Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado.

Autores:
Andrea Carolina Gamboa Ariza – 2132084
Paula Andrea Cáceres Ortiz - 2141712

Plan de proyecto de grado
Fecha de entrega: 17 de septiembre de 2018

Director:
PhD. Henry Lamos Díaz
Codirector:
Ing. David Esteban Puentes Garzón

Universidad Industrial de Santander
Facultad de Ingenierías Físico-Mecánicas
Escuela de Estudios Industriales y Empresariales
Bucaramanga
2018

Tabla de Contenido

Lista de Figuras

Lista de Apéndices

(Apéndices adjuntos en CD)

Introducción

La agricultura es una de las actividades de mayor contribución al crecimiento económico de la población en Colombia. Esta actividad económica es una de las que más aporta al crecimiento del PIB, gracias al fuerte ritmo que tomó durante el 2017, fue el sector que lideró el porcentaje con una variación de 4,9% DANE (n.d.). Para 2018, el aporte de la agricultura al PIB no se quedó atrás pues según las últimas cifras presentadas "En el primer trimestre de 2018, el valor agregado de agricultura tuvo una serie positiva de 2,0%" *DANE (2018).*

El cultivo de cacao contribuye en gran medida a dicho crecimiento, debido a que fue el cultivo que más creció porcentualmente en producción, pues pasó de producir 56.785 toneladas de cacao en el 2016 a producir 60.535 en el 2017, traduciéndose en un incremento del 6,6% y creándose un récord para el país en este sector FEDECACAO (2018). Del mismo modo este cultivo ha venido sustituyendo gradualmente hectáreas ilícitas brindando tranquilidad e ingresos dignos a los agricultores Semana- Comercio, (2018), por lo cual se ratificó como "el cultivo de la paz" según lo confirma el gobierno nacional ante agricultores y diferentes agremiaciones del sector cacaotero SAC-Sociedad de Agricultores de Colombia (n.d.)

Este importante crecimiento en los cultivos de cacao contribuye positivamente al estudio realizado en la Universidad Nacional sobre los proyectos productivos, en donde se expone que "el 60% de los excombatientes quieren especializarse en actividades agropecuarias" (Semana-Agricultura, 2017), esto se puede llevar a cabo gracias a las más de 40 millones de hectáreas aptas para la siembra con las que cuenta Colombia delimitadas a partir del mapa básico realizado por la Unidad de Planificación Rural Agropecuaria (UPRA) UPRA (n.d.).

La situación presentada anteriormente, favorece al departamento de Santander el cual tiene una alta participación en el sector agrícola "Teniendo presente que de los 87 municipios del Departamento 78 de ellos basan sus actividades en el Sector Agropecuario como principal renglón económico y cerca del 50% de la población vincula sus ingresos con actividades originadas en el Sector Rural" (Secretaría de agricultura, 2017).

El sector cacaotero está catalogado como uno de los sectores estratégicos en el departamento de Santander con un porcentaje de participación en el área nacional sembrada del 24% y una producción de aproximadamente 26.431,64 Ton/Año. (Plan de Desarrollo Departamental, 2016-2018).

En aras de contribuir con uno de los objetivos del Plan de Desarrollo Departamental, el cual busca "Fortalecer la agricultura familiar de tal forma que se garantice la Seguridad Alimentaria en el Departamento de Santander, integrando redes de intercambio de bienes y servicios rurales locales, regionales y/o nacionales" (Plan de Desarrollo Departamental, 2016-2018) y gracias a los avances recopilados en los trabajos:

- "Development of soft computing and applications in agricultural and biological engineering". (Huang et al., 2010).
- "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper" (Mishra, Mishra, & Santra, 2016).

Se origina el presente trabajo con el propósito de predecir los rendimientos de los cultivos a partir de la construcción de modelos de Machine Learning (Modelos de aprendizaje automático) que ayuden a agricultores, entidades gubernamentales y demás actores en la toma de decisiones importantes a fin de lograr un incremento en la productividad, calidad, sostenibilidad y competitividad de este sector.

1. Definición del proyecto

1.1. Título del proyecto

Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado.

1.2. Modalidad

Trabajo de investigación

1.3. Responsables

Nombre del autor: Andrea Carolina Gamboa Ariza
E-mail autor: andreagamboar.14@hotmail.com
Teléfono: 3164363156
Firma autor: _____

Nombre del autor: Paula Andrea Cáceres Ortiz
E-mail autor: pauandrea96_26@hotmail.com
Teléfono: 3153109505
Firma autor: _____

Nombre director: Henry Lamos Díaz
E-mail director: hlamos@uis.edu.co
Firma director: _____

Nombre codirector: David Esteban Puentes Garzón
E-mail codirector: dpuentesgarzon@gmail.com
Firma codirector: _____

Nombre del director del grupo de investigación: Carlos Eduardo Díaz Bohórquez
E-mail director del grupo de investigación: cediazbo@uis.edu.co
Firma director del grupo de investigación: _____

1.4. Nombre del grupo de investigación

Grupo de Optimización y Organización de Sistemas Productivos, Administrativos y Logísticos (OPALO).

2. Revisión de la literatura

2.1. Análisis bibliométrico

La búsqueda y el análisis de la información para el desarrollo del presente proyecto se hizo mediante la plataforma Web of Science, la cual pertenece a los recursos electrónicos de la Universidad Industrial de Santander, con el fin de encontrar referencias bibliográficas que permitieran extraer la mayor cantidad de información relacionada con diferentes modelos para la predicción de rendimientos agrícolas.

De una previa revisión de literatura realizada en la etapa inicial del proyecto se identificaron palabras clave tales como: *Agriculture, crop, yield, forecast, production y Machine Learning,* las cuales fueron utilizadas para construir la siguiente ecuación de búsqueda

TS = ((Agricult OR Crop OR Farm) AND (Forecast*) AND (Yield) AND (Product*) AND (Machine learning OR Regression))*

Figura 1. Ecuación de búsqueda.

El número de referencias bibliográficas que arrojó dicha ecuación fue de 101 investigaciones, sin embargo, se utilizó la herramienta de refinado de categorías excluyendo: *Energy fuels, remote sensing, economics, water resources y oceanography* con el fin de generar una búsqueda más refinada y alineada al presente trabajo; la cual arrojo 79 resultados delimitados en su mayoría en el idioma ingles entre los años 2001- 2018 y que a su vez tuvieran acceso institucional.

Con el objetivo de brindar mayor precisión a la búsqueda, se tuvieron en cuenta los siguientes criterios de selección:

- Artículos que tuvieran como resultado la predicción de rendimientos de cultivos agrícolas.
- Trabajos en los que se evidenciara el uso de modelos de aprendizaje automático y/o regresión para el tratamiento de los datos.

A partir de los criterios mencionados anteriormente se obtuvo un total de 23 referencias bibliográficas. Cabe mencionar que se utilizó la técnica *bola de nieve* para incluir algunos artículos útiles en la investigación para finalmente contar con un total de 27 artículos, los cuales sirvieron para extraer y recopilar la información relevante y necesaria para el presente proyecto de investigación. En la figura 2 se muestra el resumen gráfico de la selección de los artículos a analizar.

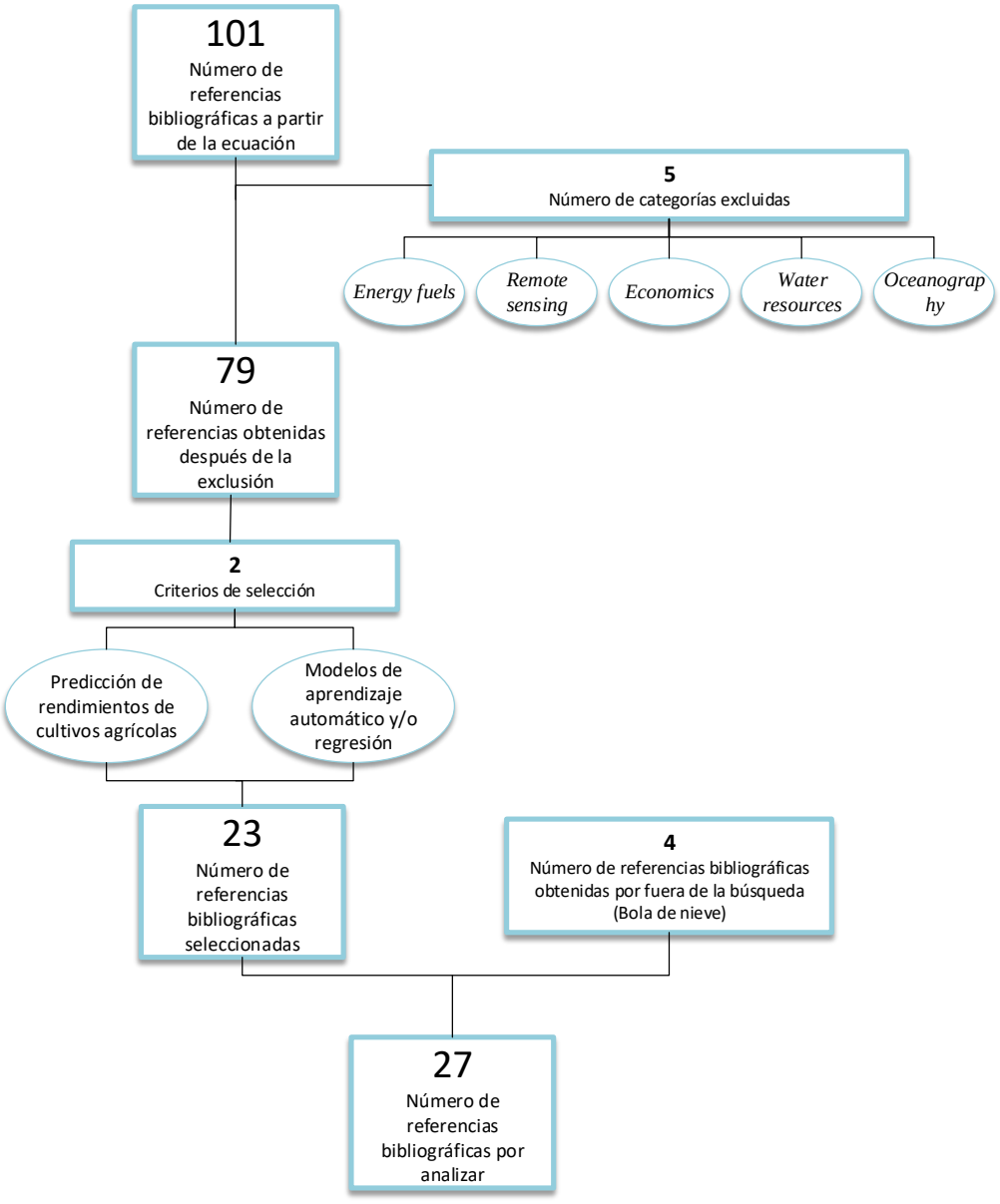

Figura 2. Artículos por analizar.

Previa a la lectura de los documentos se realizó un análisis descriptivo de la temática de investigación empleando el software Vantage Point, se observa en el análisis que el número de publicaciones anuales ha venido en descenso en los últimos 3 años tal y como se observa en la figura 2. Sin embargo, en el año 2014 el número de publicaciones fue mayor con respecto a los demás años. También se evidencia que el número de publicaciones con mayor frecuencia es 1, situación que se presenta en los años 2001, 2004, 2007, 2008, 2010, 2012, 2017, 2018, mientras que, en los años 2002, 2003, 2006 y 2013 no se muestra ningún registro de publicaciones relacionadas con el tema. Por otra parte, los años 2005, 2009 y 2011 coinciden en un total de 2 publicaciones en el año y en el 2015 y 2016 se registraron 3 y 4 publicaciones respectivamente.

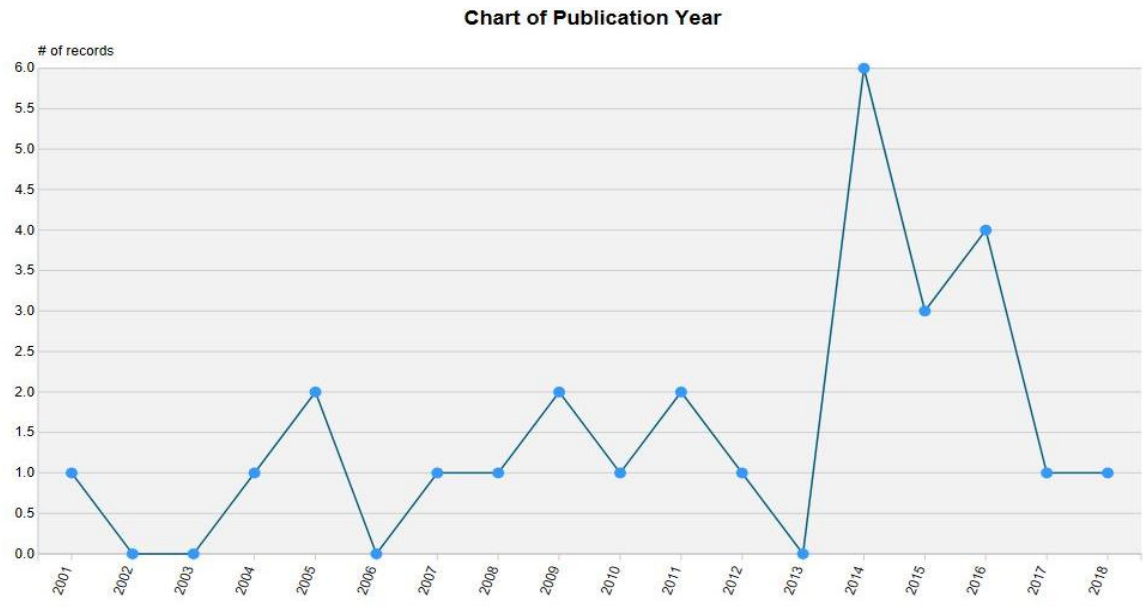

Figura 3. Número de publicaciones anuales. Adaptado de Vantage Point (2018).

Por otro lado, en la figura 3 se observa que, en relación al número de artículos publicados por país, Estados Unidos realiza una mayor cantidad de investigación en el campo del presente proyecto; situación soportada por una cifra de 8 artículos publicados. Seguido por Italia, el cual realizó un total de 5 publicaciones, mientras que Brasil, Canadá y España disminuyen en el número de publicaciones por año con 3 de ellas. El número de publicaciones disminuye a 2 para los países China, Alemania e India.

Finalmente, los países que investigan en menor proporción en temas relacionados con el presente proyecto son Argentina, Australia, Bélgica, Dinamarca, Grecia, Nueva Zelanda, Portugal,

Sur África, Corea del Sur, Sri Lanka, Suazilandia, Tunisia y Reino Unido. Cabe resaltar que, como se observa, Colombia no se encuentra dentro de los países con investigaciones en el presente tema.

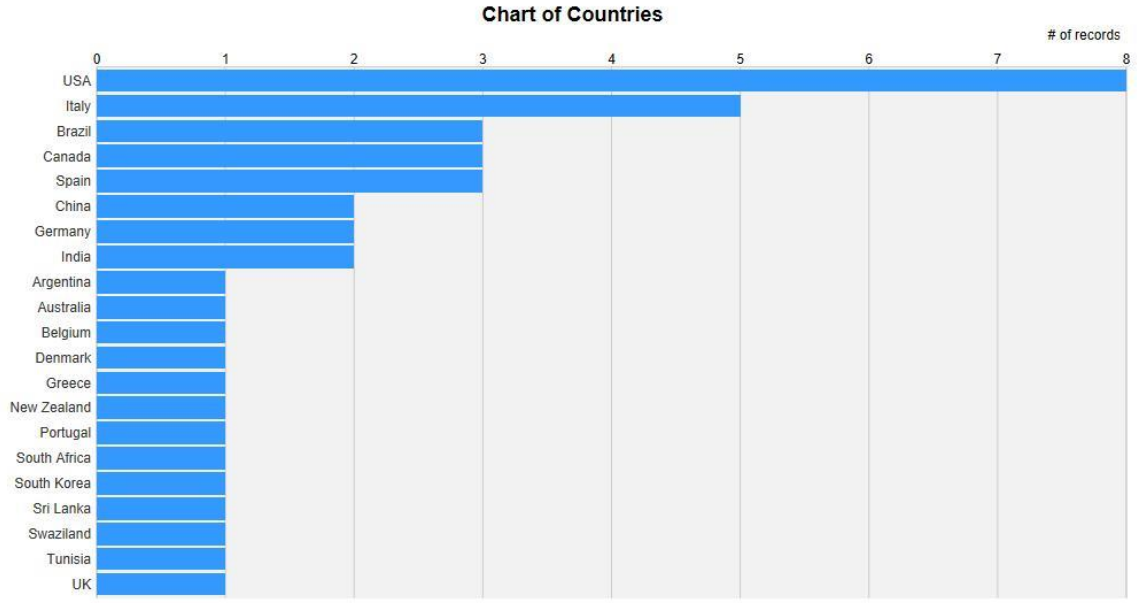

Figura 4. Numero de publicaciones por país. Adaptado de Vantage Point (2018).

La figura 4 ilustra un comportamiento detallado del número de publicaciones por año de cada país. Para el caso de Estados Unidos, se observa un comportamiento constante en el número artículos publicados, a diferencia de Argentina, Australia y España en los que el total de sus publicaciones se dieron en el año 2009, 2004 y 2014 respectivamente. En los años 2012, 2014, 2016 y 2017 Italia ha publicado al menos un artículo.

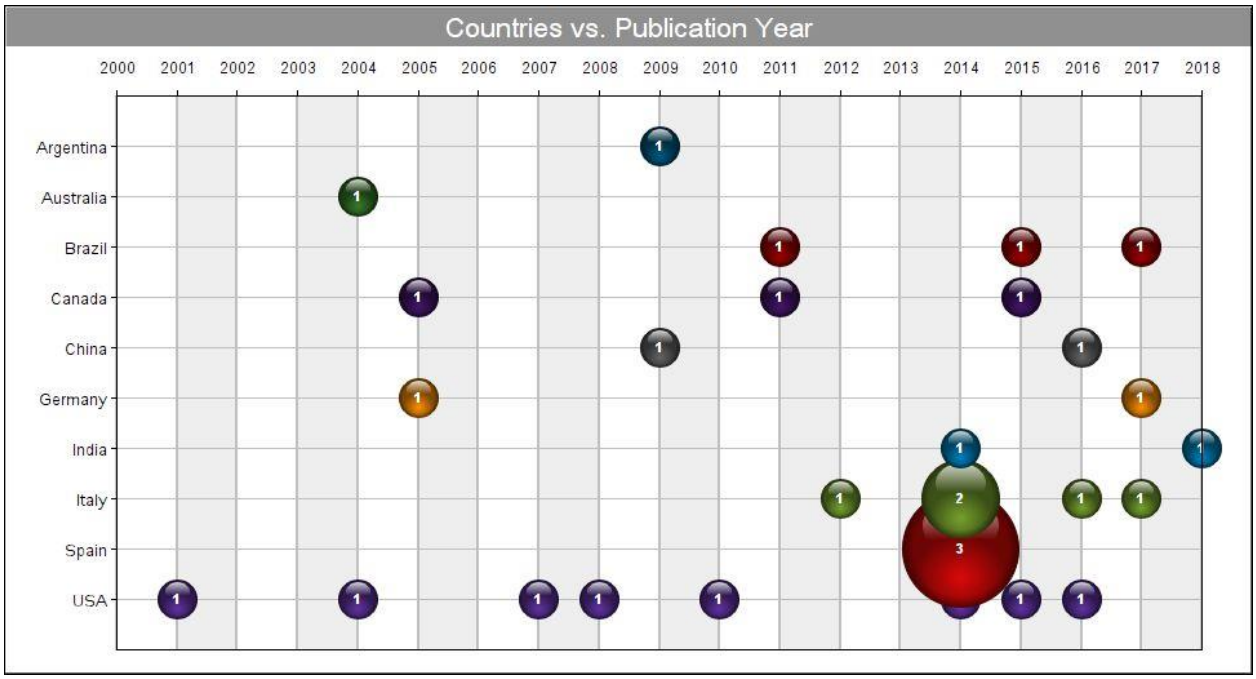

Figura 5. Número de artículos anuales por autor. Adaptado de Vantage Point (2018).

En la figura 5 se observa el número de publicaciones de los 10 principales autores, donde se muestra que Aguilera y Ruiz realizaron 2 publicaciones mientras que Abichou, Abreu, Acutis, Ainsworth, Alvarez, Barkley, Bedard Y Ben Dhiab solo hicieron una publicación.

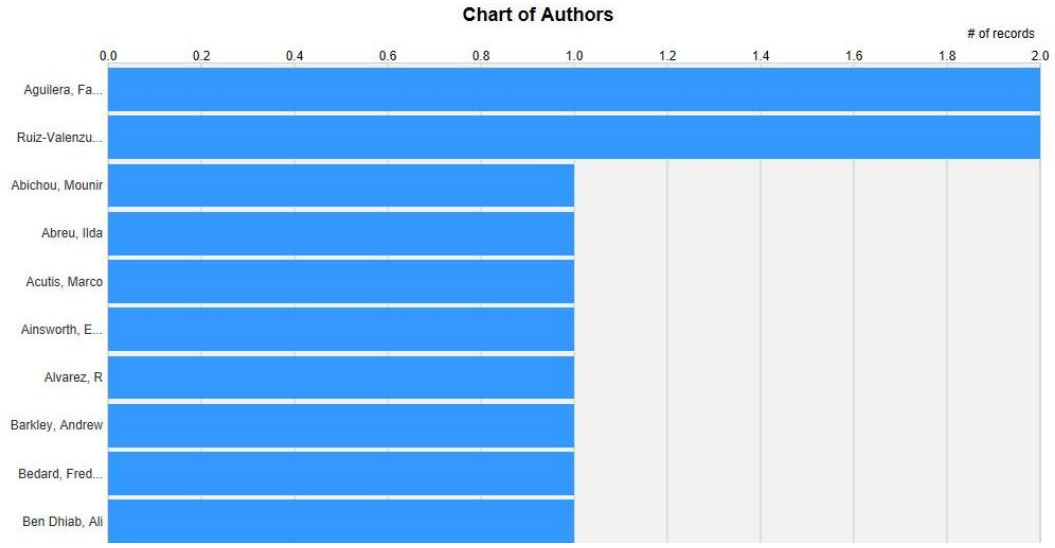

Figura 6. Numero de publicaciones por autor. Adaptado de Vantage Point (2018).

De los artículos seleccionados en la revisión de literatura, la Figura 6 presenta las principales palabras claves destacadas en todos ellos. En el gráfico de nube de palabras se expone que los términos más sobresalientes son: *Crop Forecasting, crop Model, regression, y yield*. De igual manera también resaltan palabras como: *Agriculture, wheat y crop*.

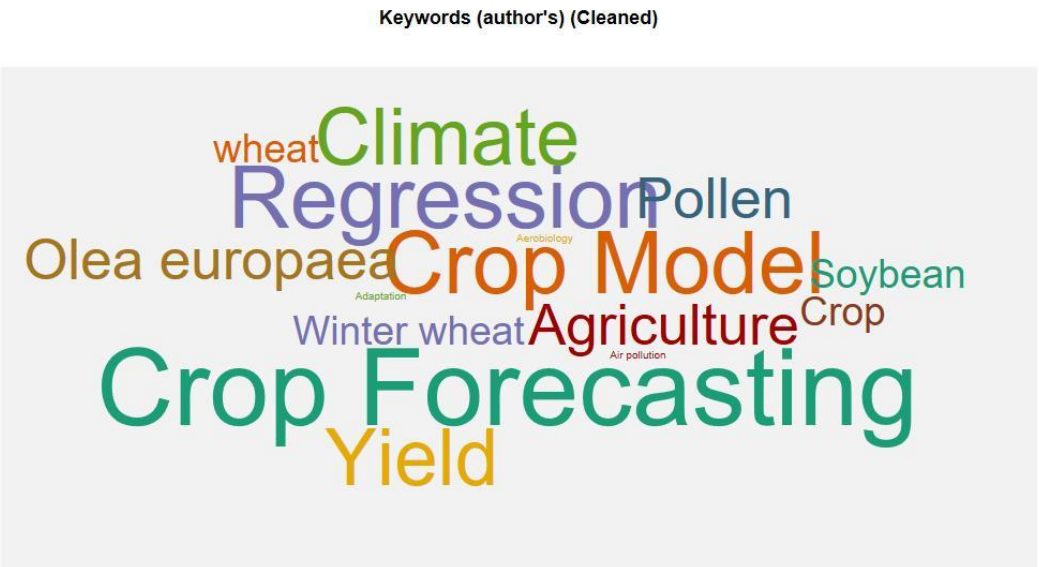

Figura 7. Nube de palabras clave. Adaptado de Vantage Point (2018).

En la gráfica de Aduna ilustrada en la figura 7, se observa el número de publicaciones realizadas producto de las investigaciones encontradas en la revisión, las cuales aparecen con mayor frecuencia palabras clave tales como: *Crop Forecasting(10), crop model(6), regresion(6), yield(5), climate(5), agriculture(3), wheat(2).*

A su vez, se observa que todas y cada una de las palabras clave están relacionadas entre sí. Situación que justifica nombres de trabajos como “Evaluation of the integrated Canadian crop yield forecaster model for in- season prediction of crop yield across the Canadian agricultural landscape” (Chipanshi et al., 2015)

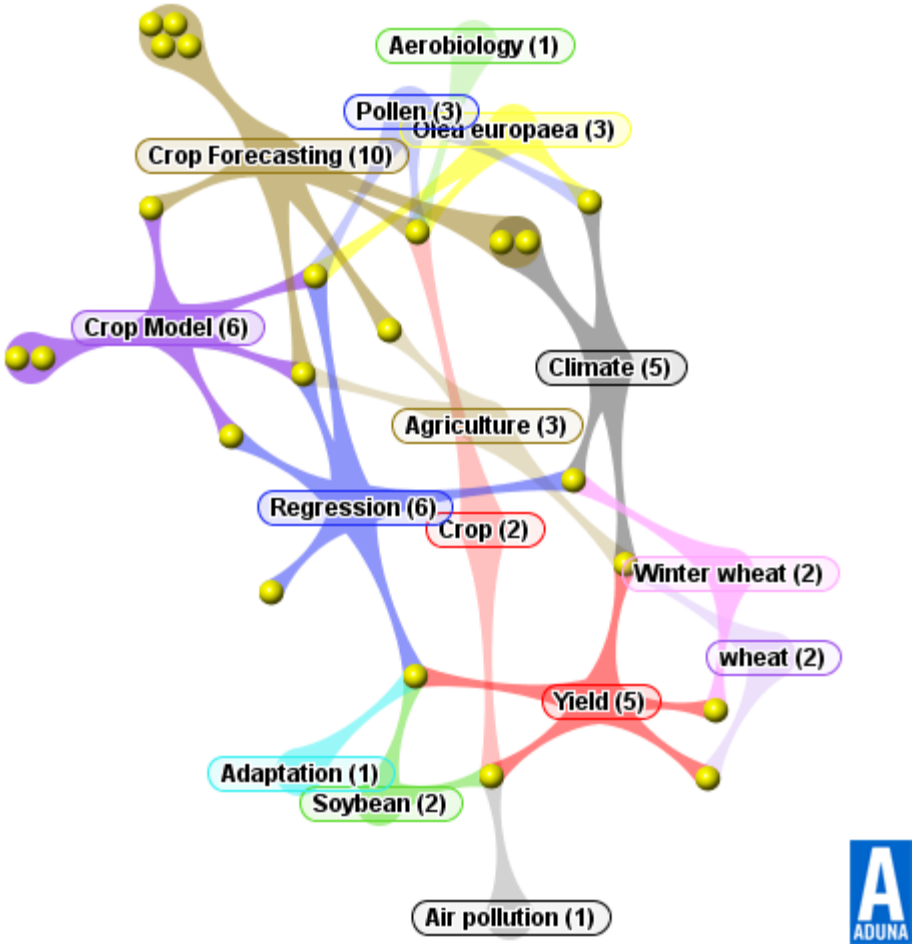

Figura 8. Aduna de palabras clave. Adaptada de Vantage Point (2018).

La figura 8, muestra la relación de los países con las áreas de investigación mayormente trabajadas por los mismos. En primera instancia, se observa que Estados Unidos, el país que realiza un mayor número de artículos publicados, se ha enfocado en investigar en mayor proporción el área de agricultura seguido por las áreas de meteorología y ciencias atmosféricas, ciencias ambientales y ecología, ciencia y tecnología, biotecnología y microbiología aplicada, investigación de operaciones y ciencia de gestión; comportamiento que sigue Italia en igual proporción.

Se observa que, por otro lado, países como India relacionan sus investigaciones en áreas de investigación de operaciones y ciencia de gestión aplicadas con agricultura y ciencias atmosféricas o Brasil que se enfoca exclusivamente en la temática de agricultura.

Países como China, España y Canadá comparten su enfoque de investigación en áreas como ciencias aplicadas agricultura y medio ambiente.

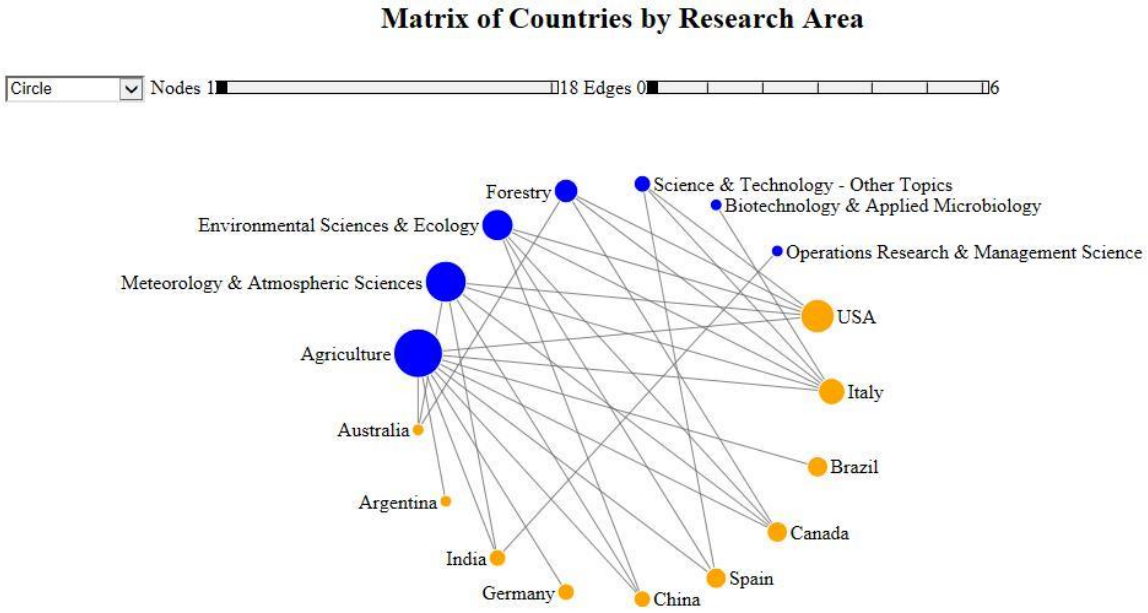

Figura 9. Matriz áreas de investigación por país. Adaptada de Vantage Point (2018).

2.2. Análisis preliminar de la literatura

En la búsqueda de literatura se observan diferentes tipos de cultivos objetos de estudio tales como maíz, trigo, uva, cebada, nuez, soja, arroz, aceituna, manzana, jilo, coco a partir de los cuales se pronostican tendencias en los rendimientos haciendo uso de modelos de aprendizaje supervisado.

Los autores Motha y Heddinghaus (1986), Stephens (1988), Walker (1989), Genovese y Terres (1999), ABARE (2004), coinciden en que “Las grandes fluctuaciones anuales en los rendimientos y la producción son motivo de gran preocupación para agricultores, políticos, entidades transportadoras. Es por esto que, para abordarlas, varios países han desarrollado métodos operativos para pronosticar los rendimientos de cultivos” (Hansen, Potgieter, & Tippett, 2004).

“La información anticipada sobre la producción probable y su distribución geográfica es útil para las agencias de manejo y comercialización que gestionan la logística de almacenamiento y

transporte y las ventas de exportación en el entorno de comercialización recientemente desregulado, y al gobierno en relación con las intervenciones de política" (Hammer et al., 2001) y que debido a los bajos valores de error "la confiabilidad y la capacidad de estos modelos de pronóstico justifican su uso para apoyar el proceso de toma de decisiones y mejorar la eficiencia agronómica y económica" (Cunha, Ribeiro, & Abreu, 2016).

De la primera investigación que arrojó la búsqueda, se observa que los investigadores Naylor, Falcon, Rochberg & Wada hicieron uso de modelos de regresión lineal para predecir la producción y las variables que más influyen en los rendimientos de cultivo de arroz, al igual que los autores Kar & Kumar. Ambas investigaciones se realizaron con el objetivo de "evaluar la productividad de la cosecha de arroz por adelantado utilizando datos de atributos meteorológicos y fisiológicos de la planta" (Kar & Kumar, 2014) y así "proporcionar una herramienta adicional para gestionar la seguridad alimentaria" (Naylor, Falcon, Rochberg, & Wada, 2001).

La creciente variabilidad climática ha venido afectando los cultivos de la región de Indonesia, especialmente para el de arroz el cual es el principal alimento básico para los habitantes de la región y una de las actividades que más genera ingresos y empleo. El estudio pretende mostrar el impacto del fenómeno de El Niño en los cultivos utilizando datos estacionales recolectados de 1971-1998 sobre cultivos plantados, arroz cosechado, producción y rendimientos de arroz. "Para medir el impacto del ENSO[1] en la producción de arroz, primero examinamos dos vínculos intermedios: las asociaciones entre el ENSO y la lluvia y entre la lluvia y la producción de arroz, y luego examinamos la asociación directa entre el ENSO y la producción de arroz. Los enlaces entre variables se cuantifican utilizando análisis de correlación y regresión"

El análisis demuestra que el 84%de la varianza en el área sembrada en septiembre-diciembre y el 81% de la varianza en el área sembrada en enero-abril se explica por las precipitaciones presentadas en esos meses, lo cual explica que la disminución en las precipitaciones de dicho año retrasa las plantaciones de arroz hasta que la lluvia sea adecuada para los cultivos. A su vez, las fechas de siembra tienen un impacto directo en la cosecha. Luego de la información suministrada,

[1] El Niño/Southern Oscillation (El fenómeno de El Niño).

los agricultores y formuladores de políticas podrán seleccionar los cultivos y la fecha de siembra para estabilizar cuestiones de demanda en los productos.

"Para abordar las preocupaciones de los mercados de productos básicos y los programas de alivio de la sequía, varios países exportadores de granos, incluida Australia, han desarrollado métodos operativos para pronosticar los rendimientos de cultivos regionales y la producción agregada" (ABARE, 2004). A partir de lo anterior, otra de las investigaciones derivadas de la preocupación por las temporadas de sequía producidas por el fenómeno de El Niño, fue la de los investigadores Hansen et al. (2004) quienes al mismo tiempo utilizaron un modelo de regresión lineal para predecir los rendimientos de trigo en función de los predictores climáticos estacionales con datos de precipitaciones tomados del distrito y estableciendo una distribución de probabilidad alrededor de cada pronóstico, generando como resultado una influencia significativa de las variables climáticas en los cultivos de trigo; cultivo que también fue objeto de investigación para Ceglar, Toreti, Lecerf, Van der Velde, & Dentener (2016), en el que mediante regresión de mínimos cuadrados parciales se identificaron las variables meteorológicas y el periodo en el cual estas tienen máxima influencia en los rendimientos de trigo durante su crecimiento en diferentes zonas de Francia. Para el desarrollo de este proyecto se utilizaron datos reportados en 92 regiones de Francia, en donde se presentaban cuatro tipos de clima principales (marítima, mediterránea, continental y montañosa), todo esto para lograr una comparación de las relaciones climáticas con el rendimiento y la variabilidad de los cultivos de trigo a partir de los diferentes tipos de clima.

Dentro de la búsqueda se observa un gran número de investigaciones en las que se utilizan los modelos de regresión lineal. Autores como Mkhabela, Mkhabela, & Mashinini (2005) y Nadler & Bullock (2011) enfocaron sus esfuerzos en pronosticar los rendimientos de cultivos de maíz con ayuda de estos modelos.

Los cultivos de maíz del sur de África, más específicamente de Swaziland, fueron objeto de estudio para los investigadores Mkhabela et al. (2005) donde el clima varía de templado con veranos calurosos a inviernos secos, lo que tiene como consecuencia inseguridad alimentaria y pérdida de ingresos debido a la reducción de los rendimientos, motivo por el cual se buscó "proporcionar estimaciones precisas y oportunas de la producción de maíz a los gobiernos y otras partes interesadas en la seguridad alimentaria para una intervención oportuna en caso de déficit"

(Mkhabela et al., 2005) utilizando elementos como NDVI (Índice de vegetación de diferencia normalizada) derivados del NOAA Advanced Very Resolution Radiometer (AVHRR). De la relación lineal positiva entre el rendimiento de maíz y el índice de vegetación de diferencia normalizada para 3 regiones (Middleveld, Lowveld y Lubombo), se concluye que las precipitaciones afectan directamente al cultivo. A su vez se puede observar que los factores climáticos no son los únicos que tiene incidencia en los cultivos, sino que además existen otros factores que determinan su rendimiento los cuales son de vital importancia involucrar para futuras investigaciones.

El trabajo de Nadler & Bullock (2011) se concentró en estudiar los cambios en las características climáticas para las praderas canadienses. Los datos de entrada para la investigación fueron registros entre 1921 y 2000 de temperatura máxima diaria, temperatura mínima diaria y cantidad diaria de precipitación, los cuales se adquirieron del Centro de Investigación de Cerealera y Oleaginosas del Este de Agricultura y Agroalimentación de Canadá.

Cinco de las siete regiones de estudio mostraron tendencias significativas en la temporada de crecimiento del cultivo, lo que indica que el calentamiento durante esas temporadas fue más pronunciado y determinó la variabilidad del cultivo.

Un descubrimiento importante derivado de las tendencias a largo plazo y resultados del modelo fue que, a lo largo de los años, debido a la variabilidad climática (Precipitaciones más variables y aumentos en la temperatura) hicieron que los cultivos se adaptaran a las condiciones y se generara una mayor variedad en los cultivos con mayor potencial de rendimiento.

Por otro lado, los cultivos de nogal (nuez) fueron objeto de estudio para los autores Lobell, Cahill, & Field (2007) que pretenden, en primer lugar, proporcionar una compresión cuantitativa de las relaciones entre el rendimiento de los cultivos y la incidencia de los cambios climáticos en estos últimos y en segundo lugar evaluar el impacto neto del clima en las tendencias de rendimiento observadas durante el periodo de estudio (1980-2003) analizando tres variables climáticas: temperatura mínima, temperatura máxima y precipitación las cuales se obtuvieron del Servicio Nacional de Estadísticas Agrícolas (NASS) del Departamento de Agricultura de los Estados Unidos (USDA). Esta vez para la región de California, en la cual la agricultura representa una actividad económica importante, motivo por el cual despierta el interés de estudiar las relaciones

de rendimiento en el clima las cuales pueden proporcionar una base para pronosticar la producción de cultivos dentro de un año para proyectar el impacto de los futuros cambios climáticos.

Para efectos de esta investigación, se ajusta una tendencia lineal para producir series de tiempo utilizando una regresión múltiple utilizando las 2 variables más importantes seleccionadas para cada cultivo como predictoras. Teniendo como resultado un alto valor de R^2 (Coeficiente de determinación), lo cual indica una estrecha relación entre las variables y los rendimientos de los cultivos.

Por otra parte según Tack, Barkley, & Nalley (2015) también coinciden en que existe una fuerte relación de los rendimientos en los cultivos y el clima. Para este caso en concreto, se cuantifica la relación entre las variables climáticas y los rendimientos de trigo los cuales evidencian un crecimiento rápido. Los avances más recientes han utilizado gran variedad de especificaciones que han presentado cambios en la forma de incluir los datos de temperatura en la ecuación de rendimiento estadístico.

Se utilizó análisis de regresión para analizar el efecto del clima en el rendimiento del trigo con la recopilación de un conjunto de datos que combina resultados de ensayos de campo de variedad de trigo de Kansas para en los periodos de 1985 al 2013.

Trabajos como los de Alvarez (2009), Mavromatis (2014) y Kouadio, Djaby, Duveiller, El Jarroudi, & Tychon (2012) coinciden en la realización de un modelo adecuado para estimar el rendimiento de trigo. Para el primero de ellos las variables analizadas fueron las características del suelo y los factores climáticos; utilizados como variables de entrada y tomados de registros meteorológicos para la provincia de Pampa, Argentina en el intervalo de tiempo de 1995-2004. Se utilizaron Redes Neuronales Artificiales y modelos lineales de regresión como metodologías para estimar el rendimiento del cultivo, el cual se correlacionó con el agua disponible en el suelo y el contenido de carbono orgánico. Comparando el modelo pronosticado con el observado y a partir de un RMSE[2] (0,05) se concluye que, en comparación con los modelos de regresión lineal, las Redes neuronales artificiales pudieron realizar una predicción más ajustada la cual explica el 64%

[2] Root-mean-square error: Error cuadrático medio.

de la varianza en el rendimiento de los cultivos obteniendo como resultado que el factor climático con mayor efecto sobre el rendimiento fue la precipitación. Mientras que para el segundo trabajo se utilizaron modelos de regresión lineal en relación a los procesos bióticos y abióticos en la situación de producción para los cultivos de trigo ubicados en Bélgica.

Los modelos lineales generalizados (GLM) fueron utilizados dentro de su metodología por los autores Park, Hwang & Vlek, los cuales coinciden dentro de su investigación en que "Los procesos de toma de decisiones en la agricultura a menudo requieren modelos fiables de respuesta de cultivos para evaluar el impacto de la gestión específica de la tierra" (Park, Hwang, & Vlek, 2005).

Tack et al. (2015) También utilizaron modelos de regresión lineal para analizar el efecto del tiempo en el rendimiento del trigo mediante datos específicos de ubicación. Para este caso se encuentran con limitaciones provenientes del conjunto de datos meteorológicos obtenidos dado que el clima en cada lugar de interés es muy variante. El proyecto presentado proporcionó importantes ideas para el mejoramiento del trigo y la toma de decisiones agrícolas relacionadas con el clima cambiante, también se ofrecen oportunidades para intensificar los esfuerzos de investigación para aumentar la resistencia al estrés por calor durante el desarrollo de cada una de las etapas de crecimiento.

La calidad del grano de trigo se ve directamente afectada por varios factores agroambientales y ambientales, es por esto que Toscano et al. (2014) tienen como objetivo determinar los principios generales que indican como en ambientes mediterráneos el Contenido de Proteína de Grano (GPC) se ve afectado por estos factores planteando un modelo de sistema con alta capacidad de predicción.

Inicialmente evaluaron la capacidad del sistema Delphi; modelo utilizado en el desarrollo del proyecto, para simular el GPC, en las principales cuencas de suministro italianas, también se analizaron las relaciones entre los errores Delphi y las variables durante las etapas de floración y llenado de grano.

Los resultados encontrados en dicho proyecto se evaluaron mediante regresión con GPC observada, mientras que los errores se calcularon realizando un análisis de correlación lineal con variables ambientales.

Los autores Oteros et al. (2014), Aguilera & Ruiz-Valenzuela (2014) y García-Mozo, Yaezel, Oteros, & Galán (2014) coinciden su estudios en los cultivos de aceituna utilizando diferentes modelos de regresión lineal como el de tipificación y modelado de regresión de mínimos cuadrados para el pronóstico de la producción de dicho cultivo. Algunos de los resultados que se obtuvieron en estos proyectos fueron que los índices más altos de polen y la disponibilidad de agua durante la primavera están relacionadas con un aumento en la producción, además, una disminución de la producción de la aceituna se relaciona con el aumento de la temperatura del aire durante el invierno y el verano.

De la búsqueda se encontraron trabajos en los que se realizan modelos para la predicción de cultivos, además de los anteriormente mencionados, de coco, soja, jilo, caña de azúcar, maní, manzanas abordados por los investigadores Toggweiler & Key (2001), Fishman et al. (2010), Rolim, Novo, Pantano, & Trani (2011), Pagani et al. (2017) y Moreto & Rolim (2015) Logan, McLeod, & Guikema (2016) para Sri Lanka, Estados Unidos, Brasil, España y Nueva Zelanda respectivamente. Los investigadores coinciden en utilizar modelos de regresión lineal para las predicciones de dichos cultivos encontrando que para el primer trabajo las precipitaciones son las condiciones climáticas que más influyen en el cultivo, en el segundo las concentraciones elevadas de ozono en el suelo y en el tercero las características de la planta (altura, numero de hojas) están relacionadas directamente con la producción del cultivo.

En la más reciente investigación, Cunha et al. (2016) implementaron modelos AHP y regresión paso a paso para modelar y ajustar pronósticos en la producción de los cultivos de uva en la región de Alentejo, Portugal utilizando datos del periodo de 1998 a 2014 de variables agronómicas y climáticas de la floración para estimar el potencial de producción y las variables que afectan al cultivo.

A partir de la revisión de literatura se puede observar que en la mayoría de los artículos las variables utilizadas son meteorológicas, es decir variables que describen el clima, por lo tanto, se evidencia que existe un fuerte interés en determinar qué tipo de relaciones hay entre el clima y los rendimientos del sector agrícola. Sin embargo, se observa que con el paso del tiempo las investigaciones involucran cada vez más variables (climáticas, no climáticas, morfología de la planta) y modelos para el pronóstico de rendimientos de los diferentes cultivos, aunque la mayoría

de los modelos utilizados por los investigadores son de regresión lineal, se percibe de las últimas investigaciones la implementación otro tipo de modelos (Máquinas de soporte vectorial, Redes neuronales…) los cuales permiten comparar sus resultados con el ánimo de elegir el modelo más acertado en el pronóstico de rendimientos y de esta manera aumentar su precisión.

3. Planteamiento del problema

Colombia, gracias a su ubicación geográfica; ubicada en la zona ecuatorial, con diversidad de pisos térmicos que van desde los 0 m.s.n.m [3] (>24ºC) hasta los 4.000 (<6ºC) Earthtrends (2011) siendo el primer país latinoamericano con mayores tasas de precipitación anuales, el décimo a nivel mundial y el cuarto país en América Latina con disponibilidad de tierras para la producción agrícola FAO (n.d.), lo convierten en un país con amplias alternativas para la producción agroindustrial. Así lo ratifican las cifras del DANE para el año 2017, las cuales indican que la agricultura fue la actividad económica que más le aportó al PIB con 4,9 puntos porcentuales DANE (n.d.).

Sin embargo, para el primer trimestre de 2018 "De las siete actividades que presentaron crecimiento por encima del promedio de la economía, la de primer lugar con 6,1% fueron las actividades financieras y de seguros. La agricultura se posicionó en octavo lugar, quedando por debajo del promedio del PIB (el cual fue de 2,2% para el primer trimestre de 2018) con una tasa de crecimiento de 2,0 puntos porcentuales *DANE* (2010).

De lo mencionado anteriormente se evidencia la necesidad de apoyar mediante métodos, modelos enfocados al crecimiento en estas áreas.

Además, debido a la situación de posconflicto por el que atraviesa el país un 60% de excombatientes de las FARC desean formarse en agricultura, según un estudio de la Universidad Nacional Semana-Agricultura (2017), lo cual podría ser posible gracias a las más de 40 millones

[3] Metros sobre el nivel del mar.

de hectáreas aptas para la siembra UPRA (n.d.) más específicamente a los cultivos de cacao, los cuales pueden jugar un papel clave en el desarrollo del posconflicto dado a que "el mapa del país donde hay más cacao y el mapa de las zonas más conflictivas es más o menos el mismo" tal y como lo afirmó el embajador de Estados Unidos en Colombia, Kevin Whitaker, en su visita al CIAT[4] en 2017 (CIAT, 2017).

El cultivo de cacao en Colombia, el cual se consideró como el cultivo de la paz, tiene un alto potencial especialmente por ser reconocido mundialmente como "Fino y de aroma en el mundo" según lo indica la International Cocoa Organization PROCOLOMBIA (n.d.). Situación que beneficia considerablemente al departamento de Santander, debido a que es la región con más áreas en hectáreas productoras del cultivo Ministerio de Agricultura (2014).

Para 2017, este cultivo alcanzó una producción de 60.535 toneladas dentro de las cuales 11.688 fueron para exportación. A pesar de las cifras alentadoras, "el sector sigue desarrollándose por debajo de su potencial" (CIAT & Universidad de Perdue, 2016). La falta de herramientas e investigaciones en el campo que ayuden a los agricultores del sector cacaotero a tomar decisiones acertadas y que en consecuencia puedan aumentar la productividad y competitividad, podrían ser algunas de las casusas que expliquen este fenómeno.

Aun cuando "Se evidencia la necesidad de aplicar un modelo de desarrollo económico sostenible soportado en la innovación y la tecnología" (Secretaría de agricultura, 2017), de la revisión de literatura y trabajos previos no se evidencian investigaciones que promuevan el mejoramiento de la agricultura que involucren nuevas tecnologías en el país. A diferencia de otros países como Indonesia, donde se encontraron investigaciones que datan de años atrás en los cuales usan "Las predicciones cuantitativas a partir de datos estadísticos previos de los efectos de factores climáticos en las cosechas para proporcionar herramientas en aras de gestionar la seguridad alimentaria" (Naylor et al., 2001).

[4] Centro Internacional de Agricultura Tropical.

La falta de uso de herramientas que utilizan datos dentro de sus modelos podría ser explicada a partir del análisis que realizó el DNP[5] del gobierno de Colombia en la "Política de explotación de datos: Big Data" (Mejía, 2018), el cual presenta un panorama de 4 factores: datos digitales, cultura de datos, capital humano para la explotación de datos y valor social y económico de los datos. Entre otros, uno de los grandes problemas es que tan solo el 37% de las entidades, de una muestra de 150 entidades colombianas, utilizan datos para la predicción (DNP, 2017).

Otra de las situaciones encontradas es que solo el 9% de las 150 entidades tiene al menos 1 proyecto de explotación de datos en los que involucra el uso de algoritmos (DNP, 2017).

Uno de los indicadores que espera conseguir el DNP para el 2020 es que el 90% de las entidades públicas tengan al menos 1 proyecto de aprovechamiento de datos que involucre el uso de algoritmos.

En aras de contribuir con los escenarios presentados anteriormente, se tiene el objetivo de desarrollar, comparar y aplicar modelos que más se ajusten a los rendimientos observados para la predicción del rendimiento de cultivos de cacao, mediante herramientas del Aprendizaje Automático, con el fin de beneficiar el crecimiento y mejoramiento de dichos cultivos.

[5] Departamento Nacional de Planeación.

4. Objetivos

4.1. Objetivo general

Plantear un modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado.

4.2. Objetivos específicos

- Realizar una revisión de literatura sobre la aplicación de los modelos de regresión generalizados y máquinas de soporte vectorial para la predicción de rendimientos agrícolas.
- Aplicar modelos de regresión generalizados y máquinas de soporte vectorial para la predicción de rendimiento agrícola en cultivos de cacao.
- Validar los modelos con métricas de ajuste para determinar la alternativa que mejor represente los rendimientos agrícolas en cultivos de cacao.
- Elaborar un artículo publicable resumiendo los hallazgos encontrados en el proyecto.

5. Resultados esperados

Los resultados esperados al culminar el presente proyecto, en relación con los objetivos planteados inicialmente, se presentan a continuación.

- Dos modelos predictivos para el rendimiento de cultivos de cacao
- Modelo que más se ajuste a los rendimientos de cacao
- Documento final con las actividades desarrolladas en la ejecución del proyecto
- Artículo científico de carácter publicable

6. Marco de referencia

6.1. Marco de antecedentes

A pesar de la importancia que tiene involucrar nuevas tecnologías para el mejoramiento de los cultivos en Colombia presentadas anteriormente, la búsqueda de trabajos dentro de la Universidad Industrial de Santander que utilicen herramientas de Aprendizaje Automático para la predicción agrícola no arrojó ningún resultado.

Las herramientas de aprendizaje automático han sido aplicadas para la predicción de diferentes cultivos agrícolas en trabajos como:

"Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions" (Zheng, Chen, Han, Zhao, & Ma, 2009). Dentro de sus objetivos están: comparar la importancia de las propiedades del suelo y las prácticas de manejo para determinar la variabilidad del cultivo e identificar las variables de mayor influencia en el rendimiento bajo condiciones de sequía de los cultivos de soja ubicados en el noreste de China, donde la sequía a menudo causa un bajo rendimiento del grano.

El aprendizaje automático supervisado es el protagonista en esta investigación, por lo que para llevarse a cabo es necesaria una gran cantidad de datos de entrada para predecir la variable respuesta.

Para la recolección de datos, se tomaron muestras aleatorias de la capa superior del suelo y se midieron variables como: pH del suelo, conductividad eléctrica del suelo (CE), carbono orgánico del suelo (SOC), nitrógeno del suelo, fosforo y potasio disponible (TK). Para variables de prácticas de manejo agrícola, se desarrolló un cuestionario a todos los cultivadores del grano en la zona sobre prácticas detalladas de gestión del campo como métodos de la preparación del suelo, fecha de siembra, entre otros. Además, se recopiló información sobre las variables socioeconómicas, como la edad y el nivel de educación de los jefes de hogar seleccionados, el área de tierras de cultivo y el ingreso medio del hogar.

Acto seguido, se utilizó SYSTAT 12 (Systat Software, San José, CA) para tratar los datos de entrada y analizarlos estadísticamente. La importancia relativa de las variables prácticas de manejo y parámetros del suelo se analizaron con modelos lineales generalizados (GLM) y árboles de clasificación de regresión (CART).

En primera instancia, los parámetros del suelo medidos se utilizaron como variables independientes con ayuda de la regresión paso a paso y los arboles de regresión para descubrir las relaciones e interacciones entre los rendimientos del cultivo y las variables agronómicas.

Los resultados obtenidos por los modelos muestran una variabilidad del rendimiento en gran medida por las variaciones tanto del suelo como del manejo de los cultivos; soportadas en primera instancia con el modelo CART por una relación de 61% entre la variabilidad del rendimiento del cultivo y las condiciones de este último; en mayor medida con un 47% la conductividad eléctrica del suelo tuvo mayor porcentaje de incidencia en la variabilidad del cultivo. Del resultado se concluye que el fosforo y el nitrógeno son los elementos que más limitan el crecimiento y la producción de la planta. A diferencia de la regresión lineal, la cual indicó que solo el 24,2% de la variabilidad del rendimiento podría explicarse por los parámetros del suelo medidos a pesar de los coeficientes de regresión para el modelo.

De los resultados de la investigación se llega a que, los modelos CART muestran una precisión de predicción mayores a los del modelo lineal generalizado. Sin embargo, se analiza que los resultados de este último fueron consistentes y pudieron predecir la variabilidad del cultivo en función de las variables de entrada.

Finalmente, a partir de los hallazgos mencionados anteriormente, se pudo indicar que los sistemas de cultivo se pueden adaptar a las condiciones de sequía alterando las opciones de manejo. A pesar de ello, a partir de los datos de las encuestas utilizados como variables de entrada, se evidenció que el uso de fertilizantes por parte de los agricultores no eran los adecuados respecto a las características de los suelos. La información que proporcionaron los modelos es de gran utilidad para capacitar a los agricultores en la selección de fertilizantes adecuados que pudiesen mejorar las condiciones del suelo y así obtener mejores rendimientos.

Por otro lado, los investigadores Chen, Wu, & Liu (2016) en su trabajo "Assessing the relative importance of climate variables to rice yield variation using support vector machines" pretenden comparar el rendimiento predictivo de los modelos de Máquinas de Soporte Vectorial (SVM) con los de Redes Neuronales Artificiales (ANN) con variables de entrada derivados de datos estadísticos del rendimiento del cultivo en el periodo de 1965-2012, tomados del China Meteorological Administration (CMA). Durante la investigación se observa en primera instancia que una sequía extrema en el 2006 produjo una disminución del rendimiento del cultivo. Los valores de correlación positiva indican la estrecha relación de factores climáticos y del suelo tales como temperatura media, humedad relativa, horas de sol, días de lluvia para los cultivos de arroz en el sudoeste de China encontrando que los factores climáticos del que más al que menos influyen en la variabilidad de los cultivos están en orden de: horas de sol, rango de temperatura diaria, lluvia, humedad relativa, temperatura media y días lluviosos.

Los valores de Error absoluto medio (MAE), error absoluto relativo (MRAE), error cuadrático medio (RMSE), error relativo del cuadrado medio de la raíz (RRMSE) para los modelos de SVM fueron de 0.39, 5.97, 0.47 y 7.19 respectivamente y para los modelos de ANN los valores fueron 0.49, 7.29, 0.57, 8.71 respectivamente lo cual indica que, en promedio, el modelo SVM supera el método de ANN debido a los bajos valores de los diferentes tipos de error. Además, el alto coeficiente de determinación del modelo SVM en comparación al de ANN de 0,56 y 0,20

respectivamente, ratifican la alta precisión de predicción generada por las SVM la cual puede ser útil para investigar la importancia relativa de las variables climáticas que influyen en los cultivos.

La más reciente investigación encontrada es de los autores Chattopadhyay & Mitra (2018) los cuales derivan su estudio de la necesidad de predecir la producción de granos alimenticios en el sur de Asia, debido a que el aumento de las temperaturas en los últimos años ha ocasionado efectos negativos en los cultivos de la región. El objetivo de dicha predicción es encontrar las influencias de los factores climáticos para facilitar a que científicos, agricultores, legisladores, empresarios y al gobierno puedan formular estrategias adecuadas para hacerle frente a la influencia de la variabilidad climática en la producción de alimentos.

La investigación se lleva a cabo mediante el uso de datos estadísticos tomados del Departamento Meteorológico de la India y el Banco de Reserva de la India del rendimiento de los cultivos con variables climáticas tales como: lluvias estacionales, temperaturas estacionales y área de producción de los granos. Cabe resaltar que el conjunto de datos tiene 65 muestras de las cuales 45 observaciones se utilizaron para el desarrollo del modelo y 20 para probar la generalización de este.

El estudio que se llevó a cabo en la investigación se realizó con ayuda de 3 tipos de modelos: modelo lineal generalizado (GLM) y dos modelos no lineales; regresión adaptativa multivariada y modelo aditivo generalizado los cuales se evaluaron en función de su rendimiento en la predicción de productividad del cultivo Chattopadhyay & Mitra (2018).

Al utilizar medidas de bondad de ajuste para evaluar la predicción y el rendimiento de los modelos, se obtuvo que el modelo de regresión adaptativa arrojó resultados más precisos soportado por el valor de error cuadrático medio (RMSE) de 9,89. Sin embargo, el modelo lineal GLM (con un valor de RMSE de 12,99) predijo bastante bien y se concluye que es un modelo con buen ajuste para predecir el rendimiento del cultivo.

De los modelos aplicados, se encontró que la variable con mayor influencia en el cultivo fue la de lluvias durante junio, julio, agosto y septiembre y la variable que menos influencia tuvo en el cultivo fue la de lluvias durante marzo, abril y mayo.

De lo anterior, se concluye que predecir el rendimiento para los cultivos alimenticios utilizando variables tanto climáticas como no climáticas, es de gran ayuda para que agricultores, empresas y gobierno puedan comprender las condiciones más influyentes y en consecuencia mejorar la adaptabilidad de los cultivos tomando decisiones acertadas anticipadamente con el objetivo de obtener mayor rendimiento en los cultivos.

6.2. Marco teórico

6.2.1. Predicción.

"La predicción numérica es una estimación del valor de una variable continua y ordenada a partir de un modelo utilizando un conjunto de entrenamiento" (Berzal, n.d.).

Para la predicción es necesario otorgar datos históricos de entrada a un modelo diseñado para transformar dichos datos en variables de salida que den idea de un evento futuro. Se debe tener en cuenta que para la evaluación y construcción de un modelo son necesarios una serie de pasos (Cayuela, 2010) los cuales se detallan a continuación:

Figura 10. Modelos lineales generalizados. Adaptado de Modelos lineales generalizados (GLM) Cayuela (2010).

6.2.2. Predicción del rendimiento cultivos. El rendimiento, al verse afectado por muchos factores interrelacionados: riegos, fertilizantes, rotación de cultivos, morfología de la planta, condiciones de la tierra, factores climáticos entre otros, es de gran interés en los planificadores agrícolas estimar una predicción para el rendimiento de los cultivos, los cuales ayuden a las diferentes entidades (agricultores, empresarios, gobierno) a tomar decisiones acertadas a partir de las tendencias y variables que más afecten el cultivo.

A lo largo de los años se han usado estimadores simples, como el promedio de rendimientos o el ultimo rendimiento obtenido. Sin embargo, estos últimos no han sido de alta precisión. "Por lo tanto se han desarrollado métodos más eficientes que se pueden clasificar como modelos basados en datos (…). Las técnicas Aprendizaje automático se basan en estructuras no paramétricas y semiparamétricas, y la validación se basa en la precisión de predicción (Breiman, 2001). Trabajos previos sugieren que los modelos basados en datos tienen una mejor adaptabilidad para la planificación de cultivos debido a su implementación y rendimiento" (Gonzalez-Sanchez, 2014).

6.2.3. Aprendizaje automático. El aprendizaje automático o "Machine Learning es una rama de la inteligencia artificial que proporciona métodos con la capacidad de aprender o hacer predicciones sobre datos. Estos métodos crean un modelo a partir de entradas de ejemplo para hacer predicciones o tomar decisiones" (Mitchell, 1997).

> "ML no hace suposiciones sobre la estructura correcta del modelo de datos, lo que permite la construcción de modelos complejos" (Díaz, Mazza, Combarro, Gimenez & Gaid, 2017).

Los algoritmos de aprendizaje automático encuentran patrones naturales en los datos que generan información y ayudan a tomar mejores decisiones y predicciones (Baratta, 2016). En la literatura, a menudo se llaman predictores a las variables independientes o bien a las entradas y a las salidas o variables dependientes se les llama variables respuesta.

A continuación, se visualiza un esquema de clasificación para el aprendizaje automático.

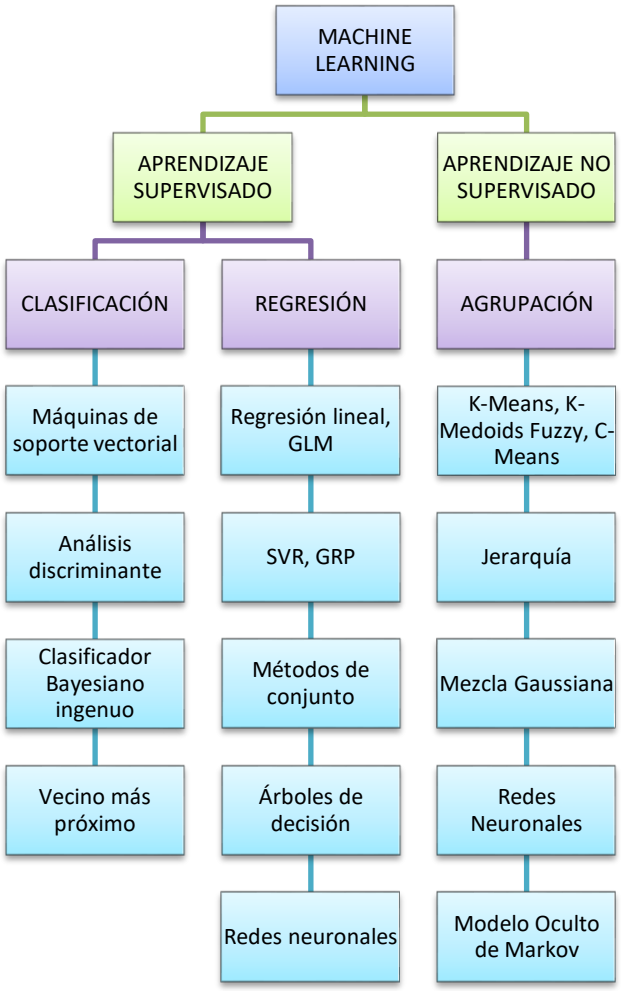

Figura 11. Métodos de aprendizaje automático. Adaptado de Introducing Machine Learning Math Works (2016). https://www.mathworks.com/tagteam/89703_92991v00_machine_learning_section1_ebook_v12.pdf

6.2.3.1. Aprendizaje no supervisado. En la categoría de reconocimiento de patrones, existe el tipo de problema no supervisado (también llamado aprendizaje no supervisado), el problema es descubrir la estructura del conjunto de datos si los hay. Esto generalmente significa que el usuario desea saber si hay grupos en los datos, y qué características hacen que los objetos sean similares dentro de un grupo. La elección de un algoritmo es una cuestión de preferencia del diseñador. Diferentes algoritmos pueden presentar diferentes estructuras para el mismo conjunto de datos. Una característica de este tipo de aprendizaje es que no hay ninguna verdad fundamental contra la

cual comparar los resultados. La única indicación de qué tan bueno es el resultado es la estimación subjetiva del usuario (Kuncheva, 2004).

6.2.4. Aprendizaje supervisado. Gracias a los avances en la tecnología informática, se tiene la capacidad de almacenar y procesar grandes cantidades de datos lo cual es posible gracias al aprendizaje automático (Baratta, 2016).

Se le llama al aprendizaje supervisado debido a la presencia de una variable resultado para guiar el proceso de aprendizaje. A diferencia del no supervisado, donde solo se observan las características y no se tiene en cuenta las mediciones del resultado.

Cada objeto en el conjunto de datos viene con una etiqueta de clase preasignada. La labor en este tipo de aprendizaje es entrenar a un clasificador para que haga el etiquetado. Con mucha frecuencia, el proceso de etiquetado no se puede describir en una forma algorítmica. Entonces se le proporciona a la máquina las habilidades de aprendizaje y a su vez se le presentan los datos etiquetados (Kuncheva, 2004).

El aprendizaje supervisado usa técnicas de clasificación y regresión para desarrollar modelos predictivos (*Introducing Machine Learning*, n.d.) (Baratta, 2016).

- Clasificación: Esta técnica clasifica los datos de entrada en categorías. Las técnicas de clasificación predicen respuestas discretas.
- Regresión: Las técnicas de regresión predicen respuestas continuas. Es decir, para predecir valores numéricos desconocidos.

6.2.4.1. Máquinas de soporte vectorial. Es una técnica de clasificación que hace parte de la rama de aprendizaje supervisado del Aprendizaje automático basada en la idea de minimización de riego estructural (SRM).

Una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de

características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento (Betancour, 2005).

La base biológica de las máquinas de soporte (SVM) es una clasificación que permite maximizar el espacio en blanco en ambos lados del hiperplano para garantizar la precisión de la clasificación. La regresión de vectores de soporte (SVR) es la extensión de SVM cuando se aplican para tratar problemas de regresión (Vapnik et al., 1997; Basak et al., 2007). La función lineal f (x) se puede expresar como:

$$f(x) = \langle \omega, x \rangle + h$$

Donde $\langle , \rangle$ indica el producto escalar, ω es un vector de peso ajustable, x es la información de entrada y h es el umbral escalar. La función de pérdida $L^{\varepsilon}(y_{j,} fx_j)$), introducida en SVR, describe un modelo que indica que no existe diferencia entre los valores actuales y predictivos si el valor de diferencia entre ellos es menos, que es la principal diferencia entre las funciones de regresión lineal (Yoon et al. 2011). Con la condición de permitir la existencia de un error de ajuste, se pueden obtener minimizando la fusión de riesgo $M_{\omega,\zeta,\zeta*}$ con las variables positivas de holgura (ζ, ζ*) introducidas en ella.

$$M_{\omega,\zeta,\zeta*} = \frac{1}{2} ||\omega||^2 + C \sum_{j=1}^{1} (\zeta_j, \zeta_j *)$$

$$\begin{cases} y_j - \langle \omega, x \rangle - h \leq \varepsilon + \zeta_j \\ \langle \omega, x \rangle + h - y_j \leq \varepsilon + \zeta_j * \\ \zeta_j, \zeta_j * \geq 0 \end{cases} \qquad j = 1, \ldots, n$$

Donde C es un factor de penalización, que controla el compromiso entre el error de capacitación y la complejidad del modelo, conciliando el riesgo empírico y el riesgo de confianza ζ, ζ* son variables de holgura que calculan el error de los lados hacia arriba y hacia abajo, respectivamente.

Los multiplicadores de Lagrange ($\alpha_j - \alpha_j *$) se introducen para resolver el problema dual. Se puede obtener el mejor hiperplano de regresión.

$$f(x) = \sum_{j=1}^{1} (\alpha_j - \alpha_j *) \langle x, x_j \rangle + h$$

donde $\alpha_j \, y \, \alpha_j *$ satisfacen la igualdad de $\alpha_j \times \alpha_j * = 0, \alpha_j \geq 0, \alpha_j * \geq 0,$ y j = 1, … n y se pueden determinar maximizando la forma dual.

La clave para mejorar la precisión de predicción es la función SVM kernel y la optimización de parámetros.

Las funciones que satisfacen el teorema de Mercer pueden ser usadas como productos punto y por ende como kernels. A continuación, un kernel polinomial de grado d:

$$K(x_i x_j) = (1 + x_i \bullet x_j)^d$$

Para construir un clasificador SVM.

6.2.4.2. Regresión lineal generalizada, GLM. Los modelos lineales generalizados (GLM de las siglas en inglés de Generalized Linear Models) son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc.) y varianzas no constantes. Estos modelos son una alternativa a la transformación de la variable respuesta y a la falta de normalidad en el modelo.

Ciertos tipos de variables respuesta sufren invariablemente la violación de estos dos supuestos de los modelos normales y los GLM ofrecen una buena alternativa para tratarlos (Cayuela, 2010).

El supuesto central que se ha hecho hasta el momento con los modelos lineales es que la varianza es constante (Figura 1a). En el caso de los conteos, sin embargo, donde la variable respuesta está expresada en números enteros y en dónde hay a menudo muchos ceros en los datos, la varianza podría incrementar linealmente con la media (Figura 1b). Con proporciones, donde hay un conteo del número de fallos de un evento, así como del número de éxitos, la varianza tendrá una forma de U invertida en relación a la media (Figura 1c). Cuando la variable respuesta siga una

distribución Gamma, entonces la varianza incrementa de una manera no lineal con la media (Figura 1d).

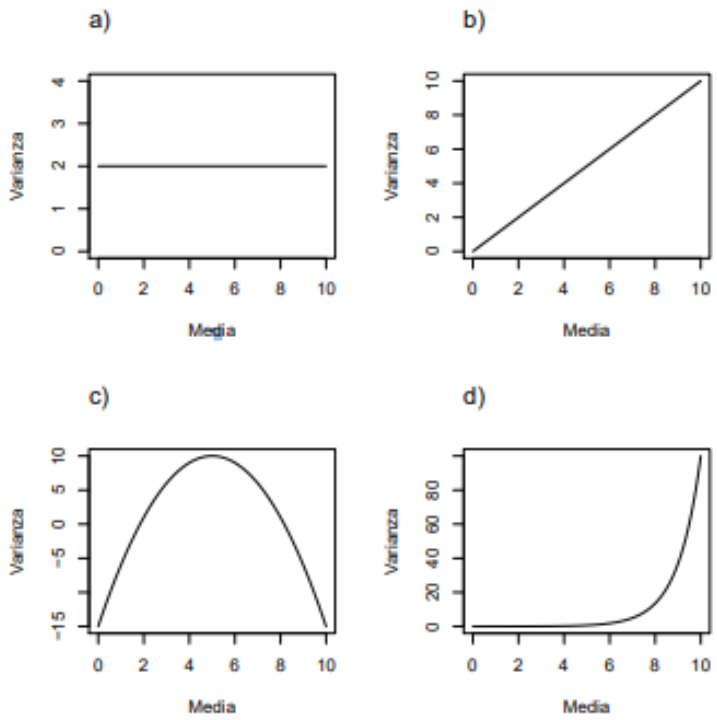

Figura 12. Relación entre la media y la varianza de los datos bajo distintos supuestos. Adaptado de Modelos lineales generalizados (GLM) Cayuela (2010).

6.2.5. Métricas de ajuste. "Para la misma tarea, puede haber varios algoritmos y podemos estar interesados en encontrar el más eficiente" (Cayuela, 2010), motivo por el cual se hacen necesarias las métricas de ajuste aplicadas al modelo con el fin de validar y escoger el que más se ajuste a los datos.

La evaluación de modelos se realiza por medio de unas medidas de error, las fórmulas se expresan como:

Error cuadrático medio relativo (RRMSE):

$$RRMSE = \frac{1}{A}\sqrt{\sum_{i=1}^{N}\frac{(A_i - P_i)^2}{N}}$$

Error absoluto medio (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | A_i - P_i |$$

Error absoluto relativo medio (MRAE):

$$MRAE = \frac{1}{n}\sum_{i=1}^{n} | \frac{A_i - P_i}{A_i} |$$

Coeficiente de masa residual (CRM):

$$CRM = \frac{\sum_{i=1}^{n} A_i - \sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} A_i}$$

Coeficiente de determinación (R^2):

$$R^2 = \frac{\sum_{i=1}^{N}(A_i - A)^2(P_i - P)^2}{\sum_{i=1}^{N}(A_i - A)^2 \bullet \sum_{i=1}^{N}(P_i - P)^2}$$

Donde $A_i \; y \; P_i$ son valores medios y pronosticados de los i-esimo, respectivamente.

A y P son los valores medios del rendimiento observado y del rendimiento previsto, respectivamente.

N es el número de datos de la validación.

Importancia relativa (RI):

$$RI = \frac{RMSE_R - RMSE_E}{RMSE_R}$$

El $RMSE_R$ y el $RMSE_E$ son los errores cuadráticos medios para el mejor rendimiento y los métodos de referencia, respectivamente.

6.3. Conjunto de datos

El conjunto de datos es el insumo fundamental de todo análisis predictivo, el cual es necesario para encontrar la variable respuesta, puesto que es la fuente que contiene la información de variables y su relación entre ellas. Para efectos de la presente investigación, el conjunto de datos de entrada

que se emplearán para el desarrollo y validación del modelo fueron suministrados por AGROSAVIA (Corporación Colombiana de Investigación Agropecuaria) los cuales fueron tomados a partir de un cultivo experimental ubicado en el centro de investigación La Suiza para los años 2015, 2016 y 2017.

El cultivo experimental de cacao está compuesto por 3 factores: fertilización, clon y exposición. El primero de ellos cuenta con 3 niveles: fertilización al 50%, 100%, 150%. En segundo lugar, se tienen los 10 tipos de clones más representativos de Santander clasificados en 5 regionales (SCC-19, SCC-52, SCC-61, SCC-64 Y SCC-83) y 5 universales (ICS-95, CNN-51, ETT-8, TSH565 y ICS-1). Y por último, la exposición del cultivo es a sol o a sombra. Considerando estos factores y niveles se cuenta con un total de 60 tratamientos.

A cada uno de los tratamientos mencionados anteriormente, le corresponden 15 plantas, por lo que se tiene un total de 900 de ellas a las cuales se les miden sus características fotosintéticas, morfológicas, físicas y químicas del suelo. De igual manera, el cacao producido por estas plantas se evalúa en términos de altura de la planta, número de ramas, diámetro del tronco, peso de la almendra, peso de la mazorca completa, entre otros.

En primer lugar, de las características fotosintéticas fueron tomadas 3 muestras, 1 vez al año durante el periodo de 2015-2017 para los 60 tratamientos, para un total de 540 observaciones. Por otro lado, se cuenta con 2160 observaciones para las características morfológicas de la planta tomadas una vez por semestre, realizando 3 repeticiones para 2 árboles por repetición para los 60 tratamientos. A su vez, se cuenta con 7239 observaciones de condiciones ambientales del cultivo experimental, las cuales fueron medidas con ayuda de sensores que realizaron recuentos para el periodo comprendido entre 2015 y 2017.

En el apéndice C, se presenta la descripción de las variables independientes clasificadas dentro de las características mencionadas anteriormente.

7. Metodología

Para dar cumplimiento a los objetivos anteriormente planteados, se llevan a cabo las siguientes etapas:

7.1. Primera Fase: Revisión de literatura (Objetivo 1)

- Definir las bases de datos en las cuales se realizará la búsqueda.
- Identificar palabras clave y términos relacionados con la aplicación de herramientas de aprendizaje automático en el sector agrícola.
- Definir la ecuación de búsqueda para las bases de datos seleccionadas.
- Realizar un análisis bibliométrico para establecer el estado actual del tema de investigación.
- Realizar una revisión sobre la aplicación de modelos de regresión generalizada y máquinas de soporte vectorial en el sector agrícola.

7.2. Segunda Fase: Aplicación de modelos (Objetivo 2)

- Levantamiento de información en bases de datos sobre cultivos de cacao en Santander.
- Programación de modelos de regresión generalizada y máquinas de soporte vectorial utilizando el lenguaje de programación Python.

7.3. Tercera Fase: Validación de modelos (Objetivo 3)

- Definir las métricas de ajuste a utilizar para probar la bondad de ajuste de los modelos.
- Evaluar los modelos con las respectivas métricas.
- Establecer las variables con mayor influencia en la predicción de los rendimientos de cacao.
- Comparar y seleccionar el modelo con el mejor ajuste a los rendimientos de cacao observados.
- Validar las variables identificadas en los modelos con expertos en cultivos de cacao.

7.4. Cuarta Fase. Documentación del trabajo (Objetivo 4)

- Elaboración de un documento final resultado de las actividades desarrolladas en la ejecución del proyecto.
- Redacción de un artículo científico de carácter publicable.

8. Estructura del proyecto

Introducción

1. Definición del proyecto

1.1. Planteamiento del problema

1.2. Título del proyecto

1.3. Modalidad

1.4. Responsables

1.5. Nombre del grupo de investigación

1.6. Objetivos

1.6.1. Objetivo general

1.6.2. Objetivos específicos

2. Revisión de literatura

2.1. Análisis bibliométrico

2.2. Análisis preliminar de literatura

2.3. Selección de técnicas para aplicación de modelos

3. Metodología

3.1. Primera Fase: Revisión de literatura

3.2. Aplicación de modelos

3.3. Validación de modelos

3.4. Documentación del trabajo

4. Marco teórico

4.1. Predicción

9. Cronograma

Actividad	Abril				Mayo				Junio				Julio				Agosto				Septiembre				Octubre				Noviembre				Diciembre				Enero				Febrero			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Aprobación ficha																																											
Primera fase: Revisión de literatura																																											
Definir las bases de datos en las cuales se realizará la busqueda																																											
Identificar palabras clave																																											
Definir la ecuación de búsqueda																																											
Realizar análisis bibliométrico																																											
Elaboración del plan																																											
Revisión y corrección del plan																																											
Capacitación Python																																											
Entrega del plan																																											
Segunda fase: Aplicación de modelos																																											
Levantamiento de información																																											
Limpieza de datos																																											
Identificar modelos																																											
Programación de modelos																																											
Tercera fase: Validación de modelos																																											
Definir métricas de ajuste																																											
Evaluar modelos																																											
Analizar modelos																																											
Comparar y seleccionar el mejor modelo																																											
Cuarta fase: Documentación del trabajo																																											
Elaboración del artículo																																											
Elaboración del libro																																											
Revisión y corrección del libro																																											
Entrega del libro																																											

Figura 13. Cronograma.

10. Presupuesto del trabajo de grado

RECURSO	RUBRO	DESCRIPCIÓN	ESTUDIANTE	UIS
RECURSO HUMANO	Dirección del proyecto	Tiempo dedicado a la dirección del proyecto		X
	Codirección del proyecto	Tiempo del codirector dedicado al proyecto		X
	Autores del proyecto	Tiempo de los autores para el desarrollo del proyecto	X	
PAPELERIA E INSUMOS	Equipos de computo	Recurso electrónico empleado en el desarrollo de la investigación.	X	X
	Impresora	Recurso necesario para la impresión de documentos fisicos inherentes al desarrollo del proyecto.	X	
	Papeleria	Recurso de uso para reuniones con director y codirector.	X	
	Internet	Recurso empleado para llevar a cabo la recolección de información, y acceso a bases de datos.	X	
	Bases de datos	Acceso a la bases de datos		X
OTROS RECUROS	Capacitación en el lenguaje estadístico de programación Python	Capacitación para adquirir conocimiento para el manejo del lenguaje Python		X

Figura 14. Presupuesto del Proyecto de Grado.

Referencias bibliográficas

Aguilera, F., & Ruiz-Valenzuela, L. (2014). Forecasting olive crop yields based on long-term aerobiological data series and bioclimatic conditions for the southern Iberian Peninsula. *Spanish Journal of Agricultural Research*, *12*(1), 215–224. https://doi.org/10.5424/sjar/2014121-4532

Alvarez, R. (2009). Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy*, *30*(2), 70–77. https://doi.org/10.1016/j.eja.2008.07.005

Baratta, A. (2016). Introducción a Machine Learning. *SUNQU*, 197–224.

Berzal, F. (n.d.). *Clasificación y predicción*. Retrieved from http://elvex.ugr.es/idbis/dm/slides/3 Classification.pdf

Betancour, G. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia Et Technica*, (27), 67–72. https://doi.org/10.22517/23447214.6895

Cayuela, L. (2010). *Modelos lineales generalizados (GLM)*. Retrieved from https://s3.amazonaws.com/academia.edu.documents/33538949/3-Modelos_lineales_generalizados.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1534283770&Signature=SwPRBT18Y6cj24fcheD0xBQUxvE%3D&response-content-disposition=inline%3B filename%3DModelos_lineales_generalizados_GLM.pdf

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., & Dentener, F. (2016). Impact of meteorological drivers on regional inter-annual crop yield variability in France. *Agricultural and Forest Meteorology*, *216*, 58–67. https://doi.org/10.1016/j.agrformet.2015.10.004

Chattopadhyay, M., & Mitra, S. K. (2018). Assessing the predictability of different kinds of models in estimating impacts of climatic factors on food grain availability in India. *Opsearch*, *55*(1), 50–64. https://doi.org/10.1007/s12597-017-0314-9

Chen, H., Wu, W., & Liu, H.-B. (2016). Assessing the relative importance of climate variables to rice yield variation using support vector machines. *Theoretical and Applied Climatology*, *126*(1–2), 105–111. https://doi.org/10.1007/s00704-015-1559-y

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., … 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. *Agricultural and Forest Meteorology*, *206*, 137–150. https://doi.org/10.1016/j.agrformet.2015.03.007

CIAT. (2017). Retrieved August 13, 2018, from https://blog.ciat.cgiar.org/es/que-papel-puede-jugar-el-cacao-para-la-paz-en-colombia/

Cunha, M., Ribeiro, H., & Abreu, I. (2016). Pollen-based predictive modelling of wine production: application to an arid region. *European Journal of Agronomy*, *73*, 42–54. https://doi.org/10.1016/j.eja.2015.10.008

DANE - PIB 2017. (n.d.). Retrieved August 9, 2018, from http://www.dane.gov.co/index.php/52-espanol/noticias/noticias/4505-pib-oferta-iv-trimestre-2017

DANE - PIB 2018. (2010). Retrieved from http://www.dane.gov.co/files/investigaciones/boletines/pib/bol_PIB_Itrim18_producion_y_gasto.pdf

FAO Marco programático Colombia (2015-2019). (n.d.). Retrieved from http://www.fao.org/3/a-bp556s.pdf

FEDECACAO. (2018). Retrieved August 9, 2018, from http://www.fedecacao.com.co/portal/index.php/es/2015-04-23-20-00-33/551-en-2017-colombia-alcanzo-nuevo-record-en-produccion-de-cacao

Fishman, J., Creilson, J. K., Parker, P. A., Ainsworth, E. A., Vining, G. G., Szarka, J., … Xu, X. (2010). An investigation of widespread ozone damage to the soybean crop in the upper Midwest determined from ground-based and satellite measurements. *Atmospheric Environment*, *44*(18), 2248–2256. https://doi.org/10.1016/j.atmosenv.2010.01.015

García-Mozo, H., Yaezel, L., Oteros, J., & Galán, C. (2014). Statistical approach to the analysis of olive long-term pollen season trends in southern Spain. *Science of the Total Environment*, *473–474*, 103–109. https://doi.org/10.1016/j.scitotenv.2013.11.142

Gonzalez-Sanchez, a. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313–328. https://doi.org/10.5424/sjar/2014122-4439

Hansen, J. W., Potgieter, A., & Tippett, M. K. (2004). Using a general circulation model to forecast regional wheat yields in northeast Australia. *Agricultural and Forest Meteorology*, *127*(1–2), 77–92. https://doi.org/10.1016/j.agrformet.2004.07.005

Huang, Y., Lan, Y., Thomson, S. J., Fang, A., Hoffmann, W. C., & Lacey, R. E. (2010). Development of soft computing and applications in agricultural and biological engineering. *Computers and Electronics in Agriculture*, *71*(2), 107–127. https://doi.org/10.1016/j.compag.2010.01.001

Introducing Machine Learning. (n.d.). Retrieved from https://www.mathworks.com/tagteam/89703_92991v00_machine_learning_section1_ebook_v12.pdf

Kar, G., & Kumar, A. (2014). Forecasting rainfed rice yield with biomass of early phenophases, peak intercepted PAR and ground based remotely sensed vegetation indices. *Journal of Agrometeorology*, *16*(1), 94–103.

Kouadio, A. L., Djaby, B., Duveiller, G., El Jarroudi, M., & Tychon, B. (2012). Cinétique de décroissance de la surface verte et estimation du rendement du blé d'hiver. *Biotechnology, Agronomy and Society and Environment*, *16*(2), 179–191.

Kuncheva, L. I. (Ludmila I. (2004). *Combining pattern classifiers : methods and algorithms*. J. Wiley.

Lobell, D. B., Cahill, K. N., & Field, C. B. (2007). Historical effects of temperature and precipitation on California crop yields. *Climatic Change*, *81*(2), 187–203. https://doi.org/10.1007/s10584-006-9141-3

Logan, T. M., McLeod, S., & Guikema, S. (2016). Predictive models in horticulture: A case study with Royal Gala apples. *Scientia Horticulturae*, *209*, 201–213. https://doi.org/10.1016/j.scienta.2016.06.033

Mavromatis, T. (2014). Pre-season prediction of regional rainfed wheat yield in Northern Greece with CERES-Wheat. *Theoretical and Applied Climatology*, *117*(3–4), 653–665. https://doi.org/10.1007/s00704-013-1031-9

Mejía, L. F. (2018). *Departamento Nacional de Planeación*. Retrieved from https://colaboracion.dnp.gov.co/CDT/Prensa/Presentación Big Data Política explotación datos.pdf

Mishra, S., Mishra, D., & Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: A review paper. *Indian Journal of Science and Technology*, *9*(38). https://doi.org/10.17485/ijst/2016/v9i38/95032

Mkhabela, M. S., Mkhabela, M. S., & Mashinini, N. N. (2005). Early maize yield forecasting in

the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. *Agricultural and Forest Meteorology*, *129*(1–2), 1–9. https://doi.org/10.1016/j.agrformet.2004.12.006

Moreto, V. B., & Rolim, G. D. S. (2015). Agrometeorological models for groundnut crop yield forecasting in the Jaboticabal, São Paulo State region, Brazil. *Acta Scientiarum. Agronomy*, *37*(4), 403. https://doi.org/10.4025/actasciagron.v37i4.19766

Nadler, A. J., & Bullock, P. R. (2011). Long-term changes in heat and moisture related to corn production on the Canadian Prairies. *Climatic Change*, *104*(2), 339–352. https://doi.org/10.1007/s10584-010-9881-y

Naylor, R., Falcon, W., Rochberg, D., & Wada, N. (2001). Using El Nino/Southern Oscillation climate data to predict rice production in Indonesia. *Climatic Change*, 255–265. https://doi.org/10.1023/A:1010662115348

Oteros, J., Orlandi, F., García-Mozo, H., Aguilera, F., Dhiab, A. Ben, Bonofiglio, T., … Galán, C. (2014). Better prediction of Mediterranean olive production using pollen-based models. *Agronomy for Sustainable Development*, *34*(3), 685–694. https://doi.org/10.1007/s13593-013-0198-x

Pagani, V., Stella, T., Guarneri, T., Finotto, G., van den Berg, M., Marin, F. R., … Confalonieri, R. (2017). Forecasting sugarcane yields using agro-climatic indicators and Canegro model: A case study in the main production region in Brazil. *Agricultural Systems*, *154*(March), 45–52. https://doi.org/10.1016/j.agsy.2017.03.002

Park, S. J., Hwang, C. S., & Vlek, P. L. G. (2005). Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. *Agricultural Systems*, *85*(1), 59–81. https://doi.org/10.1016/j.agsy.2004.06.021

PROCOLOMBIA. (n.d.). Retrieved August 13, 2018, from http://www.inviertaencolombia.com.co/sectores/agroindustria.html

Rolim, G. D. S., Novo, M. D. C. D. S. S., Pantano, A. P., & Trani, P. E. (2011). Modelagem agrometeorolÃ³gica para estimaÃ§Ã£o do desenvolvimento e da produÃ§Ã£o de jilÃ³\rAgrometeorological model to estimate development and production of "jilÃ³," (19), 832–837.

SAC-Sociedad de Agricultores de Colombia. (n.d.). Retrieved August 10, 2018, from

http://sac.org.co/es/noticias/536-el-cacao-sera-el-cultivo-de-la-paz.html

Secretaría de agricultura. (2017). Retrieved August 10, 2018, from http://www.santander.gov.co/index.php/secretaria-agricultura

Semana- Comercio. (2018). Retrieved August 10, 2018, from https://www.semana.com/contenidos-editoriales/hay-campo-para-la-paz/articulo/las-hectareas-del-territorio-colombiano-donde-siembran-cacao/565714

Semana-Agricultura. (2017). Retrieved August 10, 2018, from https://www.semana.com/educacion/articulo/farc-en-que-se-quieren-formar-los-exguerrilleros-de-las-farc/531826

Tack, J., Barkley, A., & Nalley, L. L. (2015). Effect of warming temperatures on US wheat yields. *Proceedings of the National Academy of Sciences*, *112*(22), 6931–6936. https://doi.org/10.1073/pnas.1415181112

Toggweiler, J., & Key, R. (2001). Ocean circulation: Thermohaline circulation. *Encyclopedia of Atmospheric Sciences*, *4*(December 2007), 1549–1555. https://doi.org/10.1002/joc

Toscano, P., Gioli, B., Genesio, L., Vaccari, F. P., Miglietta, F., Zaldei, A., … Porter, J. R. (2014). Durum wheat quality prediction in Mediterranean environments: From local to regional scale. *European Journal of Agronomy*, *61*, 1–9. https://doi.org/10.1016/j.eja.2014.08.003

UPRA. (n.d.). Retrieved August 10, 2018, from http://www.upra.gov.co/uso-y-adecuacion-de-tierras/evaluacion-de-tierras/zonificacion

Zheng, H., Chen, L., Han, X., Zhao, X., & Ma, Y. (2009). Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. *Agriculture, Ecosystems and Environment*, *132*(1–2), 98–105. https://doi.org/10.1016/j.agee.2009.03.004