareas específicas a realizar en cada una de las etapas del proceso de modelado matemático.

5.2.1. Preparación de los datos

En esta etapa se prepara el conjunto de datos que se va a utilizar como base para el resto de trabajos. Como se menciona en el Capítulo 2, el cultivo de maíz en Iowa está presente desde abril, cuando se produce la siembra, y hasta el mes de noviembre, cuando se produce la cosecha. Es por esto, que solo se eliminarán los atributos correspondientes a los meses de enero, febrero, marzo y diciembre.

Las observaciones correspondientes al año 2019 se separarán y se mantendrán al margen de todos los procesos de modelado matemático. Esto permitirá disponer de un conjunto de datos para validación final del modelo seleccionado con la seguridad de que el modelo tiene la capacidad de generalizar su aprendizaje para un conjunto de datos nunca antes visto y que es robusto para ser utilizado en operación real.

5.2.2. Selección de atributos

A partir del conjunto de datos que contiene las observaciones correspondientes a los años 2000-2018 se realizará la selección de atributos mediante la técnica LASSO descrita en el Capítulo 2 de este trabajo.

5.2.3. Entrenamiento del modelo

Para realizar el entrenamiento del modelo se utilizarán los algoritmos propuestos y descritos en el Capítulo 2 de este trabajo. El modelo DummyRegressor servirá como referencia del objetivo mínimo a batir.

Para determinar la configuración óptima de los hiper-parámetros de cada modelo se utilizará el algoritmo GridSearchCV. De este modo, realizando una calibración previa de los algoritmos se asegura que se están obteniendo los mejores resultados que pueden ofrecer y, por tanto, la comparación del ajuste entre unos y otros será en las mejores condiciones de cada uno.

5.2.4. Evaluación del modelo

La evaluación de los modelos se realizará mediante los parámetros descritos en el Capítulo 2 de este trabajo. El modelo seleccionado será el que presente los mejores indicadores del error y tenga el funcionamiento más adecuado para las necesidades que se tienen.

Una vez seleccionado el algoritmo y su configuración que mejores niveles de error presenta, se realizará una validación final con el conjunto de datos correspondiente al año 2019. Esta prueba consistirá en modelado utilizando todo el conjunto de datos (años 2000-2018) y realizar la predicción de las productividades del año 2019 para todos los condados del estado de Iowa. Dado que conjunto de datos correspondiente al año 2019 no ha participado en ninguna etapa del proceso de modelado matemático, esta validación final permitirá determinar la capacidad de generalización del aprendizaje del modelo frente a un conjunto de datos nuevos.

5.3. Caso de uso 2: Número mínimo de años

En este caso de uso se aborda el objetivo 2, que como ya se definió en el Capítulo 1 de este trabajo, pretende determinar el número de años mínimo para un buen ajuste de los modelos. Para ello, se partirá del algoritmo ajustado validado en el caso de uso 1. A continuación, se describen las tareas específicas a realizar en cada una de las etapas del proceso de modelado matemático, para este caso.

5.3.1. Preparación de los datos

Para generar los diferentes conjuntos de datos de entramiento que son necesarios para dar solución a este objetivo, se irán eliminando las observaciones correspondientes de año en año, empezando por el año 2000 y finalizando en el año 2017. Esto generará 18 conjuntos de datos de entrenamiento, uno por cada año eliminado.

5.3.2. Selección de atributos

No se realizará un trabajo de selección de atributos y se mantendrán los mismos que se consideraron oportunos en el proceso de modelado para alcanzar el objetivo 1 de este trabajo.

5.3.3. Entrenamiento del modelo

El algoritmo utilizado y su configuración serán las mismas que las alcanzadas en el objetivo 1 de este trabajo.

5.3.4. Evaluación del modelo

La evaluación de los modelos se realizará mediante los parámetros descritos en el Capítulo 2 de este trabajo. Se realizará una comparativa de la variación de los parámetros conforme se avanza en el proceso sistemático de eliminación de datos correspondientes a cada año.

5.4. Caso de uso 3: Mes para iniciar las predicciones

En este caso de uso se aborda el objetivo 3, que como ya se definió en el Capítulo 1 de este trabajo, pretende determinar cuánto tiempo antes de la cosecha es posible predecir la producción con un nivel de certeza adecuado. Para ello, se partirá del algoritmo ajustado validado en caso de uso 1. A continuación, se describen las tareas específicas a realizar en cada una de las etapas del proceso de modelado matemático.

5.4.1. Preparación de los datos

Para generar el conjunto de datos de entrenamiento, se irán eliminando las observaciones correspondientes a cada mes, empezando por el mes de noviembre y terminando en el mes de mayo. Esto generará 8 conjuntos de datos de entrenamiento, uno por cada mes eliminado.

5.4.2. Selección de atributos

Dado que se está realizando una eliminación secuencial de atributos correspondientes a determinados meses, existe la necesidad de realizar una nueva selección de atributos en cada iteración del proceso.

El método de selección de atributos será LASSO, descrito en el capítulo 2 de este trabajo.

5.4.3. Entrenamiento del modelo

El algoritmo utilizado y su configuración serán las mismas que las alcanzadas en el objetivo 1 de este trabajo.

5.4.4. Evaluación del modelo

La evaluación de los modelos se realizará mediante los parámetros descritos en el Capítulo 2 de este trabajo. Se realizará una comparativa de la variación de

los parámetros conforme se avanza en el proceso sistemático de eliminación de atributos correspondientes a cada mes.

Capítulo 6
Resultados y discusión

De cada uno de las etapas que conforman el proceso de proceso de modelado se pueden extraer resultados y conclusiones. Así, para cada objetivo de este trabajo se comentará lo más relevante de lo descubierto durante el proceso de modelado.

6.1. Caso de uso 1

Los puntos más importantes a comentar al finalizar proceso de trabajo para alcanzar el objetivo 1 son: la selección de atributos, cuales es el mejor algoritmo y su configuración, y cuál es el resultado de la validación para el año 2019.

6.1.1. Selección de atributos

El conjunto de datos está formado por 193 columnas, de las que una es el ID de cada condado y otra la productividad a predecir, por lo que se tienen 191 atributos que pueden aportar información al modelo. Una vez se realiza la eliminación de atributos por correlación entre ellos, el conjunto de datos cuenta con 141 atributos candidatos a formar parte del modelado.

Al aplicar el método Lasso de selección de atributos, este arroja los atributos que se pueden ver en la Figura 19 como los relevantes dentro del conjunto de datos.

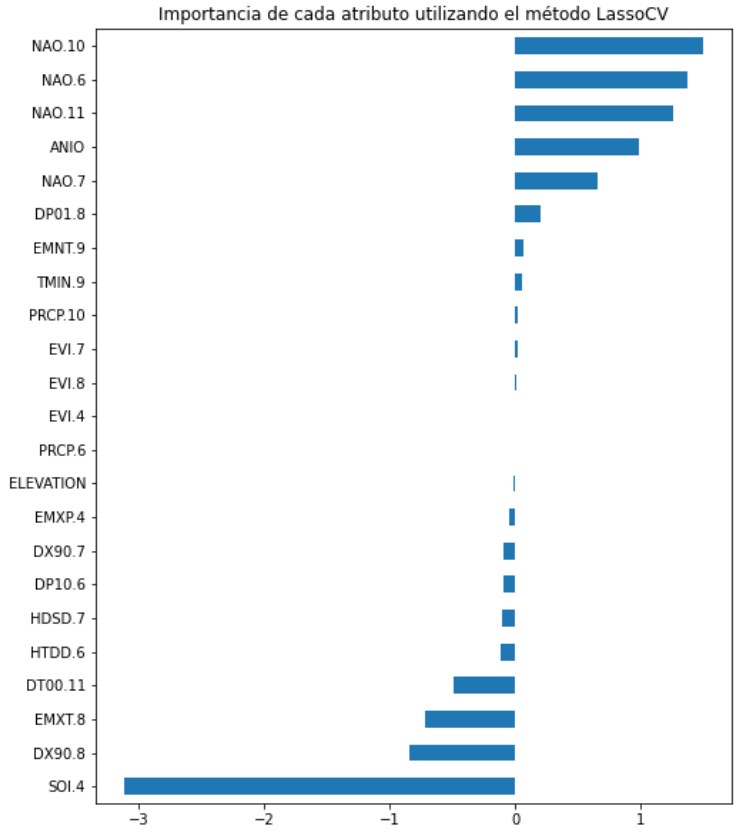


Figura 19: Importancia de cada atributo seleccionado por el método LassoCV

Los atributos que poseen un coeficiente de importancia positivo implican que un incremento en el valor de esos atributos produce un incremento de la producción de maíz. En cambio, los atributos que poseen un coeficiente de importancia negativo implican que un incremento de esos atributos produce una disminución de la producción de maíz.

Una vez seleccionados los atributos más relevantes en el proceso de selección de atributos se realiza el cálculo de la correlación de cada uno de estos atributos con la productividad de maíz. De este modo permitirá determinar con cuales de ellos existe una correlación más fuerte.

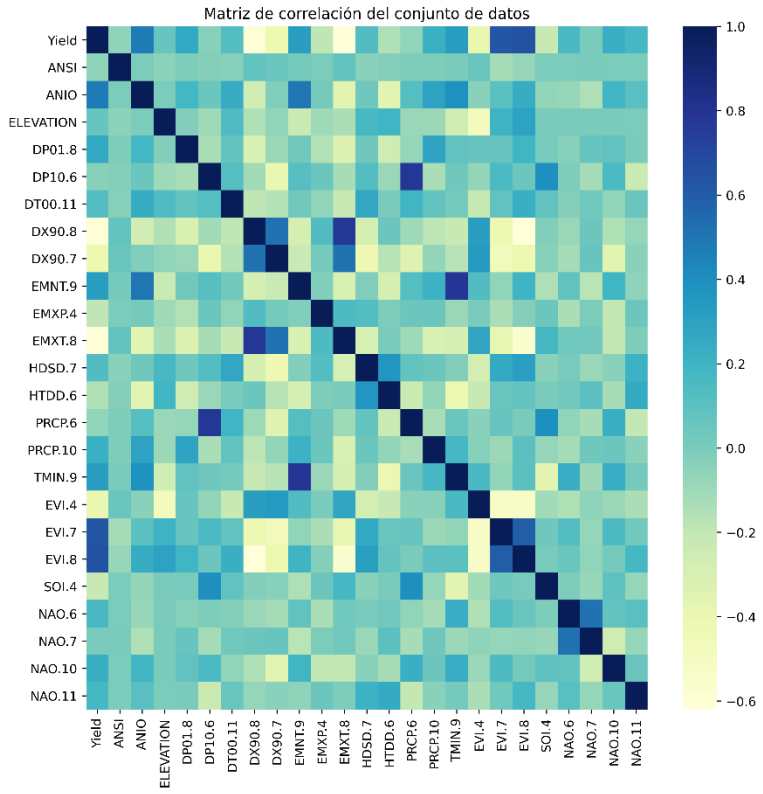


Figura 20: Matriz de correlación del conjunto de datos

Como se puede ver en la Figura 20, las correlaciones más importantes con la producción son EVI.8 (0,64), EVI.7 (0,63), ANIO (0,47), EMNT.9 (0,32) y TMIN.9 (0,31). Claramente, un mayor valor del índice EVI.8 y EVI.7 quiere decir que existe mayor masa vegetal por lo que se espera la producción sea más alta. Respecto al año, a priori no debería existir una correlación entre la producción ya que el año en que se produce no debería tener una relación directa con la producción de maíz. Sin embargo, este hecho se explica debido a las mejoras de las herramientas productivas año tras año. Conforme avanza el tiempo los productores cuentan con mejores semillas, mejores máquinas, más conocimiento y experiencia, más herramientas de gestión de cultivos, etc. Es por esto, que año tras año la productividad (bu/acre) tiende a mejorar y se genera esta correlación. Parece ser que las temperaturas mínimas del mes de septiembre (EMNT.9 y TMIN.9) tienen una especial importancia en el cultivo de maíz, ya que coincide con el estado fenológico de maduración. Analizándolo conjúntame con la importancia del atributo, que tiene

un coeficiente positivo, se puede determinar que, si las temperaturas mínimas del mes de septiembre son elevadas, se produce un incremento en la producción.

6.1.2.Selección del algoritmo y su configuración

A continuación, se hace una exploración del comportamiento de cada algoritmo realizando la tarea de modelado asignada.

Dummy Regresor - DR

Este algoritmo sirve como punto de partida, ya que simplemente su valor de predicción para cada observación es el valor medio de los valores de entrenamiento. Es por esto que se le llama "Dummy", ya que es la predicción más simple que se puede realizar. El resto de modelos tienen que mejorar los resultados de este para considerar que han conseguido realizar un correcto aprendizaje.

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018, este algoritmo presenta los siguientes resultados:

- MAE promedio: 21,08 bu/acre
- MAPE promedio: 12,76%
- R^2: -1,73

Linear Regression - LR

La búsqueda de los hiper-parámetros óptimos indica que esta debe ser la siguiente: Fit_intercept: True, Normalize: True, Copy_X: True. Puede verse en el Anexo 3 el código Python utilizado para la búsqueda automática de la mejor combinación de hiper-parámetros en este modelo.

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018 (puede verse en el Anexo 4 el código Python utilizado), este algoritmo presenta los siguientes resultados:

- MAE promedio: 10,83 bu/acre

- MAPE promedio: 6,60%
- R^2: 0,26

Support Vector Regression - SVR

La búsqueda de los hiper-parámetros óptimos indica que esta debe ser la siguiente: C=10, cache_size=200, coef0=0.0, degree=9, epsilon=0.1, gamma='scale', kernel='poly', max_iter=-1, shrinking=True, tol=0.0001, verbose=False.

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018, este algoritmo presenta los siguientes resultados:

- MAE promedio: 14,94 bu/acre
- MAPE promedio: 8,94%
- R^2: -0,43

Multi-layer Perceptron Regresor - MLPR

La búsqueda de los hiper-parámetros óptimos indica que esta debe ser la siguiente: activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999, early_stopping=False, epsilon=1e-08, hidden_layer_sizes=(200,), learning_rate='constant', learning_rate_init=0.001, max_fun=15000, max_iter=200, momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5, random_state=1, shuffle=True, solver='adam', tol=0.0001, validation_fraction=0.1, verbose=False, warm_start=False.

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018, este algoritmo presenta los siguientes resultados:

- MAE promedio: 12,51 bu/acre
- MAPE promedio: 7,63%

- R^2: -0,18

Gaussian Process Regressor - GPR

La búsqueda de los hiper-parámetros óptimos indica que esta debe ser la siguiente: alpha=1e-05, copy_X_train='True', kernel=DotProduct(sigma_0=1) + WhiteKernel(noise_level=1), n_restarts_optimizer=5, normalize_y='True', optimizer='fmin_l_bfgs_b', random_state=1

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018, este algoritmo presenta los siguientes resultados:

- MAE promedio: 10,73 bu/acre
- MAPE promedio: 6,56%
- R^2: 0,28

Random Forest Regressor - RFR

La búsqueda de los hiper-parámetros óptimos indica que esta debe ser la siguiente: bootstrap=True, ccp_alpha=0.0, criterion='mse', max_depth=None, max_features='sqrt', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.01, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=200, n_jobs=None, oob_score=False, random_state=1, verbose=0, warm_start=False

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018, este algoritmo presenta los siguientes resultados:

- MAE promedio: 12,46 bu/acre
- MAPE promedio: 7,58%
- R^2: -0,08

Extremely Randomized Tree Regressor - ETR

La búsqueda de los hiper-parámetros óptimos indica que esta debe ser la siguiente: ccp_alpha=0.0, criterion='mse', max_depth=None, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=1, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, random_state=1, splitter='best'

Así, realizando el proceso de validación cruzada mediante "leave one year out" de la serie 2000-2018, este algoritmo presenta los siguientes resultados:

- MAE promedio: 13,18 bu/acre
- MAPE promedio: 8,06%
- R^2: -0,29

Resumen del comportamiento de los modelos

Tabla 2: Resumen de los indicadores de error de los modelos ajustados

	DR	LR	SVR	MLPR	GPR	RFR	ETR
MAE (bu/acre)	21,08	10,83	14,94	12,51	10,73	12,47	13,18
MAPE (%)	12,77	6,61	8,95	7,63	6,56	7,59	8,06
R^2	-1,74	0,26	-0,44	-0,19	0,29	-0,09	-0,29
Var. Explicada (%)	0,00	44,79	37,72	35,06	45,87	42,66	21,84
Error máx (bu/acre)	59,52	39,95	48,4	44,63	39,24	42,23	45,92

DR: DummyRegresor
LR: Linear Regression
SVR: Support Vector Regression
MLPR: Multi-layer Perceptron Regresor
GPR: Gaussian Process Regressor
RFR: Random Forest Regressor
ETR: Extremely Randomized Tree Regressor
MAE: Error promedio
MAPE: Error promedio porcentual

Como se aprecia en la tabla anterior, todos los modelos mejoran los resultados respecto al DummyRegresor, por lo que, en mayor o menor medida, todos los modelos han conseguido completar la tarea de aprendizaje automático.

Con claridad, los modelos que mejor bondad presentan a la hora de realizar la labor de predicción de la productividad de maíz son LR y GPR. Sus números en los diferentes indicadores son muy similares, siendo los de GPR ligeramente mejores.

Sin embargo, el tiempo de computación requerido por el GPR es ostensivamente muy superior al requerido por LR. La ejecución del modelo GPR tuvo un tiempo de computación de 1.264 segundos, mientras que el de LR fue de tan solo 6 segundos. La ejecución del modelo GPR consume 210 veces más tiempo que el modelo LR.

Por otro lado, el modelo LR permite una mayor interpretabilidad de los resultados, ya que se puede consultar de una forma sencilla el peso que está asignando a cada atributo en el modelo.

Por las razones anteriormente expuestas, se determina que el modelo entrenado con el algoritmo LR es el más adecuado para cumplir con el objetivo de este trabajo.

6.1.3. Validación en el año 2019

Una vez seleccionado el algoritmo y su configuración que mejor se adapta a la solución de nuestro problema, se procede a validar su funcionamiento contra el año 2019. El conjunto de datos de 2019 no ha formado parte del proceso de modelado en ningún momento, ni en la selección de atributos ni en la validación del modelo.

De este modo, a partir del algoritmo LR y su configuración de hiperparámetros, se entrena el modelo con el conjunto de datos correspondiente a los años 2000-2018. Una vez entrenado el modelo, se le introducen el conjunto de variables explicativas para permitir predecir la productividad por condados en 2019,

que posteriormente se compararán con las reales. A continuación, puede verse una representación gráfica de la comparación entre los valores reales y predichos.

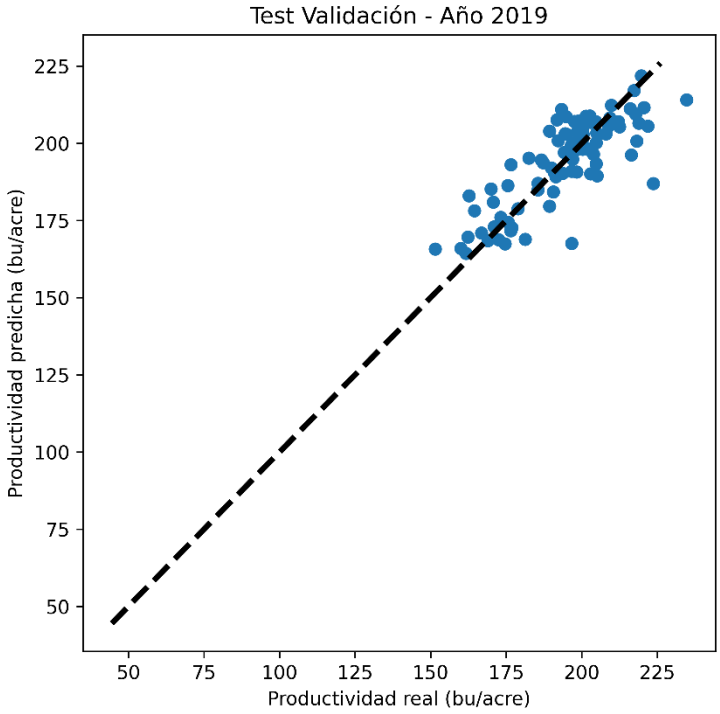


Figura 21: Representación gráfica de la validación en el año 2019

El modelo presenta los siguientes resultados para 2019:

- MAE: 7,54 bu/acre
- MAPE: 3,88%
- R2: 0,64
- Varianza explicada: 64,56%
- Error max: 36,77 bu/acre

Como se puede ver, el modelo ha cumplido satisfactoriamente la tarea de predicción de la productividad por condado en el año 2019 presentando unos valores de error muy adecuados para convertirse en una herramienta consistente para la toma de decisiones en el sector agroindustrial.

6.1.4. Análisis de los resultados

Una vez se ha seleccionado el algoritmo que mejores prestaciones ofrece para resolver la tarea asignada, se realiza un análisis de los resultados año a año del modelo para el periodo 2000-2018.

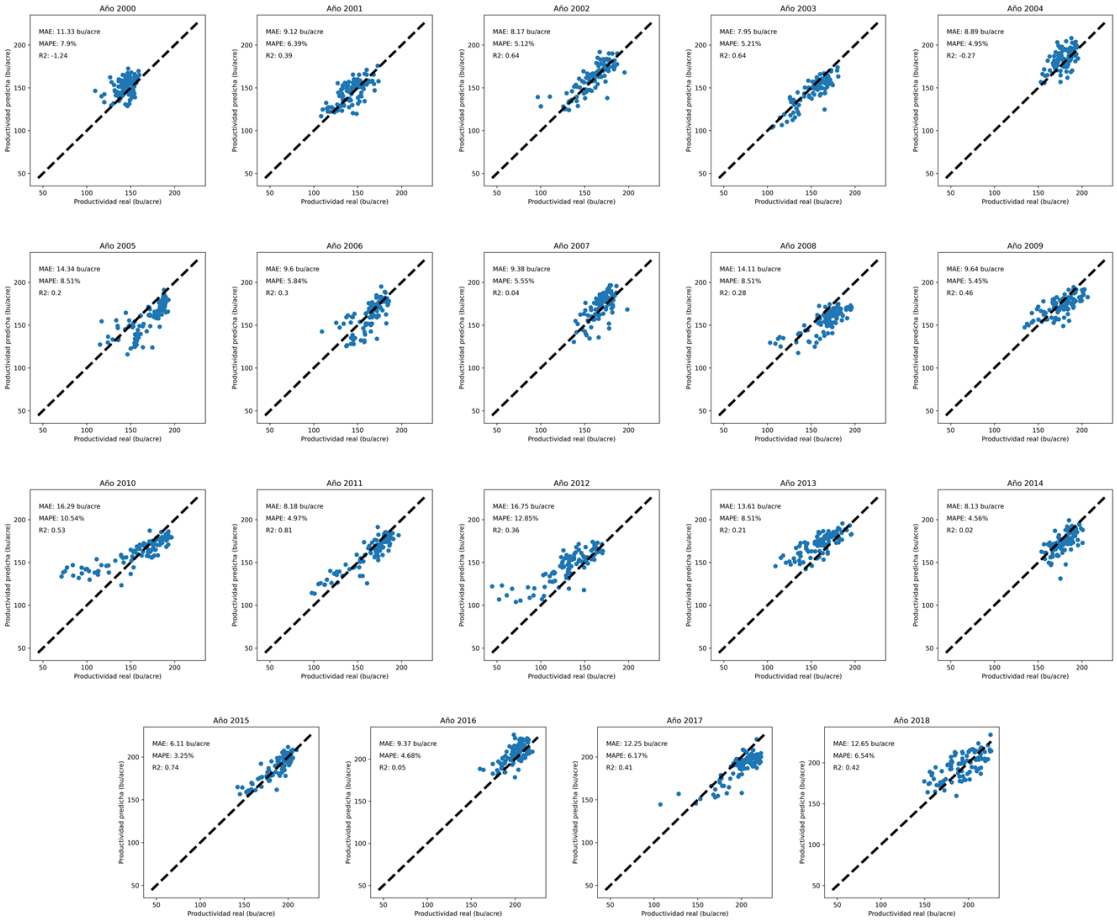


Figura 22: Análisis de los resultados del modelo para los años 2000-2018

Como puede verse en la Figura 22, el modelo tiene un buen comportamiento para los valores de productividad bien representados en el modelo, las productividades superiores aproximadamente a 100 bu/acre. Concretamente, en los años 2012 y 2010, el modelo no consigue tener un buen ajuste para las productividades anormalmente bajas y poco representadas en el histórico.

Si se analiza año por año mediante un diagrama de *box and whisker*, la dispersión de las productividades por condado pueden identificarse registros de productividad anormalmente bajos en dichos años.

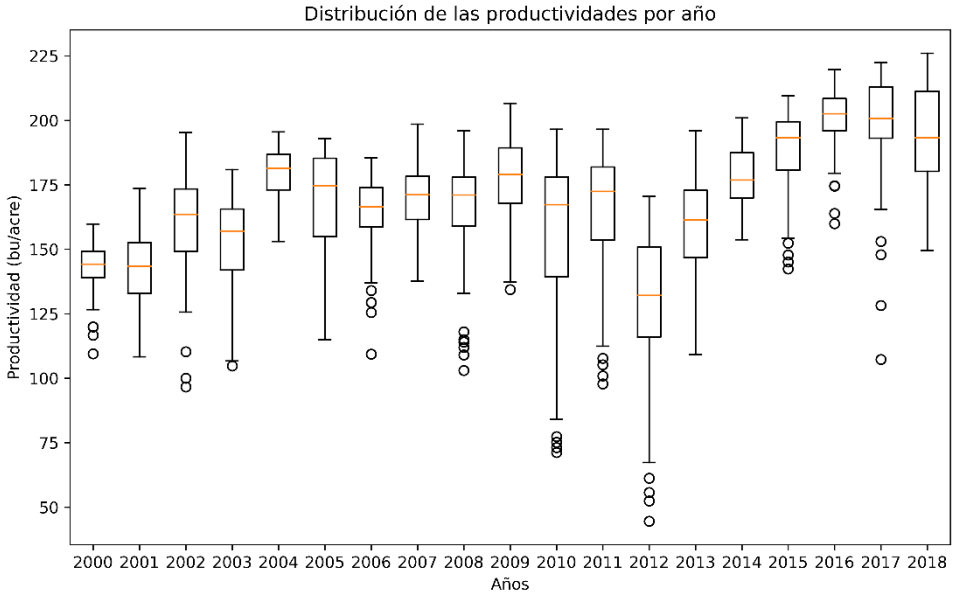


Figura 23: Diagrama box and whisker de las productividades de los condados para los años 2000-2018

Como se puede ver en la Figura 23, justo en los años 2010 y 2012 existen una serie de productividades en algunos condados que son anormalmente bajas. El modelo no es capaz de realizar unas correctas predicciones para esos valores ya que no se ven representados en otros años dentro del histórico con suficiente representatividad. Aquí se abre la línea de trabajo de eliminación de observaciones estadísticamente anómalas para ver si de ese modo se pueden mejorar los resultados del modelo.

Otra cosa que permite observar el modelo LR es el signo del valor de cada coeficiente que multiplica a cada una de las variables independientes del modelo. Es complicado dar un valor concreto a cada uno de los atributos, ya que por un lado para compararlos entre sí sería necesario normalizar el conjunto de datos, y por otro, al realizar una validación cruzada para el conjunto de años 2000-2018, se tienen diferentes coeficientes para cada atributo en función de cada año. Por esto,

se ha realizado el promedio del coeficiente de cada atributo para ver su signo, positivo o negativo, a lo largo de los distintos años. En la Figura 24 se presentan los atributos son su signo a lo largo de la campaña de producción del maíz.

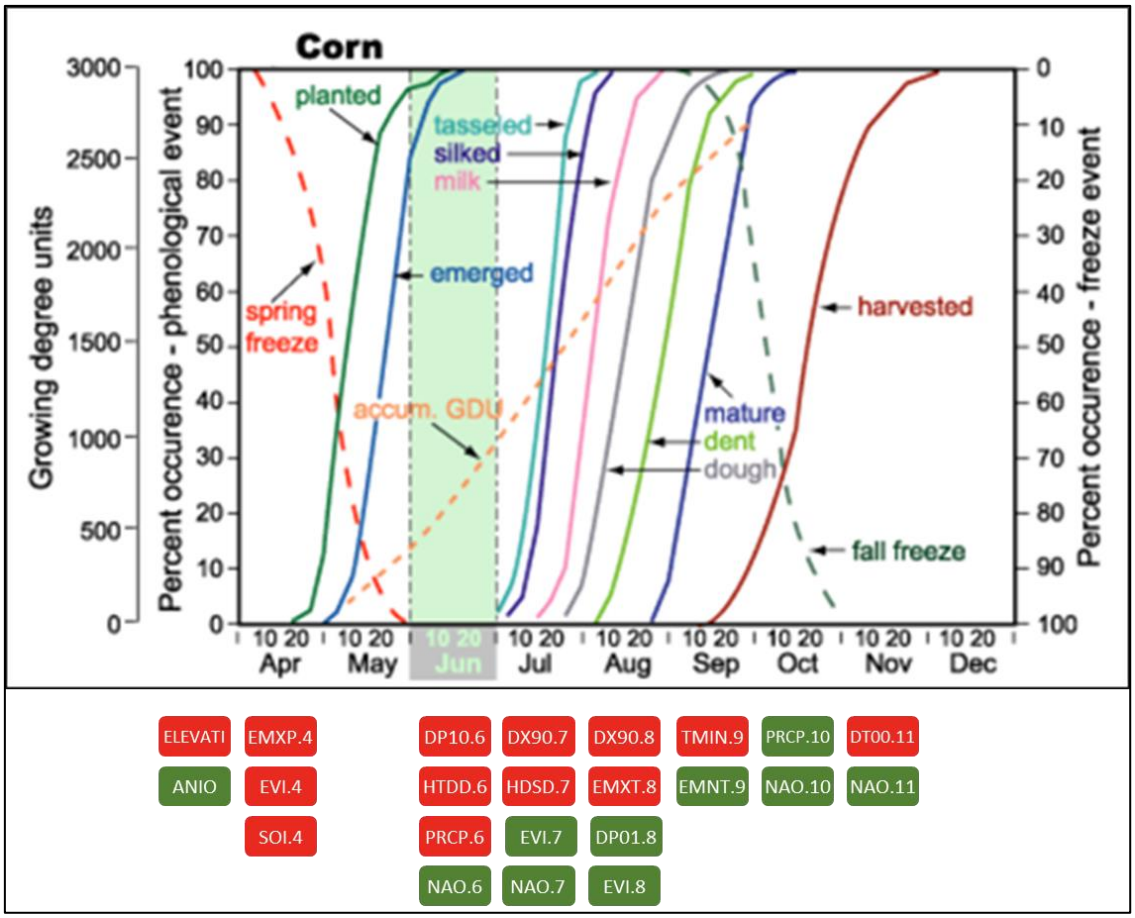


Figura 24: Atributos del modelo distribuidos a lo largo del ciclo de desarrollo fenológico del maíz.

Los atributos representados dentro de los cuadros en color rojo significan que tienen una relación negativa con la productividad, por lo que un incremento del valor de estos atributos supone una disminución del valor de la productividad final. Por el contrario, los atributos representados dentro de los cuadros en color verde significan que tienen una afección positiva con la productividad, por lo que cuanto más altos sean los valores de estos, mayor será la productividad final que se alcance.

Al realizar la validación del modelo con el conjunto de datos correspondiente al año 2019, se obtuvo un error promedio entre todos los condados de 3,88%, o lo que es igual a 7,54 bu/acre. La validación se hizo sobre 88 de los 99 condados que

conforman el estado de Iowa, ya que no se disponen de los datos para los condados restantes.

En la Figura 25 se muestra el rango del error cometido por el modelo en la validación en el año 2019:

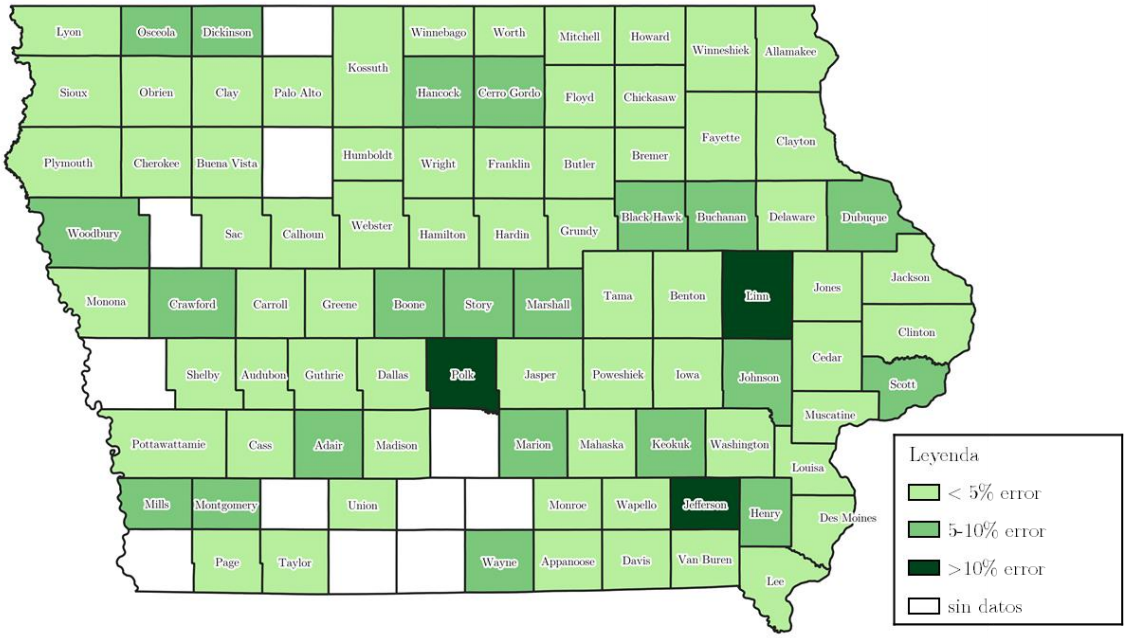


Figura 25: Rango del error cometido por el modelo en cada condado para el año 2019

Como se puede ver, únicamente en 3 condados del estado de Iowa la predicción del modelo tiene una desviación superior al 10%. En 21 estados la desviación está en un rango del 5-10%, y en los restantes 64 condados la desviación es menor al 5%.

En este punto se plantea conocer cuál ha sido el error en la predicción de la producción total en bushels en todo el estado de Iowa (88 condados considerados), ya que esta sería la información que más valor aportaría a un potencial comercializador de maíz, al que le puede interesar conocer la producción total en el estado, y no tanto la productividad en cada uno de los condados, ya que independientemente de donde se produzca el maíz, terminará el total de la producción en los mismos mercados de productos agrícolas.

Si se realiza la operación de multiplicar la productividad predicha por el número de hectáreas cosechadas en cada condado, y se realiza el sumatorio de todas estas producciones, se obtiene un total de 2.396.641.896 bushels, cuando la producción total del estado fue de 2.402.499.000 bushels. La diferencia es de 5.857.104, o lo que es lo mismo, un 0,24% de la producción total. Esto vuelve a confirmar el buen desempeño del modelo y su capacidad para predecir correctamente la producción de maíz en el estado de Iowa.

6.2. Caso de uso 2

A continuación, se puede ver una comparativa de los niveles de error que presenta la validación con el año 2019 conforme se van eliminando años, uno a uno, del conjunto de entrenamiento.

Tabla 3: Evolución de los indicadores del error en el caso de uso 2

Año Inicio Conjunto de datos	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
MAE (bu/acre)	7,54	8,36	9,1	9,35	9,13	8,36	7,84	8,37	8,25	54,53
MAPE (%)	3,88	4,3	4,68	4,81	4,7	4,3	4,03	4,31	4,25	28,06
R2	0,64	0,57	0,51	0,49	0,51	0,57	0,63	0,59	0,59	-9,89
Var. Explicada (%)	64,55	65,3	64,77	65,02	65	65,19	63,31	61,68	60,66	61,17
Error max (bu/acre)	36,77	41,83	43,62	44,08	43,91	42,08	35,58	33,97	34,69	77,94

Año Inicio Conjunto de datos	2010	2011	2012	2013	2014	2015	2016	2017	2018
MAE (bu/acre)	42,84	21,63	11,17	12,96	8,86	8,13	9,03	10,73	16,98
MAPE (%)	22,05	11,13	5,75	6,67	4,56	4,18	4,65	5,52	8,74
R2	-5,89	-0,96	0,35	0,16	0,54	0,61	0,51	0,29	-0,52
Var. Explicada (%)	59,58	60,51	61,05	58,84	59,36	61,32	51,22	31,26	-51,29
Error max (bu/acre)	70,02	49,04	35,71	39,7	32,9	34,55	36,63	51,54	52,92

Como se puede ver en la Tabla 3, el comportamiento del modelo no se ve muy afectado si se eliminan los años desde el 2000 al 2008. En los años 2009, 2010, 2011 se produce una perturbación muy importante de los indicadores del modelo. Entre los años 2012 y 2015 el resultado del modelo se mantiene bastante estable, para a partir de 2016 van empeorando progresivamente los resultados.

Es muy revelador el comportamiento de la varianza explicada, que va descendiendo conforme se eliminan años de entrenamiento, llegando más o menos constante hasta 2015. A partir de ese año, el modelo presenta un drástico empeoramiento de este parámetro.

Con este análisis se puede concluir que puede existir un periodo mínimo de 3 años de datos para que un modelo como el presentado tenga un buen comportamiento. A partir de ahí, hay que tener cuidado ya que no siempre más histórico de datos presenta mejores resultados, ya que se pueden estar incluyendo datos de años muy antiguos que no tienen mucha relación con la realidad productiva actual y pueden estar perjudicando el comportamiento del modelo cuando se aplica en la actualidad.

6.3. Caso de uso 3

De forma similar a lo realizado anteriormente, se van a ir eliminando atributos del modelo correspondientes a los diferentes meses del año para estimar desde que momento se pueden empezar a realizar predicciones de productividad con un nivel considerable de acierto.

A continuación, puede verse la evolución de los indicadores del modelo conforme se van eliminando los atributos correspondientes a cada mes. Se parte de la situación en que se disponen de todos los meses posibles (hasta noviembre incluido), eliminándose meses de forma secuencial hasta llegar a utilizar solamente los del mes de abril.

Tabla 4: Evolución de los indicadores del error en el caso de uso 3

Atributos desde mes	Nov	Oct	Sep	Ago	Jul	Jun	May	Abr
MAE (bu/acre)	7,54	10,14	8,66	11,53	9,53	13,81	16,34	16,34
MAPE (%)	3,88	5,22	4,46	5,93	4,9	7,11	8,41	8,41
R2	0,64	0,46	0,56	0,33	0,5	0,06	-0,32	-0,29
Var. Explicada (%)	64,55	63,92	61,01	59,49	53,16	43,67	36,95	35,46
Error max (bu/acre)	36,77	29,5	33,73	30,55	36,63	43,52	47,29	41,99

Como era de esperar, se puede ver en la Tabla 4, conforme los atributos utilizados perteneces a meses más lejanos a la fecha de recolección (noviembre), del momento de predicción, los indicadores del error del modelo tienden a empeorar.

Hasta el mes de julio, los valores del MAPE permanecen en un entorno del 5% o menor. A partir de junio, ya los valores del MAPE aumentan, por lo que se puede determinar qué finales de junio (ya que se tiene que disponer de los resúmenes mensuales del clima acontecido en junio) es el momento en que las predicciones tienen un nivel de error similar al que se consigue una vez finalizada la campaña de producción de maíz en noviembre.

6.3.1. Análisis de los resultados

El USDA, a través del NASS (National Agricultural Statistics Service) realiza estimaciones de producción para el cultivo de maíz en todo el país. A partir del 1 de agosto, y de forma mensual hasta el 1 de noviembre, este organismo emite 4 estimaciones de producción que pretenden aportar información sobre las expectativas de la campaña productiva a los diferentes actores del sector a nivel mundial, dado que EEUU es el primer productor mundial de maíz.

El NASS tradicionalmente ofrecía las estimaciones de productividad a nivel de estado, pero a partir del año 2019 estas estimaciones también las ofrecen a nivel distrito. El distrito es un nivel político-administrativo que se sitúa entre medias entre el condado y el estado, ya que un distrito está formado por varios condados, y el conjunto de distritos conforman el estado.

El estado de Iowa cuenta con 9 distritos que se nombran por su posición geográfica dentro del estado. En la Figura 26 puede verse la posición y los condados que abarca cada uno de estos 9 distritos.

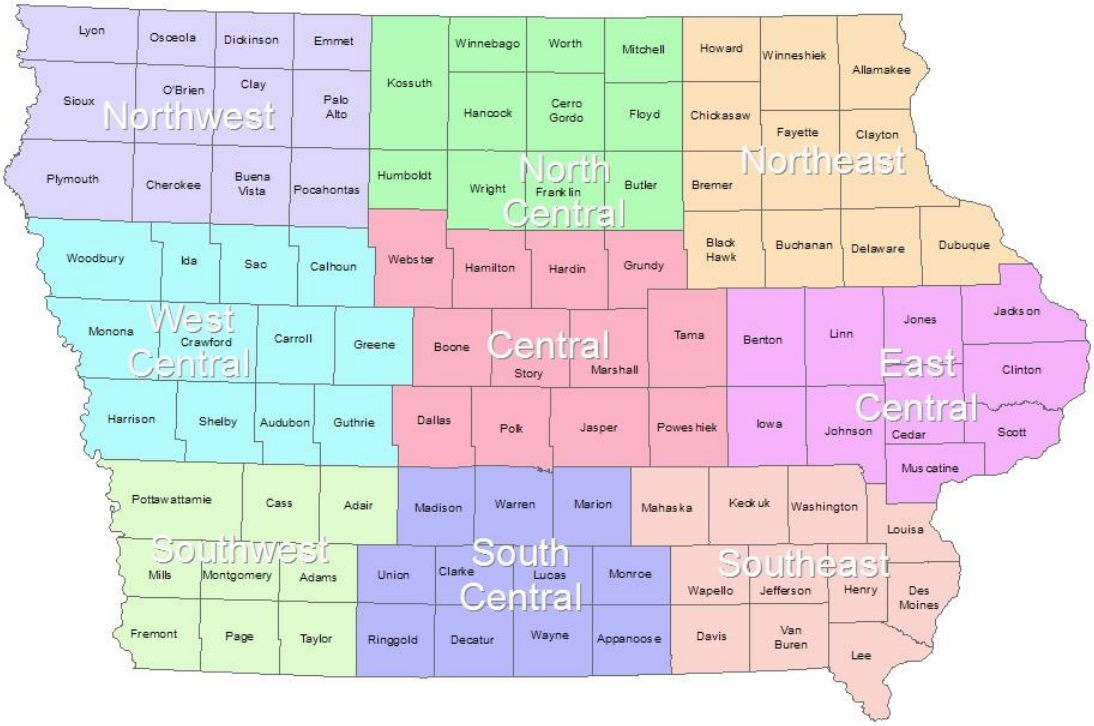


Figura 26: Distritos del estado de Iowa. Fuente: U.S. Department of Agriculture Crop Reporting Districts

Una vez identificados los distritos que conforman el estado de Iowa, se pasa a recabar los informes mensuales que emite el NASS a lo largo de la campaña. En la siguiente tabla pueden verse las predicciones de productividad en las 4 fechas en que se realiza la predicción junto con la productividad real final que se obtuvo.

Tabla 5: Productividad real y estimaciones del NASS para los distritos de Iowa para el año 2019

Distritos	Productividad real (bu/acre)	Estimaciones NASS (bu/acre)			
	20 Feb-20	1 Nov-19	1 Oct-19	1 Sep-19	1 Ago-19
Central	202,1	198	194	194	197
East Central	195,5	193	189	188	184
North Central	198,6	190	190	191	194
Northeast	205,4	204	204	200	199
Northwest	195,3	188	192	191	197
South Central	161	162	166	159	162
Southeast	174,9	165	175	173	172
Southwest	194,2	191	186	186	182
West Central	213,1	201	201	200	194

Como puede verse, las estimaciones de predicción emitidas por el NASS tienen una desviación respecto con el resultado final de la campaña de producción del año 2019.

Para conocer la bondad del modelo desarrollado el mismo trabajo que desarrolla el NASS, se comparan los resultados de las predicciones del modelo agregadas para cada distrito junto con las estimaciones del NASS.

Tabla 6: Productividad real y predicciones del modelo para los distritos de Iowa en el año 2019

Distritos	Productividad real (bu/acre)	Predicciones Modelo (bu/acre)			
	20 Feb-20	1 Nov-19	1 Oct-19	1 Sep-19	1 Ago-19
Central	202,1	209,7	205,3	211,4	203,4
East Central	195,5	197,5	194,5	200,1	190,6
North Central	198,6	212,1	208,8	215,6	209,8
Northeast	205,4	209,3	204,6	210,5	204,1
Northwest	196,0	204,5	201,9	204,9	203,1
South Central	166,9	179,0	177,3	180,7	177,0
Southeast	174,9	185,9	184,6	188,8	182,7
Southwest	195,1	200,9	196,3	200,0	192,4
West Central	213,7	214,7	210,7	214,3	208,2

Como puede verse las Tablas 5 y 6, la productividad real en los distritos de Northwest, South Central, Southeast, Southwest, West Central no coincide entre la ofrecida por el NASS y la calculada a partir de la producción y los acres cosechados de cada condado y utilizada en el modelado. Esto es debido a la situación

ya comentada que para el año 2019 solo se dispone de datos para 88 condados, por lo que faltan datos de 11 condados que pertenecen a los distritos donde hay diferencias. El NASS realiza una estimación de los condados que faltan en un agregado de "otros condados" pero no ofrece esos resultados. Esto provoca que al calcular los rendimientos promedios en los distritos en los que no se disponen los datos de todos los condados en 2019, los resultados sean ligeramente distintos al no estar considerando los mismos condados por parte del NASS y los considerados en este trabajo. Para los distrititos donde si se disponen de todos los condados, la productividad promedio del distrito es la misma que la ofrecida por el NASS, por lo que eso nos indica que la metodología del cálculo es correcta.

Es por esto que la comparación del nivel de acierto del NASS y del modelo desarrollado en este trabajo hay que contra los valores correspondientes en cada caso.

Para conocer la bondad del modelo desarrollado en este trabajo, se presentan los resultados del error para las estimaciones de productividad emitidas por el NASS y las productividades calculadas a partir de las predicciones del modelo desarrollado en este trabajo es los mismos distritos y las mismas fechas de predicción.

Tabla 7: Comparación de los errores del NASS y del modelo respecto a la producción real

Distrito	1 de Nov		1 de Oct		1 de Sep		1 de Ago	
	NASS	Modelo	NASS	Modelo	NASS	Modelo	NASS	Modelo
Central	2,05%	3,73%	4,03%	1,57%	4,03%	4,59%	2,54%	0,63%
East Central	1,30%	0,98%	3,34%	0,55%	3,85%	2,32%	5,90%	2,53%
North Central	4,34%	6,80%	4,34%	5,12%	3,84%	8,57%	2,33%	5,63%
Northeast	0,66%	1,92%	0,66%	0,38%	2,61%	2,52%	3,09%	0,60%
Northwest	3,74%	4,32%	1,69%	2,97%	2,20%	4,51%	0,87%	3,62%
South Central	0,62%	7,28%	3,11%	6,23%	1,24%	8,30%	0,62%	6,05%
Southeast	5,66%	6,33%	0,06%	5,56%	1,09%	7,98%	1,66%	4,48%
Southwest	1,65%	2,95%	4,22%	0,59%	4,22%	2,51%	6,28%	1,38%
West Central	5,68%	0,50%	5,68%	1,40%	6,15%	0,30%	8,96%	2,55%
PROMEDIO	2,85%	3,87%	3,01%	2,71%	3,25%	4,62%	3,58%	3,05%

Como se ve en la Tabla 7, las predicciones del modelo no son muy diferentes respecto a las estimaciones emitidas por el NASS. En las fechas del 1 de noviembre y el 1 de septiembre, las estimaciones del NASS tienen un error promedio a las que puede ofrecer el modelo. En cambio, en las fechas del 1 de octubre y 1 de agosto, las predicciones del modelo tienen un error promedio menor a las emitidas por el NASS. Es por esto, que los resultados del modelo pueden equipararse sin lugar a duda a las estimaciones emitidas por el NASS.

A continuación, puede verse en la Figura 27 una comparativa gráfica de los niveles de error en cada distrito para las 4 fechas que se comparan los niveles de error del NASS y del modelo.

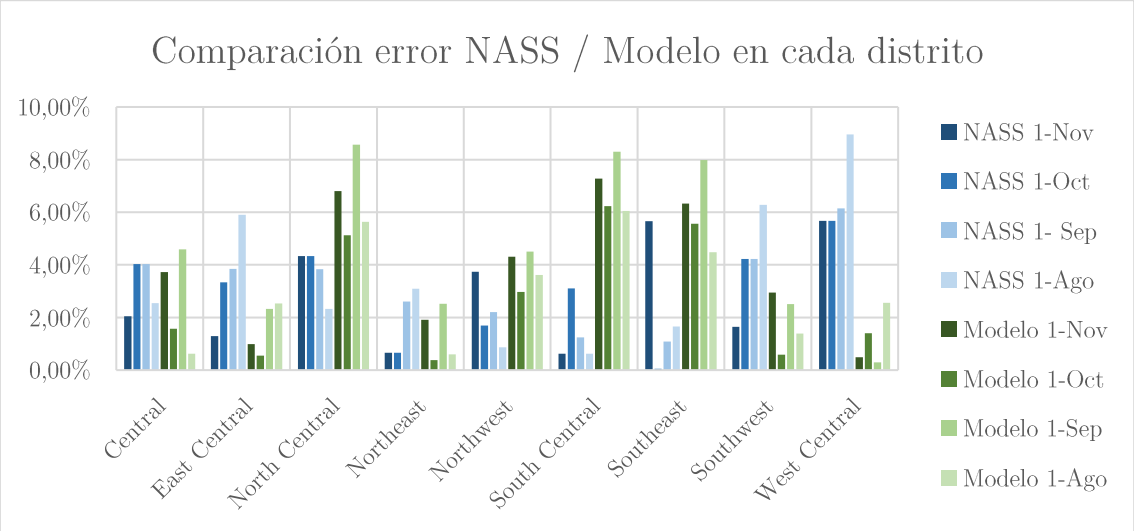


Figura 27: Comparación entre el error del NASS y del modelo para cada distrito en el 2019

Cabe destacar que para los distritos de East Central y West Central el modelo ha obtenido menores desviación para todas las fechas comparadas respecto a las estimaciones del NASS.

Capítulo 7
Conclusiones y trabajos futuros

Una vez expuestos en los capítulos anteriores los trabajos realizados y sus resultados, a continuación, se realiza una enumeración de las conclusiones extraídas del trabajo, así como una propuesta de trabajos futuros.

7.1. Conclusiones

1. Queda validado que el enfoque Machine Learning es adecuado para desarrollar herramientas de predicción de producción de maíz en Estados Unidos. Los errores que presenta el modelo desarrollado permiten su utilización como herramienta de toma de decisiones en un entorno real, ya que para 2019, ha presentado un error promedio para todos los condados del estado de Iowa del 3,8%. Además, si comparamos las predicciones del modelo contra las estimaciones que realiza el NASS, la diferencia entre ambas es mínima. De 4 fechas en las que se pueden comparar las predicciones, son mejores las emitidas por el del NASS en dos ocasiones, y mejores las emitidas por el modelo en otras dos fechas.

2. Se ha comprobado que no existe un número fijo de años de datos históricos para la predicción de producción de un año posterior. En este trabajo se ha determinado que son necesarios como mínimo 3 años. No siempre disponer de más años históricos puede ser positivo, ya que, si se utilizan datos de años históricos muy alejados en el tiempo, puede que se esté introduciendo en el modelo un aprendizaje de una realidad que ya no es válida en la actualidad, ya que las condiciones y herramientas para la producción han ido mejorándose con el paso del tiempo.

3. El análisis del momento en que las predicciones tienen un buen nivel de acierto, indica que a partir del 1 de julio pueden iniciarse las predicciones de productividad de maíz con un nivel de error que no es muy diferente al que se consigue si la predicción se realiza a finales de noviembre.

4. El desarrollo de modelos predictivos en maíz en Estados Unidos solo es posible gracias a que existen datos disponibles, por lo que, si existieran bases de datos similares en España para otros cultivos, sería posible que la cadena agroalimentaria pudiera extraer valor a los datos históricos de desarrollo de cultivos.

7.2. Propuesta de trabajos futuros

1. Este trabajo tiene potencial de mejora, ya que los datos utilizados provienen de fuentes de información que han sido seleccionadas por su facilidad de acceso y simpleza operativa. Con tiempo y recursos, se puede realizar una mejor captura de datos que se debería traducir en una mejora de los niveles de error de los modelos.

2. La agregación temporal de las variables para la construcción de los atributos del modelo ha sido mensual. Es posible que, realizando un resumen de la información en espacios temporales más pequeños como la semana o la quincena, se pudiesen lograr mejores resultados en el modelado.

3. Los cultivos no se rigen por fechas calendario, si no que tienen sus etapas fenológicas en función de las fechas de siembra y las condiciones climatológicas. Es por esto que sería interesante ligar la construcción de atributos a momentos fenológicos concretos y no a semanas o meses calendario, ya que las etapas fenológicas del cultivo no tienen por qué coincidir exactamente año tras año en las mismas fechas del calendario.

Bibliografía

1. Ameline, M., Fieuzal, R., Betbeder, J., Berthoumieu, J. -., & Baup, F. (2019). Estimation of corn yield by assimilating SAR and optical time series into a simplified agro-meteorological model: From diagnostic to forecast. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(12), 4747-4760. doi:10.1109/JSTARS.2018.2878502

2. Bhojani, S. H., & Bhatt, N. (2020). Wheat crop yield prediction using new activation functions in neural network. Neural Computing and Applications, 32(17), 13941-13951. doi:10.1007/s00521-020-04797-8

3. Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F., Reichert, G., 2015. Evaluation of the integrated Canadian crop yield forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric. For. Meteorol. 206, 137–150.

4. Guo, R., Zhu, X., Li, S., & Hou, C. (2020). Monitoring and forecasting method of winter wheat yield in shandong province. [山东省冬小麦单产监测与预报方法研究] Nongye Jixie Xuebao/Transactions of the Chinese Society for Agricultural Machinery, 51(7), 156-163. doi:10.6041/j.issn.1000-1298.2020.07.018

5. Hoogenboom, H., 2000. Contribution of agrometeorology to the simulation of crop production and its applications. Agric. For. Meteorol. 103, 137–157.

6. Hufford, M. B., Bilinski, P., Pyhäjärvi, T. & Ross-Ibarra, J. Teosinte as a model system for population and ecological genomics. Trends Genet. 28, 606–615 (2012).

7. IPCC, 2013. Climate Change 2013: The Physical Science Basis. Report. Intergovernmental Panel on Climate Change (IPCC).

8. Jiang, Z., Liu, C., Ganapathysubramanian, B., Hayes, D. J., & Sarkar, S. (2020). Predicting county-scale maize yields with publicly available data. Scientific Reports, 10(1) doi:10.1038/s41598-020-71898-8

9. Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., Ritchie, J., 2003. The DSS, cropping system model. Eur. J. Agron. 18, 235–265.

10. Kandiannan, K., Chandaragiri, K., Sankaran, N., Balasubramanian, T., Kailasam, C., 2002. Crop-weather model for turmeric yield forecasting for Coimbatore district, Tamil Nadu, India. Agric. For. Meteorol. 112, 133–137.

11. Kandiannan, K., Chandaragiri, K., Sankaran, N., Balasubramanian, T., Kailasam, C., 2002. Crop-weather model for turmeric yield forecasting for Coimbatore district, Tamil Nadu, India. Agric. For. Meteorol. 112, 133–137.

12. Kotlowski, K., 2007. Qualitative Models of Climate Variations Impact on Crop Yields. Technical Report IR-07-034. IIASA Interim Report.

13. Lobell, D., Cahill, K., Field, C., 2007. Historical effects of temperature and precipitation on California crop yields. Clim. Change 81, 187–203.

14. Newlands, N.K., Zamar, D.S., Kouadio, L.A., Zhang, Y., Chipanshi, A., Potgieter, A., Toure, S., Hill, H.S.J., 2014. An integrated, probabilistic model for improved seasonal forecasting of agricultal crop yield under environmental uncertainty. Front. Environ. Sci. 2. http://dx.doi.org/10.3389/fenvs.2014.00017.

15. Pede, T., Mountrakis, G., & Shaw, S. B. (2019). Improving corn yield prediction across the US corn belt by replacing air temperature with daily MODIS land surface temperature. Agricultural and Forest Meteorology, 276-277 doi:10.1016/j.agrformet.2019.107615

16. R.W. Malone, et al, 2009, Quasi-biennial corn yield cycles in Iowa, Agricultural and Forest Meteorology, Volume 149, Issues 6–7, Pages 1087-1094,

17. Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. Agricultural and Forest Meteorology, 284 doi:10.1016/j.agrformet.2019.107886

18. Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science, 11 doi:10.3389/fpls.2020.01120

19. Stepanov, A., Dubrovin, K., Sorokin, A., & Aseeva, T. (2020). Predicting soybean yield at the regional scale using remote sensing and climatic data. Remote Sensing, 12(12) doi:10.3390/rs12121936

20. Tack, J., Coble, K. H., Johansson, R., Harri, A., & Barnett, B. J. (2019). The potential implications of "Big ag data" for USDA forecasts. Applied Economic Perspectives and Policy, 41(4), 668-683. doi:10.1093/aepp/ppy028

21. Tannura, M.A., Irwin, S.H., Good, D.L., 2008. Weather, Technology, and Corn and Soybean Yields in the US Corn Belt. Marketing and Outlook Res. Rep. No. 2008-01.

22. Taylor, S., Carlson, R., 1997. Weather and yield trends. ICM Conference 175–188.

23. Thompson, L., 1988. Effects of changes in climate and weather variability on the yield of corn and soybean. J. Agro Prod. 1, 20–27.

24. Tian, L., Wang, C., Li, H., & Sun, H. (2020). Yield prediction model of rice and wheat crops based on ecological distance algorithm. Environmental Technology and Innovation, 20 doi:10.1016/j.eti.2020.101132

25. Jiang, Z., Liu, C., Ganapathysubramanian, B. et al. Predicting county-scale maize yields with publicly available data. Sci Rep 10, 14957 (2020). https://doi.org/10.1038/s41598-020-71898-8

26. Ceglar, A., Toreti, A., Prodhomme, C. et al. Land-surface initialisation improves seasonal climate prediction skill for maize yield forecast. Sci Rep 8, 1322 (2018). https://doi.org/10.1038/s41598-018-19586-6

27. Hansen, J.W., Indeje, M., 2004. Linking dynamic seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. Agric. For. Meteorol. 125, 143–157. http://dx.doi.org/10.1016/j.agrformet.2004.02.006.

28. Hansen, J.W., Potgieter, A., Tippett, M., 2004. Using a general circulation model to forecast regional wheat yields in Northeast Australia. Agric. For. Meteorol. 127, 77–92. http://dx.doi.org/10.1016/j.agrformet.2004.07.005.

29. Basso, B., Cammarano, D., Carfagna, E., 2013. Review of crop yield forecasting methods and early warning systems. Report Presented to First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics.

Anexos

Se exponen las partes del código Python o R utilizadas en el trabajo que más importancia tienen. Las partes del código utilizadas para la creación de figuras o otros menesteres no han sido incluidas para no saturar el trabajo con demasiadas páginas con código.

Anexo 1: Código Python utilizado para la descarga de índices satelitales

```
ruta = "/content/drive/My Drive/TFM DigitalAgri/Counties/"

now=datetime.now()
current_time=now.strftime("%H:%M:%S")
print ("Hora inicio: ",current_time)

#Definimos las fechas de inicio y fin de la colección
start_date='2000-01-01'
end_date='2019-12-31'

#Buble for para calcular y extraer los índices de cada condado de
una forma automática
for file in os.listdir(ruta):
    if file.endswith(".geojson"):
        try:
          name = file.split(".")[0]
          now=datetime.now()
          current_time=now.strftime("%H:%M:%S")
          print("A las:",current_time,"se inicial el calculo de",na
me)
          path_file = os.path.join(ruta, file)
          #print(file)
          county = batch.Import.table.fromGeoJSON(path_file)
          county_1 = county[0].geometry()

          #Creamos una colección de imágenes
          coleccion = ee.ImageCollection('MODIS/006/MOD13A2').selec
t('EVI').filterBounds(county_1)\
                    .filterDate(start_date, end_date)\
          coleccion2 = ee.ImageCollection('MODIS/006/MOD13A2').sele
ct('NDVI').filterBounds(county_1)\
                    .filterDate(start_date, end_date)\
```

```
            escala=10
            grafico=chart.Image.series(**{
                  'imageCollection': coleccion,
                  'region': county_1,
                  'scale': escala,
              })
            grafico2=chart.Image.series(**{
                  'imageCollection': coleccion2,
                  'region': county_1,
                  'scale': escala,
              })

            formatter = dt.DateFormatter('%Y-%m-
%d')  # Specify the format - %b gives us Jan, Feb...
            locator = dt.MonthLocator()  # every month

            df=grafico.dataframe
            df2=grafico2.dataframe
            #df = df.append(df2)

            df.reset_index(inplace=True)
            df2.reset_index(inplace=True)

            df = df.append(df2)

            df.rename(columns={'index': 'Fecha'}, inplace=True)

            df['Anno'] = pd.DatetimeIndex(df['Fecha']).year
            df['Mes'] = pd.DatetimeIndex(df['Fecha']).month

            r = df.groupby(['Anno', 'Mes']).mean()

            r['NDVI'] = r['NDVI'].round(0)
            r['EVI'] = r['EVI'].round(0)

            ruta_csv = ruta+'NDVI Data por County/'
            r.to_csv (r'{}{}.csv'.format(ruta_csv, name), index = Tru
e, header=True)
            print("Todo bien con {}".format(name))
          except:
            print("ERROR: Algo ha ido mal con {}".format(name))

now=datetime.now()
current_time=now.strftime("%H:%M:%S")
print ("Hora fin: ",current_time)
```

Anexo 2: Código Python para los tratamientos de limpieza y reducción dimensional

```
import os
import pandas as pd

ruta = '/content/drive/My Drive/TFM DigitalAgri/superdataset.csv'
path_file = os.path.join(ruta)
SDS = pd.read_csv(path_file)

#Se eliminan las columnas con valores únicos.
valor_unico = SDS.columns[SDS.nunique() <= 1]
SDS = SDS.drop(columns = valor_unico)

#Se eliminan las columnas con >10% de valores ausentes
porcentage_MAX_NAN = 0.1
max_cel_NAN = SDS.count()*porcentage_MAX_NAN
SDS = SDS[SDS.columns[SDS.isnull().sum()<max_cel_NAN]]

#Se imputan los valores ausentes con la media de la columna
SDS = SDS.fillna(SDS.mean())
```

Anexo 3: Código Python para la búsqueda automática de la mejor combinación de hiper-parámetros en un algoritmo.

Se muestra a modo de ejemplo el código utilizado para el algoritmo LinearRegression. Para otros modelos se utiliza del mismo modo, y solo hay que modificar el *grid* que le corresponda a cada modelo.

```
from sklearn.model_selection import GridSearchCV

# defining parameter range
param_grid = {'fit_intercept': ['True','False'],
              'normalize': ['True','False'],
              'copy_X': ['True','False']}

grid = GridSearchCV(LinearRegression(), param_grid, refit = True, v
erbose = 3)

# fitting the model for grid search
grid.fit(X, y)


# print best parameter after tuning
print("************************************")
print(grid.best_params_)

# print how our model looks after hyper-parameter tuning
print(grid.best_estimator_)

print(grid.best_index_)
```

Anexo 4: Código Python para la validación cruzada “leave one year out” del modelo.

Se muestra a modo de ejemplo el código utilizado para el algoritmo LinearRegression. Para otros modelos se utiliza del mismo modo, y solo hay que definir el algoritmo que corresponda.

```
MAE_list = []
MAPE_list = []
R2_list = []
Explicacion_var = []
Error_max = []

for año in range(2000,2019):

    filtro_año_test = X['ANIO'] == año
    filtro_año_train = X['ANIO'] != año

    Año_train = SDS[filtro_año_train]
    Año_test = SDS[filtro_año_test]


    X_train = Año_train.drop(columns = 'Yield')
    X_train = X_train.drop(columns = 'ANSI')
    y_train = Año_train['Yield']

    X_test = Año_test.drop(columns = 'Yield')
    X_test = X_test.drop(columns = 'ANSI')
    y_test = Año_test['Yield']


    linear_regression = LinearRegression(fit_intercept=True, normal
ize=True, copy_X=True, n_jobs=None)
    linear_regression.fit(X_train,y_train)

    y_pred = linear_regression.predict(X_test)
    error = abs(y_test - y_pred)

    print("***ANALISIS DEL ERROR***")

    MAE = mean_absolute_error(y_test, y_pred)
    print("El MAE es:",MAE)

    MAPE = MAE/y_test.mean()
    print("El MAPE es:",MAPE)
```

```
R2 = r2_score(y_test, y_pred)
print("El R2 es:",R2)

explicación_var = explained_variance_score(y_test, y_pred)
print("La explicación de la varianza es:",explicación_var)

error_max = max_error(y_test, y_pred)
print("El error máximo es:",error_max)


MAE_list.append(MAE)
MAPE_list.append(MAPE)
R2_list.append(R2)
Explicacion_var.append(explicación_var)
Error_max.append(error_max)
```

Anexo 5: Código en R para la construcción del SDS

```
#install.packages("tidyverse")
library("tidyverse")

#install.packages("foreign")
library("foreign")

#install.packages("data.table")
library("data.table")

#install.packages("reshape2")
library("reshape2")

input_path_wea <- "C:/Users/Miguel Angel/OneDrive/Master Digital-
Agri/TFM - Trabajo Final de Máster/Data Sources Nuevas/Weather/"
output_path <- "C:/Users/Miguel Angel/OneDrive/Master Digital-
Agri/TFM - Trabajo Final de Máster/Preproceso/Event tables/"
input_path_sat <- "C:/Users/Miguel Angel/OneDrive/Master Digital-
Agri/TFM - Trabajo Final de Máster/Data Sources Nuevas/Satelital
Indexes/"
input_path_yield <- "C:/Users/Miguel Angel/OneDrive/Master Digital-
Agri/TFM - Trabajo Final de Máster/Data Sources Nuevas/Datos USDA
Corn/"

#######
# Wea #
#######
weas <- list.files(input_path_wea, pattern=".csv")

for(r in 1:length(weas)){
data <- as.data.table(fread(paste0(input_path_wea, weas[r])))

data[, ANIO:=as.numeric(substring(DATE,1,4))]

data[, MES:=as.numeric(substring(DATE,6,8))]

data[,c("STATION","NAME","DATE"):=NULL,with=F]

# Media por fecha de todas las estaciones (na.rm=T)

data <- data[, lapply(.SD,function(x){mean(x, na.rm=T)}),
by=c("ANIO","MES")]

att_estatico <- c("LATITUDE", "LONGITUDE", "ELEVATION")
```

```
data <- data[, (att_estatico) :=lapply(.SD,function(x){mean(x,
na.rm=T)}), .SDcols = att_estatico]

data[,ANSI:=str_sub(weas[r],end=-5)]

setcolorder(data, c("ANSI","ANIO","MES"))

# Darle la vuelta: Trasponerla
latitude <- unique(data$LATITUDE)
longitude <- unique(data$LONGITUDE)
elevation <- unique(data$ELEVATION)

data <- data[,-c("LONGITUDE","LATITUDE","ELEVATION")]

for(i in 1:length(names(data[,-c("ANIO","ANSI","MES"),with=F]))){
  data_estudio <- cbind(data[,c("ANIO","ANSI","MES"),with=F],
                        data[,names(data[,-
c("ANIO","ANSI","MES"),with=F])[i],with=F])

  data_estudio_t <- reshape(data_estudio,
                            timevar = "MES",
                            idvar = c("ANSI","ANIO"),
                            direction = "wide")

  if(i==1){ data_new<- data_estudio_t}else{data_new <-
merge(data_new, data_estudio_t,by=c("ANIO","ANSI"))}
}

data_new[, LONGITUDE:=longitude]
data_new[, LATITUDE:=latitude]
data_new[, ELEVATION:=elevation]

setcolorder(data_new,c("ANSI","ANIO","LONGITUDE","LATITUDE","ELEVAT
ION"))

fwrite(data_new, paste0(output_path, "wea_", weas[r]))
print(paste0("Va por el condado ", weas[r],"\n"))
}

# Tabla de eventos

############
# Satelite #
############

sates <- list.files(input_path_sat, pattern=".csv")
for(r in 1:length(sates)){
```

```
#Dato enero con la media de los demás eneros

data <- as.data.table(fread(paste0(input_path_sat, sates[r])))

att_estatico <- c("NDVI","EVI")
data_enero <- data[Mes==1, lapply(.SD,function(x){mean(x,
na.rm=T)}), .SDcols = att_estatico]

data<-
rbind(data,as.list(c(2000,1,data_enero$EVI,data_enero$NDVI)))

# Ansi

data[,ANSI:=str_sub(sates[r],end=-5)]

setnames(data, "Anno", "ANIO")

setnames(data, "Mes", "MES")


setorder(data, "ANSI","ANIO", "MES")
setcolorder(data, c("ANSI","ANIO", "MES"))

# Trasponerla

for(i in 1:length(names(data[,-c("ANIO","ANSI","MES"),with=F]))){
  data_estudio <- cbind(data[,c("ANIO","ANSI","MES"),with=F],
                        data[,names(data[,-
c("ANIO","ANSI","MES"),with=F])[i],with=F])

  data_estudio_t <- reshape(data_estudio,
                            timevar = "MES",
                            idvar = c("ANSI","ANIO"),
                            direction = "wide")

  if(i==1){ data_new<- data_estudio_t}else{data_new <-
merge(data_new, data_estudio_t,by=c("ANIO","ANSI"))}
}

fwrite(data_new, paste0(output_path, "sat_", sates[r]))
print(paste0("Va por el condado ", sates[r],"\n"))
}

# Tabla de eventos

#########
# Yield #
#########
```

```
yield_name <- list.files(input_path_yield, pattern=".csv")

data_yield <-
as.data.table(fread(paste0(input_path_yield,yield_name)))
setnames(data_yield,c("Data Item","County
ANSI","Value","Year"),c("Data_Item","ANSI","Yield","ANIO"))
names_buenos <- c("ANIO","ANSI","Yield")

data_yield_new <- data_yield[Data_Item %in% c("CORN, GRAIN - YIELD,
MEASURED IN BU / ACRE"),names_buenos,with=F]

for(i in 1:length(unique(data_yield_new$ANSI))){
  data_estudio <- data_yield_new[ANSI %in%
unique(data_yield_new$ANSI)[i],]

  fwrite(data_estudio, paste0(output_path,
"yie_",unique(data_yield_new$ANSI)[i],".csv"))
  print(paste0("Va por el condado ",
unique(data_yield_new$ANSI)[i],"\n"))
}

##################
# UNIÓN POR ANSI #
##################

output_path_ansi <- "C:/Users/Miguel Angel/OneDrive/Master Digital-
Agri/TFM - Trabajo Final de Máster/Preproceso/Final event tables/"


ansi <- unique(as.numeric(str_sub(list.files(output_path,
pattern=".csv"),5,-
5)))[!is.na(unique(as.numeric(str_sub(list.files(output_path,
pattern=".csv"),5,-5))))]


files <- list.files(output_path, pattern=".csv")

data_ONI <- as.data.table(fread(paste0(output_path,"ONI.csv")))
data_QBO <- as.data.table(fread(paste0(output_path,"QBO.csv")))
data_NAO <- as.data.table(fread(paste0(output_path,"NAO.csv")))
data_SOI <- as.data.table(fread(paste0(output_path,"SOI.csv")))

for(i in 1:length(ansi)){
  data_yield <-
as.data.table(fread(paste0(output_path,"yie_",ansi[i],".csv")))
  data_wea <-
as.data.table(fread(paste0(output_path,"wea_",ansi[i],".csv")))
  data_sat <-
as.data.table(fread(paste0(output_path,"sat_",ansi[i],".csv")))
```

```
  data_aux <- merge(data_yield,data_wea,by=c("ANSI","ANIO"))
  data <- merge(data_aux, data_sat, by=c("ANSI","ANIO"))

  data <- merge(data, data_SOI, by=c("ANIO"))
  data <- merge(data, data_NAO, by=c("ANIO"))
  data <- merge(data, data_ONI, by=c("ANIO"))
  data <- merge(data, data_QBO, by=c("ANIO"))


  fwrite(data, paste0(output_path_ansi, ansi[i],".csv"))
  print(paste0("Va por el condado ", ansi[i],"\n"))
}

######################
# Final superdataset #
######################

output_path_final <- paste0("C:/Users/Miguel Angel/OneDrive/Master
Digital-Agri/TFM - Trabajo Final de
Máster/Preproceso/","Superdataset/")

superdataset <- c()

for (i in 1:length(list.files(output_path_ansi,pattern=".csv"))){

  data <-
data.table(fread(paste0(output_path_ansi,list.files(output_path_ans
i,pattern=".csv")[i])))

  if(i==1){superdataset <- data}else{superdataset <-
rbind(superdataset, data,fill=T)}

  print(paste0("Va por el fichero ",
list.files(output_path_ansi,pattern=".csv")[i]))
}

fwrite(superdataset, paste0(output_path_final, "superdataset.csv"))
```