

Optimización de parámetros en redes neuronales artificiales para la predicción del rendimiento agrícola de cultivos de cacao mediante métodos de selección e ingeniería de características.

Autores:

María Alejandra Gómez Castellanos

Estefanía Jaimes Gil

Trabajo de Grado para Optar el título de Ingeniero Industrial

Director:

PhD En Física Matemática. Henry Lamos Díaz

Codirector:

Ingeniero Industrial. David Esteban Puentes Garzón

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2019

DEDICATORIA

A Dios, mi refugio, que me ha dado tanto y me sigue sorprendiendo con sus infinitas bendiciones, por poner en mi camino las personas indicadas.

A mi madre Alba Luz, por confiar y creer en mí como nadie más sabe hacerlo, y por su inmenso sacrificio sin el cual no podría cumplir los sueños que Dios puso en mi corazón.

A mi padre Luis Carlos por resaltar siempre mis habilidades y talentos, y su enorme apoyo a mi formación personal y profesional.

A Nicolás por su paciencia, consejos y cariño incondicional.

A mi compañera y amiga Estefanía por ser mi complemento perfecto.

María Alejandra Gómez Castellanos

DEDICATORIA

A Dios por darme vida, salud, amor, por guiar cada paso que doy, por llenar mi vida de bendiciones y por haber puesto en mi camino a personas maravillosas, llenas de luz que me ayudan ser mejor cada día.

A mis padres Zoraida Gil y Martin Jaimes por el apoyo, el amor que me brindan, por inculcarme valores y principios valiosos, por la motivación constante y necesaria que ha permitido cumplir lo que me propongo. Ustedes son el motor de mi vida.

A Sharick Sofía Jaimes por estar pendiente de todo lo que ocurre en mi vida, por amarme tanto y llenar mi vida de alegría.

A Luis Carlos Gamas por apoyarme en todo momento sin cuestionar nada.

A Alejandra, mi compañera de tesis, por compartir conmigo todo este tiempo, porque en medio de todos los momentos y emociones que vivimos durante el desarrollo del proyecto nos supimos complementar y logramos un buen resultado.

Estefanía Jaimes Gil

AGRADECIMIENTOS

A nuestro director Henry Lamos, agradecemos el tiempo y la orientación dedicada para la realización del presente proyecto.

A nuestro codirector, David Puentes, por el apoyo y conocimientos brindados en el desarrollo de la investigación.

A la Escuela de Estudios Industriales y Empresariales, por ser la casa del saber para nuestra formación profesional.

Al grupo de investigación ÓPALO, por los espacios y el conocimiento ofrecido en la ejecución del proyecto.

A la Corporación Colombiana de Investigación Agropecuaria, por ofrecernos la información necesaria para la elaboración de la presente investigación.

Tabla de Contenido

Introducción.....	20
1. Planteamiento del problema	23
2. Justificación del proyecto.....	25
3. Objetivos	27
3.1. Objetivo general	27
3.2. Objetivos específicos.....	27
4. Revisión de literatura	28
5. Marco teórico.....	37
5.1. Predicción	37
5.2. Predicción de rendimientos de cultivos	37
5.3. Aprendizaje automático	38
5.3.1. Aprendizaje automático no supervisado	40
5.3.2. Aprendizaje automático supervisado.....	40
5.4. Redes neuronales artificiales.....	42
5.5. Perceptrón multicapa	45
5.6. Mecanismo de aprendizaje.....	46
5.7. Validación de modelos	48
5.7.1. Holdout.	48

5.7.2. Validación cruzada KFold	49
5.8. Ingeniería de características	49
5.8.1. Análisis de componentes principales.....	50
5.9. Selección de características	51
5.9.1. Análisis de correlaciones	53
5.9.2. Análisis de función de regresión	54
5.9.3. Paso a Paso.....	54
5.10. Identificación de Hiperparámetros	55
5.10.1 Búsqueda Exhaustiva.....	56
5.10.2 Algoritmo genético	56
5.11. Métricas de bondad de ajuste	59
6. Metodología.....	60
6.1. Consolidación de la base de datos.....	61
6.1.1. Limpieza del conjunto de datos.....	62
6.2. Desarrollo del modelo predictivo basado en redes neuronales artificiales.....	63
6.3. Aplicación de métodos de selección e ingeniería de características	69
6.3.1. Análisis de correlaciones	70
6.3.2. Análisis univariante por función de regresión	72
6.3.3. Análisis por componentes principales	73
6.3.4. Selección paso a paso	75

6.4. Optimización del modelo predictivo	77
6.4.1. Desarrollo del Algoritmo Búsqueda Exhaustiva.....	78
6.4.2. Desarrollo del algoritmo genético	79
7. Resultados.....	81
7.1. Métricas de ajuste para el primer modelo de red neuronal	81
7.2. Métricas de ajuste aplicando selección e ingeniería de características	82
7.3 Métricas de ajuste para los modelos de red optimizados	83
7.3.1. Algoritmo Búsqueda Exhaustiva.....	84
7.3.2. Algoritmo Genético	84
7.3.3. Comparación de los algoritmos de búsqueda.....	85
7.4. Comparación de los modelos de red neuronal	85
7.5. Comparación del modelo red neuronal antes y después de optimización	86
8. Conclusiones.....	88
9. Recomendaciones	91
Referencias bibliográficas	93

Lista de Tablas

Tabla 1. Cumplimiento de objetivos específicos	22
Tabla 2. Hiperparámetros primer modelo de red neuronal	64
Tabla 3. Pesos y umbral asociado a la neurona oculta 1	65
Tabla 4. Pesos y umbral asociado a la neurona de salida	68
Tabla 5. Correlaciones de variables independientes con la variable dependiente y	70
Tabla 6. Correlaciones entre las 6 primeras variables	71
Tabla 7. Valores de puntuación de función de regresión lineal.....	72
Tabla 8. Varianza explicada por componente principal.....	73
Tabla 9. Posiciones variables en 10 iteraciones.....	75
Tabla 10. Valores de variables.....	76
Tabla 11. Posibles valores para Búsqueda Exhaustiva.	78
Tabla 12. Solución por Búsqueda Exhaustiva.	79
Tabla 13. Parámetros de búsqueda Algoritmo Genético	79
Tabla 14. Posibles valores para Algoritmo Genético.....	80
Tabla 15. Solución por Algoritmo Genético	81
Tabla 16. Métricas de ajuste Primer modelo de red neuronal	82
Tabla 17. Métricas de ajuste según método de selección.....	82
Tabla 18. Métricas de ajuste solución por Búsqueda Exhaustiva.....	84
Tabla 19. Métricas de ajuste solución por Algoritmo Genético	84
Tabla 20. Tiempos computacionales.....	85
Tabla 21. Comparaciones métricas de ajuste.....	86

Tabla 22. Incremento porcentual R^2	86
Tabla 23. Comparaciones métricas de ajuste primer modelo y modelo optimizado	87

Lista de figuras

Figura 1. Proceso Aprendizaje Automático.....	39
Figura 2. Proceso SML.	41
Figura 3. Modelo neuronal McCulloch-Pitts.	42
Figura 4. Principales funciones de activación.	44
Figura 5. Función de activación Relu.	44
Figura 6. Arquitectura de perceptrón multicapa.	45
Figura 7. Validación Holdout.	48
Figura 8. Validación cruzada KFold (k=5).....	49
Figura 9. Ciclo evolutivo.	57
Figura 10. Cruce de 1 punto.	58
Figura 11. Cruce de 2 puntos.....	58
Figura 12. Cruce uniforme.	58
Figura 13. Diagrama de flujo proceso de limpieza de datos.	63
Figura 14. Pesos asociados a la neurona oculta 1.	67
Figura 15. Pesos asociados a la neurona de salida.....	69

Lista de Apéndices

(Apéndices en CD adjunto)

Apéndice A. Análisis Bibliométrico.....	28
Apéndice B. Descripción variables.....	61
Apéndice C. Pesos y umbrales primer modelo de Red Neuronal.....	65
Apéndice D. Selección de variables Paso a Paso.....	77
Apéndice E. Artículo publicable.....	20

Resumen

Título del proyecto: Optimización de parámetros en redes neuronales artificiales para la predicción del rendimiento agrícola de cultivos de cacao mediante métodos de selección e ingeniería de características.*

Autores: María Alejandra Gómez Castellanos, Estefanía Jaimes Gil**

Palabras clave: Agricultura, redes neuronales artificiales, optimización de parámetros, selección e ingeniería de características, predicción.

Descripción: El presente trabajo de investigación tiene como objetivo principal la optimización de un modelo de red neuronal para la predicción de rendimiento agrícola de un cultivo de cacao mediante la aplicación de métodos de selección e ingeniería de características y la aplicación de un dos algoritmos de búsqueda, uno de ellos el algoritmo de búsqueda exhaustiva y la aplicación del algoritmo genético para determinar los hiperparámetros (función de activación, solver, número de neuronas en la capa oculta, momentum y tasa de aprendizaje) adecuados de la red neuronal artificial (Perceptrón Multicapa).

Para el objetivo principal se propone la comparación de dos modelos de red, el primero sin selección e ingeniería de características y con una arquitectura establecida a prueba y error por los autores del proyecto basados en la revisión de literatura realizada, y el segundo modelo con la aplicación de métodos de selección e ingeniería de características y la arquitectura basada en los resultados de la búsqueda adaptativa que realiza el algoritmo de búsqueda (algoritmo genético), con la finalidad de obtener un mejor ajuste o coeficiente de determinación en el segundo modelo

* Trabajo de grado.

** Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz, Ph.D en Física-Matemática. Codirector: David Esteban Puentes Garzón, M.Sc (c) en Ingeniería Industrial.

de red neuronal artificial. Cabe resaltar que los datos sujetos al estudio fueron obtenidos durante el periodo entre 2015 y 2017 de un cultivo experimental de cacao ubicado en el centro de Investigación la Suiza Santander.

Abstract

Project title: Parameters optimization in artificial neural networks for cocoa yield prediction using selection and feature engineering.*

Authors: María Alejandra Gómez Castellanos, Estefanía Jaimes Gil**

Keywords: Agriculture, artificial neural networks, parameters optimization, selection and feature engineering, prediction.

Description: The main objective of this research work is the optimization of a neural network model for the prediction of agricultural performance of a cocoa crop through the application of methods of selection and engineering of characteristics and the application of a two search algorithms, one of them the algorithm of exhaustive search and the application of the genetic algorithm to determine the hyperparameters (function of activation, solver, number of neurons in the hidden layer, momentum and rate of learning) of the artificial neural network (Multilayer Perceptron).

For the main objective we propose the comparison of two network models, the first one without selection and engineering of characteristics and with an established architecture to trial and error by the authors of the project based on the literature review, and the second model with the application of methods of selection and engineering of characteristics and the architecture based on the results of the adaptive search performed by the search algorithm (genetic algorithm), in order to obtain a better adjustment or coefficient of determination in the second model of the neural

* Degree project.

** Faculty of Physicomechanical Engineering. Industrial and Business School. Director: Henry Lamos Díaz, Ph.D in Physics-Mathematics. Co-director: David Esteban Puentes Garzón, M.Sc (c) in Industrial Engineer.

network artificial. It should be noted that the data subject to the study were obtained during the period between 2015 and 2017 of an experimental cocoa crop located at the Santander Switzerland Research Center.

Introducción

Actualmente los países en desarrollo, Colombia incluido, han comprendido la necesidad de definir y estructurar estrategias y políticas que propendan a impulsar el desarrollo de la agricultura y sus territorios rurales, por lo cual, dichas estrategias y políticas deben ser rediseñadas y adaptadas en coherencia con el entorno actual (Ministerio de Agricultura y Desarrollo Rural, 2018). Tras la firma del acuerdo de paz, y gracias a la iniciativa de Colombia como el primer país de América Latina y el octavo en el mundo en adoptar una política nacional de explotación de datos (que surgió como un requerimiento del Plan de Desarrollo Nacional 2014-2018) y bajo la premisa “la explotación de datos es el aprovechamiento de los mismos para generar valor social y económico” (Mejía, 2018), se logró establecer una herramienta clave que contribuye a el procesamiento y análisis de los datos, desarrollo de nuevos conocimientos, creación de valor, así como a la consolidación de información, más amplia y de mayor calidad que permita la definición de políticas y la mejora de la toma de decisiones (Departamento Nacional de Planeación, 2014).

Alrededor del mundo el pronóstico de cultivos y las herramientas de decisión se utilizan cada vez más en el campo agrícola para mejorar la eficiencia de la producción (Khairunniza-Bejo, Mustaffha, & Ishak, 2014). Sin embargo, en Colombia el sector agropecuario sigue siendo un sector poco potencializado debido a que se caracteriza por bajos niveles de tecnología, acción que conlleva a una baja productividad en un mercado cada vez más exigente (Acosta Bonilla & Villarraga Rubio, 2006).

Inicialmente el pronóstico de cultivos se vio como una herramienta para comprender el rendimiento de sistemas complejos, sus factores externos y condiciones climáticas hacia el rendimiento del cultivo. Las redes neuronales artificiales han demostrado ser de gran utilidad en

ello, destacándose sobre otros métodos debido a que asume características de funcionamiento comunes a las redes neuronales biológicas, siendo capaces de procesar la información de manera paralela, potenciar las interconexiones y facilitar la adaptación (Ballesteros-ricaurte, 2015). Sin embargo, dado a la complejidad de su estructura, es importante la correcta selección de sus parámetros. Decidir cuál será su tasa de aprendizaje, momemtum, función de activación, tolerancia, número de capas ocultas y tamaño de capas ocultas es clave para la implementación y optimización de la red neuronal, ejercicio que requiere el uso de métodos de búsqueda adaptativa para encontrar la arquitectura óptima según cada caso de estudio porque de lo contrario se convierte en un ejercicio de prueba y error sujeto a criterio del investigador.

De la creciente necesidad de fortalecer las capacidades de Santander en materia de agronomía (ciencia cuyo objetivo es mejorar la calidad de la producción agrícola con base en principios científicos y tecnológicos) y el rol de la Ingeniería Industrial en los sistemas organizacionales como medida y estrategia de respuesta ante eventos repentinos y externos en el mejoramiento de la toma de decisiones (Montoya Restrepo, Montoya-Restrepo, & Lowy Ceron, 2015), se origina el presente trabajo que busca diseñar una red neuronal artificial mediante un algoritmo genético capaz de predecir de forma óptima el rendimiento agrícola de un cultivo experimental de cacao ubicado en el departamento, realizando también selección e ingeniería de características sobre las variables de las que depende el modelo para así obtener como resultado mayor precisión en la predicción del rendimiento y mayor conocimiento sobre los elementos que influyen en el cultivo.

Tabla 1

Cumplimiento de objetivos específicos

Objetivos específicos	Cumplimiento
Realizar una revisión de la literatura sobre métodos de selección e ingeniería de características en redes neuronales artificiales aplicados en la predicción del rendimiento agrícola de cultivos.	Capítulo 4
Elaborar un modelo predictivo basado en redes neuronales artificiales.	Numeral 6.2.1
Seleccionar y aplicar métodos de selección e ingeniería de características para la predicción del rendimiento agrícola de cultivos de cacao basados en redes neuronales artificiales.	Numeral 6.2.2
Evaluar y comparar el desempeño del modelo predictivo después de implementar métodos de selección e ingeniería de características basadas en el modelo predictivo propuesto	Numeral 7.5
Elaborar un artículo de carácter publicable sobre los resultados obtenidos en la investigación.	Apéndice E

1. Planteamiento del problema

Las redes neuronales son modelos que intentan reproducir el comportamiento del cerebro humano. Como tal, el modelo realiza una simplificación estableciendo cuales son los elementos relevantes del sistema, bien porque la cantidad de información de la que se dispone es excesiva o bien porque es redundante. Una elección adecuada de sus características más una estructura conveniente, es el procedimiento convencional utilizado para construir redes capaces de realizar una determinada tarea (Hilera González & Martínez Hernando, 1995).

Las redes neuronales artificiales poseen ventajas bien definidas sobre otros modelos de predicción, que las hacen más adecuadas para ciertos problemas y situaciones. Tienen la capacidad de aprender y modelar relaciones no lineales y complejas, lo que es realmente importante para modelar sistemas reales; también pueden generalizarse, es decir, que después de aprender de las entradas iniciales y sus relaciones, puede inferir relaciones de datos no observados, lo que hace que el modelo pronostique sobre datos no presentes durante el aprendizaje; a diferencia de muchas otras técnicas de predicción, las redes neuronales artificiales no imponen ninguna restricción en las variables de entrada (como la forma en que deben distribuirse). Además, muchos estudios han demostrado que pueden modelar mejor la heterocedasticidad, es decir, datos con alta variabilidad (volatilidad) y varianza no constante, dada su capacidad para aprender relaciones ocultas en los datos sin imponer ninguna relación fija (Mahanta, 2017).

Las redes neuronales deben ser entrenadas para cada problema que se quiera abordar. Además, es necesario realizar múltiples pruebas para determinar la arquitectura adecuada. Debido a que las redes se entrenan en lugar de programarlas, estas requieren la definición previa de sus hiperparámetros definidos como variables de configuración que son internas al modelo y cuyo

valor se puede estimar a partir de los datos (número de capas ocultas, número de nodos por capa, interconexiones, función de activación, etc.), y su entrenamiento puede llegar a ser más complejo de acuerdo a lo que se quiera aprender.

Comprendiendo la dificultad que implica la construcción de una red neuronal artificial y la importancia de decidir correctamente sus hiperparámetros, el presente proyecto propone la optimización de una red neuronal aplicada a la predicción de rendimiento agrícola de cultivos, mediante el uso de un algoritmo genético como método de búsqueda adaptativa, que deberá hallar la arquitectura optima en un conjunto de soluciones candidatas, también llamada población. Se escogió el algoritmo genético ya que “aplicado a la estructura de una red neuronal puede encontrar las topologías óptimas específicas del dominio que faciliten el aprendizaje y la generalización de la misma”(Fiszelew & García-Martínez, 2012) y sumado a esto puede encontrar soluciones adecuadas en poco tiempo, la mutación aleatoria permite tener una mayor diversidad de soluciones y codificarlos es relativamente fácil en comparación con otros algoritmos que hacen el mismo trabajo; y adicionalmente la selección e ingeniería de características de entrada del respectivo modelo, dado que proporciona múltiples beneficios, como: mejorar el rendimiento de predicción de los modelos, proporcionar modelos predictores más rápidos, proporcionar una mejor comprensión del proceso subyacente que generó los datos, facilitar la visualización y comprensión de los datos, reducir los requisitos de medición y almacenamiento y reducir los tiempos de entrenamiento.

2. Justificación del proyecto

El aumento de la productividad , por medio de la diversificación, la modernización y la innovación, como motores para promover el crecimiento económico, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos ha impulsado a muchos países y territorios a centrar sus esfuerzos en el desarrollo de su agricultura (Naciones Unidas de Colombia, 2015). Y siendo Colombia un país en vía de desarrollo, gran parte de su población depende aún de la agricultura local como sustento económico y alimenticio, por ello es necesario el apoyo de la investigación y la extensión agrícola para mejorar el aprovechamiento de los recursos locales destinados a estas prácticas (Organización de las Naciones Unidas para la Alimentación y la Agricultura, 2002).

Reconociendo el bajo aporte investigativo del país en cuanto a la implementación de herramientas de aprendizaje automático para pronóstico de cultivos y con la finalidad de aumentar la eficiencia de la producción (Khairunniza-Bejo et al., 2014) surge la idea de emplear redes neuronales artificiales en la predicción del cultivo agrícola de cacao, ya que éstas han demostrado ser de gran utilidad en ello, destacándose sobre otros métodos debido a que asume características de funcionamiento comunes a las redes neuronales biológicas, tienen la capacidad de aprender y modelar relaciones no lineales y complejas, lo que es realmente importante para modelar sistemas reales y además pueden generalizarse. Sin embargo, la selección e ingeniería de características y la arquitectura de la red neuronal artificial (hiperparámetros) son trascendentales para que el modelo sea un buen predictor y minimice errores. Reconociendo que esta tarea requiere mucho esfuerzo y tiempo para su identificación, se hace necesaria la utilización de herramientas de búsqueda, como el algoritmo genético, ya que puede encontrar las topologías óptimas específicas del dominio que faciliten el aprendizaje y la generalización de la misma (Fiszelew & García-

Martínez, 2012) y puede encontrar la solución adecuada en muy poco tiempo. Se busca optimizar los parámetros de un modelo de red neuronal artificial para la predicción del rendimiento agrícola de cultivos de cacao mediante métodos de selección e ingeniería de características.

3. Objetivos

3.1.Objetivo general

Optimizar los parámetros de las redes neuronales artificiales para la predicción del rendimiento agrícola de cultivos de cacao mediante la aplicación de métodos de selección e ingeniería de características.

3.2.Objetivos específicos

- Realizar una revisión de la literatura sobre métodos de selección e ingeniería de características en redes neuronales artificiales aplicados en la predicción del rendimiento agrícola de cultivos.
- Elaborar un modelo predictivo basado en redes neuronales artificiales.
- Seleccionar y aplicar métodos de selección e ingeniería de características para la predicción del rendimiento agrícola de cultivos de cacao basados en redes neuronales artificiales.
- Evaluar y comparar el desempeño del modelo predictivo después de implementar métodos de selección e ingeniería de características basadas en el modelo predictivo propuesto.
- Elaborar un artículo de carácter publicable sobre los resultados obtenidos en la investigación.

4. Revisión de literatura

De la búsqueda de literatura se encontró que los principales cultivos objeto de estudio para la predicción de su rendimiento fueron el maíz, la cebada, el tomate, el trigo, la soja, la caña de azúcar y el arroz. En algunos, la predicción fue bajo un método específico de aprendizaje automático y en otros se realizó una comparación de varios de ellos para encontrar cual se ajustaba mejor al problema. Además, los principales países que realizan investigación en el tema son Estados Unidos, Canadá, Irán, China e India (Ver Apéndice A).

Los métodos más recurrentes en las investigaciones fueron las redes neuronales artificiales (ANN, por sus siglas en inglés) y la regresión lineal múltiple (MLR, por sus siglas en inglés). Algunas investigaciones tuvieron como objetivo determinar si las redes neuronales artificiales son más precisas que otros modelos, como en la investigación “The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling” (Bocca & Rodrigues, 2016), estos investigadores evaluaron el impacto de diferentes prácticas para modelar el rendimiento en el cultivo de caña de azúcar, donde se realizó una selección de 13 artículos donde se aplicaban algunos modelos de aprendizaje automático y se comparó el desempeño de los modelos (árboles de regresión, modelo lineal generalizado, bosques aleatorios, máquinas de soporte vectorial, k vecinos más cercanos, redes neuronales artificiales). Además, los investigadores (Lykhovyd, 2018) compararon el modelo de red neuronal con la regresión lineal simple, la predicción del rendimiento se realizó sobre un cultivo de maíz dulce. Los datos de estudio fueron obtenidos durante el periodo entre 2014 a 2016 en Ucrania. Se evaluó el impacto de las profundidades de arado, las tasas de aplicación de fertilizantes minerales y la densidad de las plantas en el rendimiento del cultivo mediante el análisis de varianza. El resultado de la

evaluación de estas dos técnicas fue que los métodos de redes neuronales artificiales son mucho mejores que la regresión lineal en la predicción del rendimiento de maíz dulce debido a los valores del coeficiente de determinación (R^2) 0,978 para ANN y 0.897 para MLR. Por ende, redes neuronales artificiales son una herramienta de procesamiento de datos mucho mejor, especialmente en las ciencias de la vida y para la predicción de los procesos y fenómenos naturales no lineales (Lykhovyd, 2018).

Los autores (Emamgholizadeh, Parsaeian, & Baradaran, 2015) coinciden con este argumento debido a que en su estudio para la predicción del rendimiento de semilla de sésamo emplean redes neuronales artificiales y el modelo de regresión múltiple. Los autores analizan características como: el tiempo de floración, altura de la planta, número de cápsulas por planta, el peso de 1000 semillas y el número de semillas por cápsula. Los datos mostraron que para una red neuronal el error cuadrático medio (RMSE, por sus siglas en inglés) es de 0,339 y un coeficiente de determinación (R^2) de 0,901, los cuales son valores superiores a los alcanzados con el modelo de regresión múltiple (RMSE = 0.346 y R^2 = 0.779)

En las investigación de (Safa, Samarasinghe, & Nejat, 2015) y (Irmak et al., 2006) se muestra la capacidad de las redes neuronales para el estudio del cultivos agrícolas, en la primera investigación se desarrolló un modelo de redes neuronales artificiales con datos de 40 granjas en Canterbury, Nueva Zelanda, para predecir la producción del cultivo de trigo y comparar su desempeño con una regresión lineal múltiple. De igual manera en la investigación “Artificial neural network model as a data analysis tool in precision farming” se desarrolló un modelo de red neuronal de propagación hacia atrás (BPNN, por sus siglas en inglés) para predecir la distribución espacial de los rendimientos de la soja y para comprender las causas de la variabilidad del rendimiento del mismo campo en diferentes años y para campos independientes. En la

investigación asumieron que los rendimientos de la soya variaban debido a cuatro factores ambientales y sus interacciones: clima, fertilidad del suelo, características del paisaje y factores del sitio. Los datos usados fueron de 6 campos en 3 diferentes condados de Iowa. De igual manera, se demuestra la superioridad de las redes neuronales sobre otros modelos de predicción

El uso de herramientas como las de aprendizaje automático ayuda en la toma de decisiones y planificación de operaciones al aumentar la precisión en las predicciones. En la investigación de (Charoen-Ung & Mittrapiyanuruk, 2011) se desarrolló un modelo basado en bosque aleatorio (RF, por sus siglas en inglés) para predecir el grado de rendimiento de la caña de azúcar en un lote de agricultores en Tailandia. El número de registros del conjunto de datos de entrenamiento fue de 8.765 y para la prueba fue de 3.756. Se realizó una selección de características hacia adelante junto con el ajuste de parámetros para el entrenamiento y se evaluó este método junto con otros dos que no son de aprendizaje automático ya que son predicciones según el rendimiento del año anterior y la subjetividad del experto, la precisión fue de 75,83%, 51,52% y 65,50% respectivamente. Además, se ve reflejado que la implementación de ajuste de parámetros, ingeniería y selección de características aplicados a modelos pueden mejorar la precisión de la predicción con resultados de 75,83% cuando se realiza frente a un 71,88% cuando no.

La preocupación del gobierno de Fujian China ha estado relacionada con la necesidad de técnicas simples y precisas para estimar los rendimientos de los cultivos de arroz en las regiones montañosas debido a alteraciones climáticas como son las inundaciones, los tifones y sequias que amenazan la producción (Ji, Sun, Yang, & Wan, 2007). Debido a ello los autores (Ji et al., 2007) realizaron su investigación con el objetivo de investigar si los modelos ANN podían predecir efectivamente los cultivos de arroz evaluando el rendimiento del modelo en relación con el ajuste de parámetros y comparándolo con un modelo de regresión lineal. En los modelos fueron usados

datos históricos de rendimientos en diferentes ubicaciones a lo largo de Fujian. El ajuste de parámetros del modelo de red neuronal, como la tasa de aprendizaje, el número de nodos ocultos y tolerancia de entrenamiento afectó el desarrollo del modelo y la precisión de las predicciones de rendimiento del cultivo de arroz y se evidenció que cuando el conjunto de datos es más pequeño se requiere menor número de nodos ocultos y menor tasa de aprendizaje en la optimización del modelo.

Ahora, en la comparación de los dos modelos resultó que las redes neuronales produjeron predicciones de rendimiento más precisas que los modelos de regresión ya que los valores de R^2 fue de 0,67 para ANN y 0,52 para la regresión. Sin embargo, lo más interesante de la investigación es el modelado de la red neuronal y el esfuerzo que se acarrea para que la precisión del modelo sea la mejor, por ejemplo a medida que la cantidad de datos que se modelaba disminuía se requerían menos nodos ocultos, a medida que el número de nodos ocultos disminuía la tasa de aprendizaje óptima también lo hizo, a medida que disminuía el número de nodos el entrenamiento ANN tendía a ser más lento y requería pequeños incrementos de cambio en las asignaciones de pesos, lo que se refleja en la tasa de aprendizaje asignada. Lo cual dio como resultado varios modelos ANN con ajuste que son prometedores como una técnica más precisa y tienen el potencial de ser útiles para el monitoreo y el pronóstico de rendimiento de cultivos. (Ji et al., 2007).

Tanto (Ji et al., 2007) como (Kaul, Hill, & Walthall, 2005) realizaron sus investigaciones enfocados en lo mismo, la diferencia es que en la investigación de estos últimos autores se basó en la relación de los rendimientos de los cultivos y la utilización de nutrientes en el proceso de planificación en Maryland. Coinciden en que “El ajuste de los parámetros ANN como la velocidad de aprendizaje y la cantidad de nodos ocultos afectan la precisión de las predicciones de rendimiento de los cultivos” (Kaul et al., 2005).

Además de las investigaciones mencionadas anteriormente que pretenden estudiar las aplicaciones de los modelos de redes neuronales en la predicción de cultivos y sus beneficios, se encontraron otras investigaciones con propósitos distintos a la agricultura, que hicieron uso del algoritmo genético, algunos investigadores focalizaron la búsqueda en las variables de entrada (Feature Selection) y otros en los hiperparámetros de la red.

En la investigación “Feature subset selection using a genetic algorithm” (Yang & Honavar, 1998) utilizan un algoritmo genético para el problema de selección de subconjuntos de características en la clasificación de patrones con red neuronal artificial, dicho problema corresponde a la tarea de identificar y seleccionar un subconjunto útil de características que representan patrones de un conjunto más amplio de características, dado que las características en el conjunto más grande tienen diferentes costos y riesgos asociados, convirtiendo algunas en irrelevantes o mutuamente redundantes. Los autores reconocen la selección de características como un proceso que repercute en la precisión, tiempo de aprendizaje requerido, número necesario de ejemplos y costo del algoritmo.

Los autores replicaron las características originales una vez y duplicaron el número de características para introducir redundancia en el conjunto de entrenamiento, luego generaron un conjunto adicional de características irrelevantes y les asignaron valores booleanos aleatorios. Una vez finalizada la investigación los resultados indican que las redes construidas utilizando un subconjunto de características seleccionadas por Algoritmo Genético se comparan favorablemente con las redes que usan todas las características. En particular, la selección de subconjuntos de características dio como resultado una mejora significativa de la generalización. En el caso del estudio para la minimización de costo y maximización de precisión la función de aptitud que combinó precisión y costo superó a la que se basó solo en precisión en todos los aspectos: la

cantidad de funciones utilizadas, la precisión de la generalización y la cantidad de neuronas ocultas. Hecho que no es sorprendente, porque el primero intenta minimizar el costo y maximizar la precisión; que resulta en un número menor de características. Y el último enfatiza solo la exactitud.

Los investigadores (Gholipoor, Rohani, & Torani, 2013) en su trabajo “Optimization of traits to increasing barley grain yield using an artificial neural network” realizaron la optimización de los componentes que afectan el aumento del rendimiento de la cebada utilizando una red neuronal artificial (ANN) y un algoritmo genético (GA, por sus siglas en inglés). Se cultivaron diez genotipos de cebada para medir el rendimiento del grano y componentes que lo afectan. Los componentes o regresores que se analizaron inicialmente fueron: la aspereza al grano, la altura del cultivo, el peso de 1,000 granos, el número de granos por oreja, la duración del crecimiento vegetativo, la duración del llenado de grano, la tasa de llenado de grano, el número de timón, el índice de cosecha, la longitud de la espiga y el diámetro del tallo principal. Luego, se evaluó el coeficiente de regresión múltiple y correlación de todos los regresores con el rendimiento. Los regresores con efectos significativos se consideraron participantes del rendimiento del grano de cebada para después dividirlos en 80% de entrenamiento y 20% de prueba, estos se utilizaron para estimar los parámetros del modelo ANN.

En otras investigaciones como “Sugarcane Yield Grade Prediction using Random Forest with Forward Feature Selection and Hyper-parameter Tuning” de (Charoen-Ung & Mittrapiyanuruk, 2011) se ha realizado selección e ingeniería de características, en este caso proponen la selección e ingeniería de características hacia adelante con un ajuste de parámetros para el entrenamiento del clasificador de bosque aleatorio para el conjunto de datos que se obtienen de parcelas de caña de azúcar en Tailandia. Ellos comparan la precisión de su método con dos líneas, una es utilizar el

rendimiento real del año pasado como la predicción y la otra el experto predice manualmente el rendimiento de cada parcela concluyendo que el método donde se realiza selección e ingeniería de características ayuda a los involucrados en la toma de decisiones en la planeación.

En las investigaciones “Designing a neural networks using Genetic Algorithms” y “Training Feedforward Neural Networks Using Genetic Algorithms” los investigadores (Miller, Todd, & Hegde, 1989) y (Montana & Davis, 1989) respectivamente presentan un algoritmo genético usado en el entrenamiento de la red. En la primera investigación, el algoritmo genético después llamado Innovator (fusión de Genesis 4.5 de Grefenstette y el sistema backpropagation para entrenar redes multicapa de McClelland) que evoluciona un arreglo de red neuronal para tres diferentes tareas, donde los operadores genéticos estándar modificados actúan sobre las poblaciones de estos genotipos para producir arquitecturas de red con mayor aptitud sobre arquitecturas sucesivas dependiendo de la tarea asignada, en este caso se usaron los problemas: Xor, una extensión del mismo en dos dimensiones y Pattern Copying, se evaluó el desempeño de las tres arquitecturas con la aptitud (Fitness) mediante la formación de instancias particulares de cada red y el error de rendimiento total. Finalizada la investigación, los investigadores afirman que la automatización de la búsqueda de arquitecturas de red neuronal mediante algoritmos genéticos, es el mejor camino para eliminar restricciones de sesgos humanos (human bias) en el diseño de la misma. Para la segunda investigación los autores realizan dos experimentos para un estudio de clasificación de imágenes, En el primero emplean el algoritmo genético en vez de propagación hacia atrás y en el segundo describen la evolución del algoritmo a medida que se le agrega más y más complejidad. Los datos usados en el experimento fueron de una etapa en el procesamiento de datos de sonar pasivos de matrices de receptores acústicos submarinos.

En forma similar los autores Fiszlewe & García-Martínez R, (Whitley, Starkweather, & Bogart, 1990) realizan su investigación “Genetic algorithms and neural networks: optimizing connections and connectivity” en la que desarrollaron un algoritmo genético llamado Genitor que contrario a un algoritmo estándar (en el que toda la población se reproduce en una sola generación con descendientes que desplazan a los padres) funciona de la forma: primero se seleccionan dos padres de la población, para que sus copias sean combinadas y produzcan dos descendientes. Luego uno de los descendientes se descarta de forma aleatoria, al otro se le permite reemplazar la cadena de clasificación más baja en la población (la descendencia no reemplaza una cadena principal). Esta nueva cadena se clasifica según su desempeño en relación con el resto de la población, de modo que ahora puede competir por las oportunidades reproductivas.

Para mantener la diversidad genética utilizan la mutación adaptativa que responde al nivel de diversidad genética en la población. Además, se observó que, durante las primeras etapas de búsqueda, cuando existe una amplia diversidad en la población, la mutación se produce a tasas muy bajas. Sin embargo, a medida que la diversidad disminuye en la población, la tasa de mutación aumenta. De este modo, la mutación ayuda a mantener la diversidad y, por lo tanto, ayuda a mantener el potencial de la búsqueda. El enfoque abordado fue definir una red de avance "máximo conectado" y luego usar el algoritmo genético para descubrir una combinación de conexiones que mejora la capacidad de aprendizaje. Por lo que el enfoque es un tipo de algoritmo de poda, pero es distinto de otros ya que las conexiones no se eliminan fácilmente. El espacio de "posibles conectividades" se explora y las conexiones se eliminan y se reintroducen.

Los algoritmos genéticos aportan beneficios sobre la construcción de la red gracias a que mientras el algoritmo busca un conjunto adecuado de pesos, buscará simultáneamente una conectividad que haría más fácil buscar el espacio del problema. Si estos dos procesos se realizan

para complementarse, entonces el resultado es redes más pequeñas que aprenden más rápido. Como evidencia, antes de implementar el algoritmo genético los investigadores tenían una red “conectada al máximo” con 53 conexiones y después el algoritmo redujo la red a 29 conexiones.

Los algoritmos genéticos también han sido usados para la búsqueda de parámetros en diferentes modelos, por ejemplo los investigadores (Huang & Wang, 2006) emplearon el algoritmo genético para la búsqueda de parámetros del modelo máquinas de soporte vectorial (SVM) en su investigación “A GA-based feature selection and parameters optimization for support vector machines”, sobre un conjunto de datos de clasificación de imágenes y este hizo que mejorara significativamente la exactitud de la clasificación. (Fiszelew & García-Martínez, 2012) y (Su, 2011) en sus trabajos de investigación “Generación automática de redes neuronales con ajuste de parámetros basado en algoritmos genéticos” y “Neural Network Hyperparameter Optimization with Genetic Algorithm on Handwritten Digit Problem, and the Observation on Neuron network by Distinctiveness” utilizan el algoritmo para encontrar la solución de la arquitectura óptima de red neuronal para aprender problemas. Este algoritmo evolutivo emplea codificación directa y usa el error de la red entrenada como medida de desempeño para guiar la evolución y mejorar el resultado.

A partir de la revisión de literatura se puede observar que las redes neuronales han sido un modelo de predicción muy usado por los investigadores ya que tiene ventajas sobre los demás modelos que lo hacen atractivo, fácil y muy útil en la agricultura. La selección e ingeniería de características hace que el modelo sea mucho más preciso y cuando se utiliza el algoritmo genético en el modelo para la optimización, se evidencia el mejoramiento de los resultados.

5. Marco teórico

5.1. Predicción

“Predecir es el proceso de aplicar un modelo estadístico a los datos con el fin de pronosticar observaciones nuevas o futuras.”(Shmueli, 2010).

Uno de los problemas de la ciencia experimental es encontrar funciones que ajusten o expliquen datos que se observan de fenómenos naturales. La principal ventaja de la existencia de tales funciones es la posibilidad de predecir el comportamiento del sistema natural en el futuro y controlar sus salidas mediante la aplicación de las entradas adecuadas. En ocasiones es posible encontrar un modelo matemático exacto que explique el proceso del que provienen los datos que observamos, sin embargo, muchas veces esta búsqueda se dificulta debido a que los datos observados tienden a estar acompañados de ruido y los mecanismos exactos que los generan normalmente son desconocidos. El objetivo, en este caso, es estimar el modelo subyacente que se aproxime a los datos observados. Algunos ejemplos interesantes son la predicción de valores en bolsa, la predicción meteorológica o la clasificación de masas tumorales, los cuales son posibles de predecir gracias a datos de ejemplo o experiencias pasadas, que permiten la identificación de patrones en el conjunto de datos para construir buenas aproximaciones al problema (Palma & Marín, 2008).

5.2. Predicción de rendimientos de cultivos

La predicción del rendimiento (relación de la producción total de un cultivo cosechado por unidades de área de terreno utilizada) en los cultivos es un problema importante y complejo en la agricultura dado que depende de diversos factores como: las decisiones de los agricultores, riegos

aplicados, fertilizantes, rotación de cultivos, preparación de la tierra, entre otros, además de factores incontrolables como el clima, los subsidios, el mercado, etc. La predicción de rendimiento tradicionalmente se ha basado en la experiencia a largo plazo de los agricultores para campos, cultivos y condiciones climáticas específicas, que pueden ser en su mayoría inexactos. De igual manera, se utilizan estimadores simples, como el promedio de varios rendimientos anteriores o el último rendimiento obtenido, sin embargo, el rendimiento del cultivo varía espacial y temporalmente con un comportamiento no lineal, introduciendo grandes desviaciones de un año a otro. Por lo tanto, se han desarrollado métodos más eficientes, que pueden clasificarse como modelos de crecimiento de cultivos y modelos basados en datos. Los modelos de crecimiento de cultivos realizan una abstracción del mecanismo dinámico de las etapas fisiológicas de la planta al adaptarlos a un modelo matemático, mientras que los modelos basados en datos se construyen empíricamente, sin requerir un conocimiento profundo sobre los mecanismos físicos que produjeron los datos. Tales técnicas son económicas, relativamente fáciles de aplicar y no necesitan una estructura predefinida del modelo (Gonzalez-Sanchez, Frausto-Solis, & Ojeda-Bustamante, 2014).

5.3. Aprendizaje automático

El Aprendizaje Automático es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Una máquina es un sistema organizado capaz de transformar un mensaje de entrada en otro de salida, de acuerdo con algún principio de transformación. Si tal principio está sujeto a un criterio de validez de funcionamiento, y si el método de transformación se ajusta a fin de que tienda a mejorar el funcionamiento del sistema de

acuerdo con este criterio, se dice que el sistema aprende (Bustos M., 2005). El proceso necesario para el desarrollo de técnicas de aprendizaje automático puede ilustrarse de la siguiente forma.

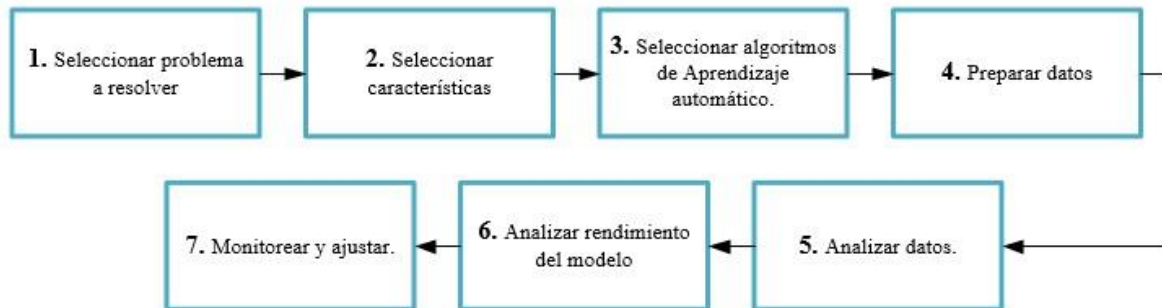


Figura 1. Proceso Aprendizaje Automático.

Hay varias razones de ingeniería por las que el aprendizaje automático es importante, algunas de estas razones son:

- Existen tareas que no se pueden definir bien, es decir, se pueden especificar pares de entrada/salida, pero no una relación concisa entre ellas. Las máquinas pueden ajustar su estructura interna para producir salidas correctas para un gran número de entradas de muestra y, por lo tanto, restringen adecuadamente su función de entrada/salida para aproximarse a la relación implícita en los ejemplos.
- Los métodos de aprendizaje automático a menudo se pueden usar para extraer estas relaciones y correlaciones importantes entre grandes conjuntos de datos (minería de datos).
- La cantidad de conocimiento disponible sobre ciertas tareas en ocasiones es demasiado grande para la codificación explícita por humanos.

5.3.1. Aprendizaje automático no supervisado. En el enfoque de aprendizaje no supervisado, las instancias no tienen etiqueta. A través de la aplicación de algoritmos no supervisados (agrupamiento), se busca obtener una agrupación coherente de los datos en función de las relaciones entre las variables definidas en los datos (Muhammad & Yan, 2015).

Entre las técnicas de aprendizaje no supervisado se encuentran: clustering; agrupamiento de objetos en regiones donde la similitud mutua es elevada; visualización, que permite observar el espacio de instancias en un espacio de menor dimensión; reducción de dimensionalidad, los datos de entrada son agrupados en subespacios de una dimensión más baja que la inicial, entre otros.

5.3.2. Aprendizaje automático supervisado. Aprendizaje automático supervisado (SML, por sus siglas en inglés) es la búsqueda de algoritmos que razonan a partir de instancias proporcionadas externamente para generar hipótesis generales, que luego hacen predicciones sobre instancias futuras. Entre los principales métodos de aprendizaje supervisado se encuentran las redes neuronales artificiales, los árboles de regresión, la regresión lineal múltiple, el modelo lineal generalizado, los bosques aleatorios, las máquinas de soporte vectorial, entre otros (Osisanwo et al., 2017).

El proceso de elaboración de un modelo de aprendizaje supervisado se constituye bajo las siguientes actividades esenciales, cuyo flujo se explica en la Figura 2.

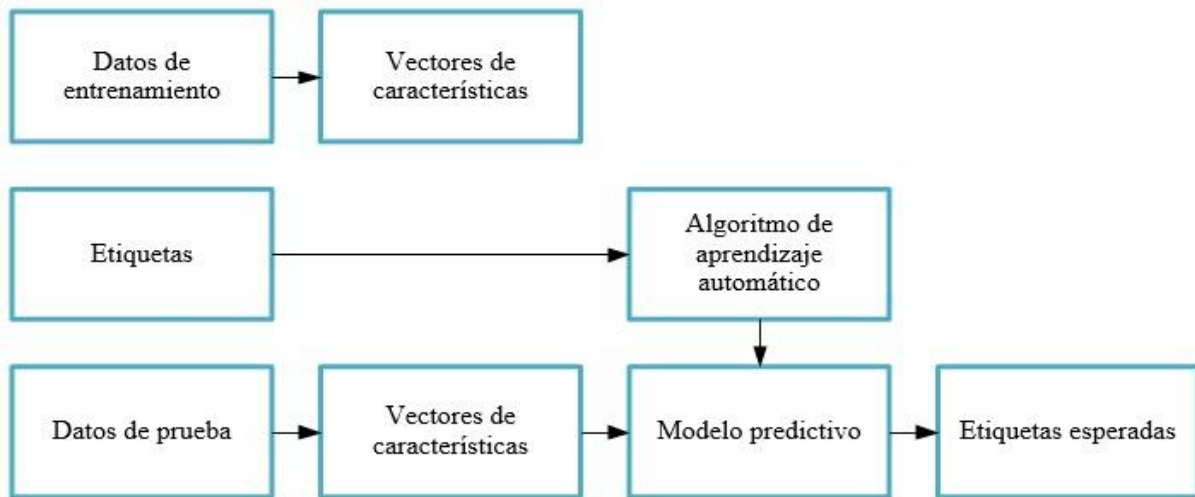


Figura 2. Proceso SML. Modificado de (Muhammad & Yan, 2015).

El aprendizaje automático supervisado tiene como objeto predecir resultados a partir de datos de entrenamiento (suministrados por pares, es decir, para cada dato de entrada existe su salida esperada), bien sea un valor numérico o una etiqueta de clase, y dependiendo de ello corresponde a un problema de regresión o clasificación respectivamente (Muhammad & Yan, 2015).

- Regresión: predecir resultados con una salida continua. Es decir, mapear variables de entrada a alguna función continua.
- Clasificación: predecir resultados con una salida discreta. Dicho de otro modo, mapear variables de entrada en categorías discretas.

5.4. Redes neuronales artificiales

Una red de neuronas artificiales (ANN, por sus siglas en inglés) es un paradigma de procesamiento de información inspirado en el cerebro humano. El elemento clave de este paradigma es su estructura (Hiperparámetros), la cual está compuesta por un cierto número de elementos de procesamiento llamados neuronas que trabajan al unísono para resolver un problema específico. Las redes neuronales actuales se basan en el modelo matemático de neurona propuesto por McCulloch y Pitts en 1943, que se representa en la Figura 3.

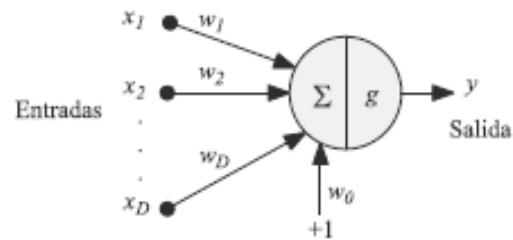


Figura 3. Modelo neuronal McCulloch-Pitts. Modificado de (Palma & Marín, 2008)

En el modelo, cada neurona recibe un conjunto de entradas $\{x_1, x_2, \dots, x_D\}$ y devuelve una única salida y . Además, dentro de una RNA existen numerosas conexiones entre las distintas neuronas que la forman, simulando de esta manera las conexiones interneuronales del cerebro y, al igual que éstas, establecer con mayor o menor intensidad dependiendo de su grado de asociación o bien pesos sinápticos. De este modo, cada entrada x_i de una neurona se encuentra afectada por un peso w_i .

Para obtener la salida y de la neurona se debe calcular la suma ponderada a de las entradas, que corresponde a la activación de la neurona:

$$a = \sum_{i=1}^D w_i x_i + w_0 \quad (1)$$

Donde, w_0 es un umbral, sesgo o bias que se utiliza para compensar la diferencia entre el valor medio de las entradas y el correspondiente valor medio de las salidas deseadas.

La salida y se obtiene a partir de este valor y mediante la aplicación de la función de activación o transferencia $g(a)$, como se muestra en la siguiente ecuación:

$$y = g(a) = g\left(\sum_{i=1}^D w_i x_i + w_0\right) \quad (2)$$

Es importante destacar que en este modelo básico de McCulloch-Pitts, la función de activación empleada es la función escalón definida por la ecuación:

$$g(a) = \begin{cases} 0 & \text{cuando } a < 0 \\ 1 & \text{cuando } a > 0 \end{cases} \quad (3)$$

En los modelos actuales se escogen otro tipo de funciones, normalmente monótonas y derivables, como la función lineal, la sigmoidea, la tangente hiperbólica y la gaussiana, cuyas ecuaciones y graficas se muestran a continuación.

Ecuación función lineal o identidad:

$$g(a) = a \quad (4)$$

Ecuación función sigmoidea o logística:

$$g(a) = \frac{1}{1 + e^{-a}} \quad (5)$$

Ecuación función tangente hiperbólica:

$$g(a) = \tanh = \frac{e^a - e^{-a}}{e^a + e^{-a}} \quad (6)$$

Ecuación función gaussiana:

$$g(a) = \exp\left(-\frac{(a - \mu)^2}{2\sigma^2}\right) \quad (7)$$

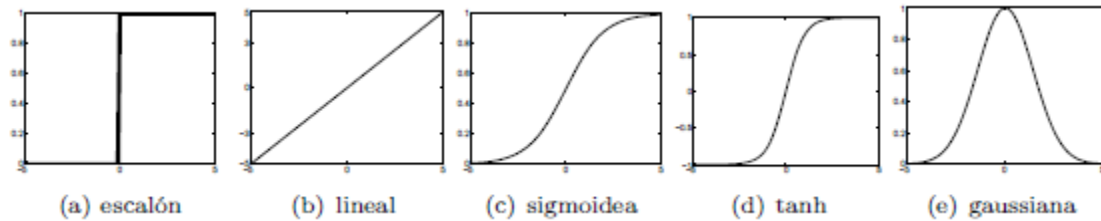


Figura 4. Principales funciones de activación. Modificado de (Palma & Marín, 2008)

Adicionalmente, existe la función de activación Relu (del inglés, Rectified Linear Unit), la cual permite el paso de todos los valores positivos sin cambiarlos, pero asigna todos los valores negativos a 0. Dicha función se comporta de la forma presentada en la Ecuación 8 y su grafica corresponde a la presentada en la Figura 5.

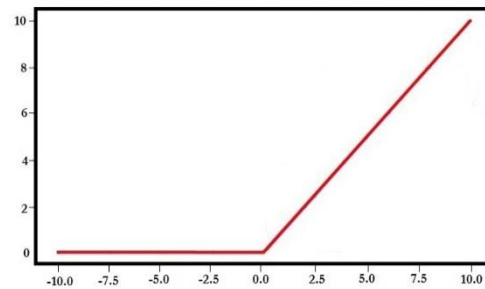


Figura 5. Función de activación Relu.

$$g(a) = \begin{cases} 0 & \text{cuando } a < 0 \\ a & \text{cuando } a \geq 0 \end{cases} \quad (8)$$

Dado que el aprendizaje supervisado se utiliza para modelar la correspondencia entre un conjunto de N ejemplos o patrones de entrada $\{x_1, x_2, \dots, x_N\}$ y las salidas deseadas $\{t_1, t_2, \dots, t_N\}$, y cuyo objetivo es aprender la correspondencia entre ambas, minimizando la función de error que mide la diferencia existente entre estas salidas y las obtenidas por la red $\{y_1, y_2, \dots, y_N\}$. Plasmado de otro modo, el conjunto de entrenamiento estará compuesto por pares de la forma:

$$X = \{(x^n, t^n)\}_{n=1}^N \quad (8)$$

Nuevamente, donde x es el vector de entradas, t el de salidas deseadas y N es el tamaño del conjunto de entrenamiento (Palma & Marín, 2008).

5.5. Perceptrón multicapa

El perceptrón multicapa (MLP, por sus siglas en inglés), surge como una extensión del perceptrón simple (compuesto únicamente por una capa de neuronas de entrada y otra capa de neuronas de salida), dado que representaba demasiadas limitaciones se propusieron entonces redes neuronales con mayor número de capas entre la capa de entrada y la capa de salida, después llamadas capas intermedias cuya función es tratar de realizar una proyección en la que resulten separables linealmente los patrones de entrada de manera que la unidad de salida pueda realizar una clasificación correcta.

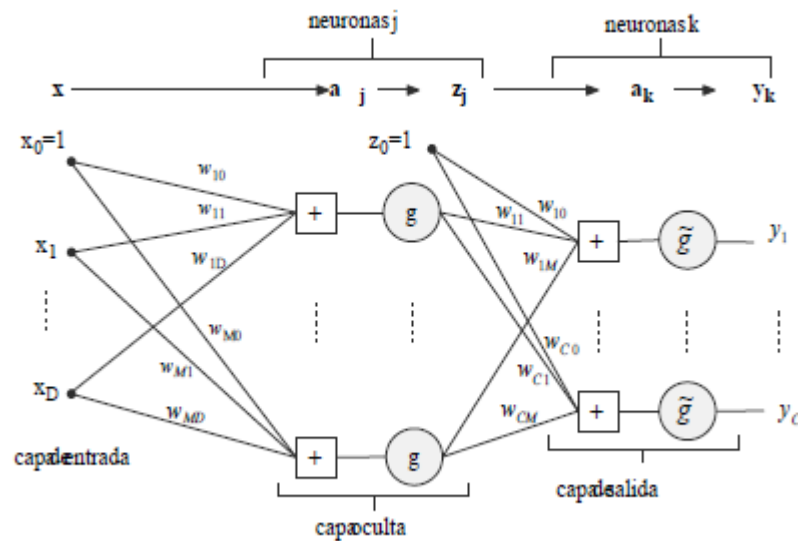


Figura 6. Arquitectura de perceptrón multicapa. Modificado de (Palma & Marín, 2008)

Como se muestra en la Figura 6, dichas capas intermedias se llaman capas ocultas dado que se caracterizan por no tener contacto con el exterior. Ahora partiendo del ejemplo, existen D entradas,

una única capa oculta, M neuronas en dicha capa oculta y C unidades de salida, la salida z_j de cada neurona de la capa oculta se define como:

$$z_j = g(a_j) = g\left(\sum_{i=1}^D w_{ji} x_i + b_j\right); \quad j = 1, 2, \dots, M \quad (9)$$

Donde a_j representa el nivel de activación de la neurona j de la capa oculta y w_{ji} el peso asociado a la neurona j y la entrada x_i y b_j representa el umbral de activación de la neurona oculta j . Ahora, repitiendo el mismo patrón (la suma ponderada de las salidas de las unidades de la capa anterior), entonces la salida de la neurona k de la capa de salida corresponde a:

$$y_k = \tilde{g}(a_k) = \tilde{g}\left(\sum_{j=1}^M w_{kj} z_j + b_k\right) = \tilde{g}\left(\sum_{j=1}^M w_{kj} g\left(\sum_{i=1}^D w_{ji} x_i + b_j\right) + b_k\right); \quad (10)$$

$$k = 1, 2, \dots, C$$

Destacando que la función de activación g suele ser de alguna de las formas presentadas anteriormente, mientras que la forma de la función $\tilde{g}$ depende del tipo de problema a abordar.

5.6. Mecanismo de aprendizaje

Las redes neuronales son capaces de aprender, gracias a que el conocimiento de las mismas se encuentra representado en los pesos de las conexiones entre las neuronas, y dicho aprendizaje implica un número determinado de modificaciones en los pesos sinápticos, dicho de otro modo, la red aprende mediante la modificación de los valores de los pesos de la red. El proceso de aprendizaje finaliza cuando los valores de los pesos permanecen estables ($dw_{ij}/dt = 0$) (Hilera González & Martínez Hernando, 1995).

El MLP basa el aprendizaje de sus pesos en una regla de ajuste del error, esto es, trata de determinar los pesos de las conexiones sinápticas de manera que las salidas de la red coincidan con las salidas deseadas, o por lo menos, sean lo más próximas posibles. Dicho ajuste se realiza comenzando por la capa de salida, según el error cometido, y propagando este error a las capas anteriores, hasta llegar a la capa de las unidades de entrada, de ahí su nombre algoritmo de retropropagación del error (*backpropagation algorithm*). El cual consiste básicamente en dos fases que se repiten hasta conseguir minimizar el error.

En la primera fase o paso hacia adelante, aplica un patrón a las entradas de la red y su efecto se propaga a través de la misma, capa a capa. Finalmente, la red presenta un conjunto de salidas como respuesta a dicho patrón de entrada. En la siguiente fase o paso hacia atrás, los pesos de la red se recalculan de acuerdo con una regla de ajuste del error, en la que se calcula el valor de la función de error que compara la respuesta actual de la red y la respuesta deseada, y este error se propaga hacia atrás (Palma & Marín, 2008).

Para la optimización de los pesos sinápticos se pueden utilizar los algoritmos:

- L-BFGS (de sus siglas en inglés Limited memory BFGS): optimizador de la familia de métodos cuasi-Newton. Permite obtener el mínimo de una función y necesita únicamente la función y su gradiente.
- Gradiente descendiente estocástico (SGD, por sus siglas en inglés): calcula el gradiente de la función de pérdida para un ejemplo en cada iteración, dicho ejemplo se elige de forma aleatoria en cada iteración.
- Adam (del inglés, Adaptive Moment Estimation): calcula un promedio móvil exponencial del gradiente y el gradiente cuadrado.

Adicionalmente, existen hiperparámetros adicionales: el momentum para la actualización de la pendiente del gradiente y la tasa de aprendizaje la cual controla el tamaño de paso en la actualización de los pesos sinápticos.

5.7. Validación de modelos

Para validar la estabilidad de los modelos de aprendizaje automático es necesario asegurarse de que el modelo haya reconocido la mayoría de los patrones en los datos, y que no detecte demasiado ruido. Para ello se requiere la aplicación de métodos de validación.

5.7.1. Holdout. Se divide aleatoriamente el conjunto de datos en dos subconjuntos: el primer subconjunto corresponde al de entrenamiento y el segundo al de prueba. El subconjunto de entrenamiento se utiliza para entrenar el modelo de aprendizaje automático, y el subconjunto de prueba se usa para evaluar el desempeño del respectivo modelo. Generalmente el 80% del total de datos corresponde al conjunto de entrenamiento y el 20% restante al conjunto de prueba.

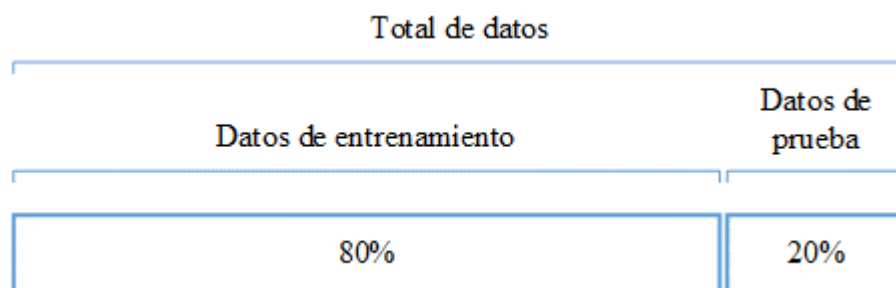


Figura 7. Validación Holdout.

5.7.2. Validación cruzada KFold. La validación cruzada con k Fold divide los datos en k subconjuntos, de manera que el respectivo modelo de aprendizaje se entrena y valida k veces. Cada una de las k veces, uno de los k subconjuntos se usa como conjunto de prueba y los otros $k - 1$ se agrupan para formar el conjunto de entrenamiento. La estimación del error se promedia en todos las k iteraciones.

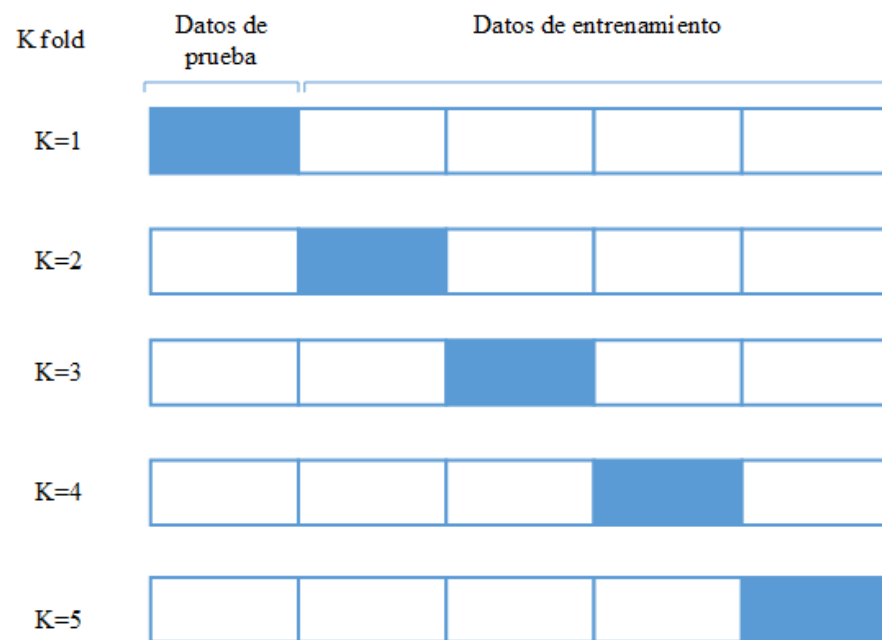


Figura 8. Validación cruzada KFold ($k=5$).

5.8. Ingeniería de características

La ingeniería de características es el proceso de utilizar el conocimiento obtenido a partir de los datos para crear características que hacen que los algoritmos de aprendizaje automático funcionen mejor. Realizada correctamente puede aumentar el poder predictivo de los algoritmos de aprendizaje automático al crear características a partir de datos sin procesar que ayudan a facilitar el proceso de aprendizaje automático (Shekhar, 2018).

El proceso de ingeniería de características suele ser iterativo: aplicar alguna técnica y, posteriormente, evaluar el modelo para ver si las nuevas características contribuyeron a mejorar el desempeño del modelo (Enesco & Keshav, 2015).

Las principales técnicas de ingeniería de características son:

- Binarización: la descomposición de cualquier tipo de atributo en dos grupos.
- Combinación de características mediante cualquier tipo de operación, como la adición o el producto. Donde también se pueden combinar características categóricas.
- Agrupación: Transformación de valores numéricos continuos en categorías con un número finito de intervalos.

5.8.1. Análisis de componentes principales. El análisis de componentes principales (PCA, por sus siglas en inglés) es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones mientras conserva su información.

Supóngase que existe una muestra con n observaciones cada una con p variables ($x_1, x_2, \dots, x_p$), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes ($z < p$) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal (Amat, 2017).

El proceso de aplicar PCA, se resume en los siguientes pasos:

- Estandarizar las variables de entrada.

- Calcular la correlación entre diferentes variables. (La correlación muestra si dos variables tienen una relación. Si están relacionados, y si esta relación es positiva o negativa).
- Calcular los vectores propios y los valores propios utilizando la matriz de correlación.
- Seleccionar las variables de entrada con la máxima variabilidad. (El valor propio determina la magnitud de la variabilidad).

Los vectores propios (o eigenvectors) son resultado de la multiplicación entre una matriz y un vector. El vector resultante de la multiplicación es un múltiplo entero del vector original. Los eigenvectors de una matriz son todos aquellos vectores que, al multiplicarlos por dicha matriz, resultan en el mismo vector o en un múltiplo entero del mismo.

En el caso de los valores propios (o eigenvalue): cuando se multiplica una matriz por alguno de sus eigenvectors se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número. Al valor por el que se multiplica el eigenvector resultante se le conoce como eigenvalue. A todo vector propio le corresponde un valor propio y viceversa.

En PCA, cada una de las componentes principales corresponde con un eigenvector, y el orden de componente se establece por orden decreciente de eigenvalue. Así pues, la primera componente es el eigenvector con el eigenvalue asociado más alto (Amat, 2017). Entonces se escogen las primeras z componentes que expliquen la variabilidad que establezca el investigador.

5.9. Selección de características

La selección de características es el proceso en el que selecciona automática o manualmente las características que más contribuyen a la variable de predicción o resultado en el que se está interesado (Shaikh, 2018).

Sus principales beneficios son:

- Reduce el overfitting (sobreajuste): menos datos redundantes significa menos oportunidad de tomar decisiones basadas en el ruido.
- Mejora la precisión: menos datos engañosos significa que la precisión del modelado mejora.
- Reduce el tiempo de entrenamiento: menos puntos de datos reducen la complejidad del algoritmo y por consiguiente se entrenan más rápido.

Es importante diferenciar la selección de características de la ingeniería de características, dado que la ingeniería de características se aplica primero para generar características adicionales, y luego se implementa la selección de características para eliminar características irrelevantes, redundantes o altamente correlacionadas.

Existen tres clases generales de algoritmos de selección de características, estos son:

- Métodos de filtro (Filter Methods): aplican una medida estadística para asignar una puntuación a cada característica. Las características se clasifican según la puntuación y se seleccionan para mantenerlas o eliminarlas del conjunto de datos.
- Métodos de envoltura (Wrapper Methods): consideran la selección de un conjunto de características como un problema de búsqueda, donde se preparan, evalúan y comparan diferentes combinaciones. Puede ser metódico (una búsqueda mejor), estocástico (algoritmo de ascenso de pendientes aleatorio) o puede usar heurísticas (como pases hacia adelante y hacia atrás para agregar y eliminar características).

- Métodos inmersos (Embedded Methods): aprenden qué características contribuyen mejor a la precisión del modelo mientras se crea el modelo. Su tipo más común son los métodos de penalización, que introducen restricciones adicionales en la optimización de un algoritmo predictivo (como un algoritmo de regresión) que desvía el modelo hacia una menor complejidad (menos coeficientes) (Brownlee, 2014).

A continuación, en las próximas subsecciones se explicarán algunos métodos de selección de características.

5.9.1. Análisis de correlaciones. El coeficiente de correlación es una medida para determinar el grado de asociación lineal entre dos variables continuas. El coeficiente de correlación puede variar desde -1 hasta 1.

La correlación de proporcionalidad directa o positiva se establece con los valores superiores a cero (> 0), y entre más cercano sea el valor a 1 mayor será la correlación positiva, la proporcionalidad inversa o negativa, se da cuando los valores son menores a cero (< 0) y entre más cercano sea el valor a -1 mayor será la correlación negativa. No existe relación entre las variables cuando el coeficiente es igual a 0 (Ramón, 2015).

Entonces el análisis de correlaciones consiste en escoger las variables independientes que mayor correlación tengan con la variable dependiente, sin olvidar, analizar las correlaciones que tienen las variables independientes entre ellas, por lo que no podrían escogerse pares de variables independientes con alta correlación, solo se deberá escoger una de ellas para evitar multicolinealidad.

5.9.2. Análisis de función de regresión. El análisis de función de regresión es un análisis univariante que funciona seleccionando las mejores características basadas en pruebas estadísticas univariadas. El análisis de regresión ayuda a entender cómo el valor de la variable dependiente varía al cambiar el valor de una de las variables independientes, manteniendo el valor de las otras variables independientes fijas, es decir, el valor promedio de la variable dependiente cuando se fijan las variables independientes.

Entonces el análisis de función de regresión consiste en realizar una función de regresión por cada variable independiente y la variable dependiente (Y). Después de evaluar y analizar el efecto individual de cada una de las variables independientes, la función asigna una puntuación a cada variable, en donde las de mayor puntuación, son las que mayor efecto tienen sobre la variable salida Y.

5.9.3. Paso a Paso. La selección de características por paso a paso (o Step Forward Feature Selection) es un método de selección envolvente, es decir, combina el proceso de selección de características con el modelo de predicción evaluando el conjunto de variables para detectar el rendimiento del modelo entre cada una de ellas y, posteriormente, selecciona una a una (en orden) las variables del modelo.

La selección paso a paso comienza con la evaluación de cada característica individual y selecciona aquella que da como resultado el mejor modelo de algoritmo seleccionado (dependiendo de las métricas fijadas por el investigador). Después, se evalúan todas las combinaciones posibles de la característica seleccionada y una característica posterior, y se selecciona una segunda característica, y así sucesivamente, hasta que se seleccione el número predefinido requerido de características (Mayo, 2018).

5.10. Identificación de Hiperparámetros

Todos los algoritmos de aprendizaje automático tienen un conjunto predeterminado de hiperparámetros, definidos como variables de configuración que son internas al modelo y cuyo valor se puede estimar a partir de los datos. Los diferentes algoritmos constan de diferentes hiperparámetros. Por ejemplo, los modelos de regresión regularizados tienen penalizaciones de coeficientes, los árboles de decisión tienen un número determinado de ramas y las redes neuronales tienen un número determinado de capas. En la construcción del algoritmo de aprendizaje se elige la configuración predeterminada de estos hiperparámetros después de ejecutar el modelo en varios conjuntos de datos.

La identificación de hiperparámetros (o Tuning en inglés) suele ser un proceso de prueba y error mediante el cual se cambian algunos hiperparámetros (por ejemplo, la cantidad de árboles en un bosque aleatorio o el valor de alfa en un algoritmo lineal), luego se ejecuta el algoritmo en los datos nuevamente y se compara su rendimiento en su conjunto de validación para determinar qué conjunto de hiperparámetros da como resultado el modelo más preciso. Además, es un proceso importante porque permite personalizar los modelos para que brinden información valiosa sobre los datos, materializándose en decisiones más efectivas (DataRobot, 2018).

5.10.1 Búsqueda Exhaustiva. El algoritmo de optimización Búsqueda Exhaustiva (o Grid Search en inglés) realiza una búsqueda de cuadrícula a través de un subconjunto de hiperparámetros especificados manualmente. La búsqueda de cuadrícula busca sistemáticamente todas las soluciones posibles (posibles combinaciones), y para ello debe guiarse por alguna métrica de rendimiento (establecida por el investigador), que se mide mediante validación cruzada en el conjunto de entrenamiento.

Para su aplicación se debe establecer una rejilla sobre el área de interés y evaluar la métrica establecida en cada nodo de la rejilla. Luego de calcular los valores de la métrica en todos los nodos de la rejilla, es posible interpolar y encontrar un máximo entre las líneas de la cuadrícula.

5.10.2 Algoritmo genético. Los algoritmos genéticos (GA, por sus siglas en inglés), son una metodología de búsqueda de optimización adaptativa general basada en una analogía directa con la selección natural y la genética darwiniana en sistemas biológicos.

El GA trabaja con un conjunto de soluciones candidatas llamadas población, y se basa en el principio darwiniano de "supervivencia del más apto", el GA obtiene la solución óptima después de una serie de cálculos iterativos, genera poblaciones sucesivas de soluciones alternativas que están representadas por un cromosoma, es decir, una solución al problema, hasta que se obtienen resultados aceptables.

Uno de sus más atractivos beneficios es que está asociado a las características de la búsqueda de explotación y exploración, tiene la capacidad de tratar grandes espacios de búsqueda de manera eficiente y, por lo tanto, tiene menos posibilidades de obtener una solución óptima local que otros algoritmos.

Los cromosomas se seleccionan para la reproducción mediante la evaluación del valor fitness (condición física). Los cromosomas en mejor forma tienen mayor probabilidad de ser seleccionados en el grupo de recombinación utilizando la rueda de la ruleta o los métodos de selección de torneos. Después, los cromosomas seleccionados son apareados para reproducir descendientes y estos descendientes tendrán la posibilidad de ser mutados. El valor fitness se evalúa mediante la función Fitness (de aptitud), que valora la calidad de una solución en el paso de evaluación (Huang & Wang, 2006).

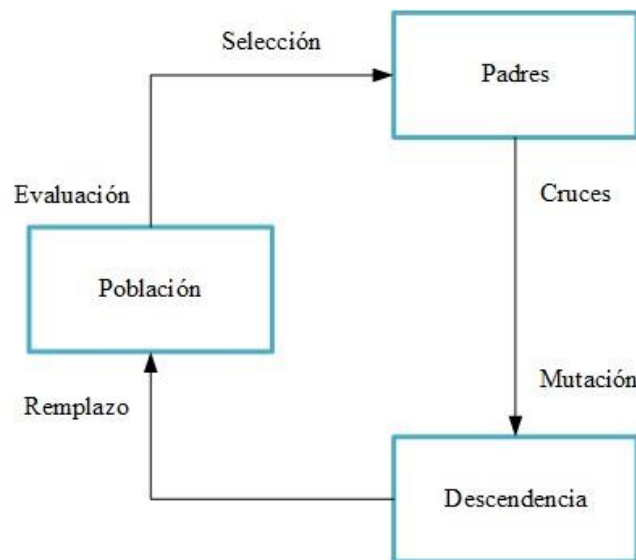


Figura 9. Ciclo evolutivo. Modificado de (Huang & Wang, 2006).

Los operadores genéticos que impactan aleatoriamente el valor fitness son los siguientes:

Selección: operador encargado de escoger qué individuos van a disponer de oportunidades de reproducirse y cuáles no. Puede ser selección por ruleta, es decir, que a cada uno de los individuos de la población se le asigna una parte proporcional a su ajuste de una ruleta, donde los mejores individuos recibirán una porción de la ruleta mayor que la recibida por los peores. O también selección por torneo, que consiste en realizar la selección en base a comparaciones directas entre individuos.

Cruce: permite explorar nuevas regiones de solución en el espacio de búsqueda, es un mecanismo aleatorio para intercambiar genes entre dos cromosomas utilizando el cruce de un punto, el cruce de dos puntos o el cruce de homólogos. Los cruces más empleados son cruce de 1 punto, cruce de 2 puntos y cruce uniforme.

Cruce de 1 punto: se cortan los cromosomas por un punto seleccionado aleatoriamente y se intercambian las colas entre los dos individuos para generar los nuevos descendientes.

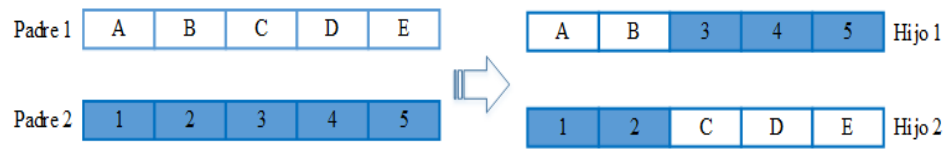


Figura 10. Cruce de 1 punto.

Cruce de 2 puntos: se realizan dos cortes, teniendo en cuenta que ninguno de estos puntos de corte coincida con el extremo de los cromosomas para garantizar que se originen tres segmentos.

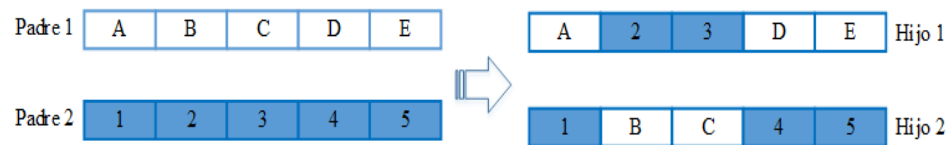


Figura 11. Cruce de 2 puntos.

Cruce uniforme o por máscara: la técnica implica la generación de una máscara de cruce con valores binarios.

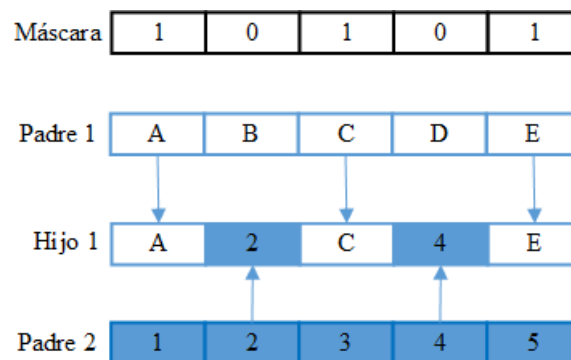


Figura 12. Cruce uniforme.

Si en una de las posiciones de la máscara hay un 1, el gen situado en esa posición en uno de los descendientes se copia del primer padre. Si por el contrario hay un 0 el gen se copia del segundo padre. Para producir el segundo descendiente se intercambian los papeles de los padres, o bien se intercambia la interpretación de los unos y los ceros de la máscara de cruce (Dorado & Gestal, 2010).

Mutación: los genes pueden alterarse ocasionalmente, es decir, provoca que alguno de sus genes, generalmente uno sólo, varíe su valor de forma aleatoria. La probabilidad de mutación es muy baja, generalmente menor al 1%. Esto se debe sobre todo a que los individuos suelen tener un ajuste menor después de mutados. Sin embargo, se realizan mutaciones para garantizar que ningún punto del espacio de búsqueda tenga una probabilidad nula de ser examinado (Dorado & Gestal, 2010).

5.11. Métricas de ajuste

El error absoluto medio (MAE, por sus siglas en inglés) es un indicador del desempeño de pronóstico que mide el tamaño del error absoluto y su fórmula se expresa de la forma:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (11)$$

El error cuadrático medio (MSE, por sus siglas en inglés) es una medida de desempeño que evalúa métodos de pronósticos. Esta medida amplifica aquellos errores de mayor magnitud cuya fórmula es:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12)$$

Para las ecuaciones (11) y (12) presentadas anteriormente, y_i corresponde al valor de variable dependiente real, $\hat{y}_i$ es el valor pronosticado y N es el número de datos.

La raíz del error cuadrático medio (RMSE, por sus siglas en inglés) mide la magnitud del error MSE. Su ecuación está representada en (13).

$$RMSE = \sqrt{MSE} \quad (13)$$

El coeficiente de determinación o R^2 , es la proporción de la varianza total de la variable explicada por la regresión. El resultado del coeficiente de determinación oscila entre 0 y 1, cuanto más cerca este el resultado del 1 mayor será el ajuste y será más confiable. Tiene como expresión la siguiente fórmula:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (14)$$

Donde y_i es la variable dependiente real, $\hat{y}_i$ es la variable dependiente estimada y $\bar{y}$ es la media de los datos.

Dado que para la comparación de los modelos elaborados en el presente proyecto se comparan las métricas descritas en las ecuaciones (11), (12), (13) y (14), se utiliza la siguiente fórmula de mejora relativa porcentual para el coeficiente de determinación.

$$RI_{R^2} = \frac{R^2_{ref} - R^2_{mejorado}}{R^2_{ref}} * 100 \quad (15)$$

6. Metodología

En el presente capítulo se describen de forma detallada la construcción de los modelos de redes neuronales, la implementación de métodos de selección e ingeniería de características a las

respectivas redes, la aplicación del algoritmo genético para la búsqueda de parámetros del segundo modelo de red neuronal y finalmente la validación de dichos modelos de redes neuronales con las métricas de bondad de ajuste mencionadas anteriormente. Cabe resaltar que para esto se utilizó el lenguaje de programación Python en su versión 3.

6.1. Consolidación de la base de datos

Para la realización de la investigación, el conjunto de datos empleado en el desarrollo y validación de los modelos predictivos 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 en Rio Negro, Santander, para los años 2015, 2016 y 2017.

La respectiva base de datos cuenta con 17 variables de entrada y 428 registros para cada una, las variables corresponden a: altura de la planta, área de copa, diámetro del tronco, área foliar no destructiva, fósforo, asimilación de CO_2 , uso eficiente del agua, transpiración, conductividad, potasio, materia orgánica, radiación, temperatura, conductividad eléctrica, lluvias acumuladas, humedad, número de días con lluvia y 1 variable dependiente correspondiente al rendimiento del cultivo (ver Apéndice B). Es importante mencionar que adicional a estas 17 variables, se cuenta con 3 que sirven de caracterización, estas son año (2015, 2016, 2017), bloque (1, 2, 3) y código de tratamiento (1- 60), esta última corresponde a los 60 tratamientos posibles del cultivo que surgen de la combinación de 3 niveles de fertilización (50%, 100%, y 150%), 10 tipos de clones de cacao (SCC19, SCC-52, SCC-61, SCC-64 Y SCC-83, ICS-95, CNN-51, ETT-8, TSH565 y ICS-1) y dos exposiciones (sol y sombra).

6.1.1. Limpieza del conjunto de datos. En primera instancia, para el tratamiento de datos perdidos se realiza una imputación de datos mediante la media de estos, es decir que, en el caso de faltar un dato en alguna de las 17 variables, este será reemplazado por la media de la respectiva variable. Adicionalmente, después de completar los datos perdidos, estos fueron normalizados para que todos se encontraran en el mismo rango (0-1), esta normalización basada en la unidad se realizó de la forma presentada en la Ecuación 16.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (16)$$

Donde:

- X' es el nuevo valor normalizado
- X es el valor a normalizar
- X_{min} es el valor mínimo de la respectiva variable
- X_{max} es el valor máximo de la respectiva variable

El respectivo diagrama de flujo del proceso de limpieza de datos se presenta en la siguiente Figura 13.

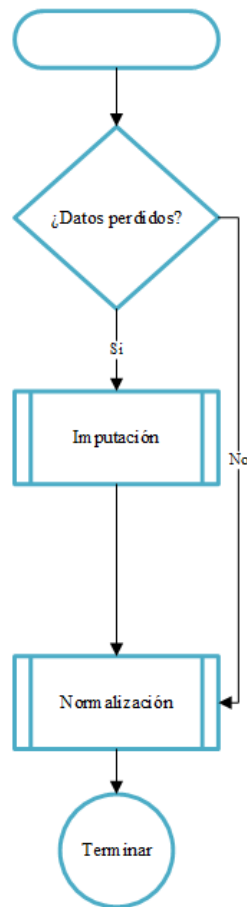


Figura 133. Diagrama de flujo proceso de limpieza de datos.

6.2. Desarrollo del modelo predictivo basado en redes neuronales artificiales

En la construcción del primer modelo de red neuronal se utilizó el lenguaje de programación Python. Lenguaje que se escogió por su simplicidad, su legibilidad, sintaxis amigable para el desarrollador y su semántica, así como la cantidad de bibliotecas y marcos de código abierto disponibles. Para acceder a los códigos de programación del presente proyecto debe consultarse el siguiente enlace: <https://github.com/alejandragoc/OPTIMIZACION-RNA>.

Para la programación del modelo de red neuronal se utilizó el paquete *MLPRegressor* de la librería *sklearn.neural_network*. El primer paso en la programación fue la importación del

conjunto de datos después de haber sido imputados y normalizados. Después se definieron dentro del modelo las variables independientes (x), previamente definidas en el Apéndice B y la respectiva variable independiente (y) correspondiente al Rendimiento Total del cultivo.

Seguido de esto se empleó *Cross Validation* como método para evaluar la exactitud, con $k=5$ (particiones del conjunto de datos) y se barajaron los datos antes de realizar las respectivas particiones, buscando que los datos de entrenamiento y prueba fueran escogidos de forma aleatoria.

Finalmente, para la selección de la arquitectura de la red se hizo un ejercicio de prueba y error variando solo uno de sus hiperparámetros al tiempo, y manteniendo los demás constantes. Buscando maximizar el coeficiente de determinación y minimizar los errores. Los hiperparámetros escogidos para la primera red fueron los presentados en la siguiente Tabla 2.

Tabla 2.

Hiperparámetros primer modelo de red neuronal

Hiperparámetro	Valor
Solver	Adam
Función de activación	Relu
Número de capas ocultas	1
Número de neuronas en capa oculta	10
Tasa de aprendizaje	0.025
Momentum	0.09

Además, cabe resaltar que la red está completamente interconectada, es decir, que todas las neuronas de la capa de entrada, están conectadas con todas las neuronas de la capa oculta.

En el caso de la neurona 1 de la capa oculta los pesos y el umbral (b_j) asociado a ella se presentan en la Tabla 3. Donde W_{ji} es el peso asociado a la neurona j (en este caso $j=1$) y la entrada x_i . Para consultar los coeficientes de todas las neuronas ocultas ver el Apéndice C.

Tabla 3.

Pesos y umbral asociado a la neurona oculta 1

Coefficiente	Valor
$W_{1,1}$	-0.000351
$W_{1,2}$	-0.000399
$W_{1,3}$	0.001321
$W_{1,4}$	-0.006548
$W_{1,5}$	0.006195
$W_{1,6}$	-0.006464
$W_{1,7}$	0.234004
$W_{1,8}$	0.006595
$W_{1,9}$	0.005937
$W_{1,10}$	-0.006012
$W_{1,11}$	0.002624
$W_{1,12}$	-0.003021
$W_{1,13}$	-0.006480
$W_{1,14}$	0.000283
$W_{1,15}$	0.000674
$W_{1,16}$	-0.007028

Coefficiente	Valor
$W_{1,17}$	0.001045
b_1	-0.172192

Entonces, la salida (z_1) asociada a la neurona oculta 1 se presenta en la Ecuación 17, donde g corresponde a la función de activación Relu.

$$\begin{aligned}
 z_1 = & g(-0.000351x_1 - 0.000399x_2 + 0.001312x_3 - 0.006548x_4 \\
 & + 0.006195x_5 - 0.006464x_6 + 0.234004x_7 + 0.006595x_8 \\
 & + 0.005937x_9 - 0.006012x_{10} + 0.002624x_{11} - 0.003021x_{12} \\
 & - 0.006480x_{13} + 0.000283x_{14} + 0.000674x_{15} - 0.007028x_{16} \\
 & + 0.001045x_{17} - 0.172192)
 \end{aligned} \tag{17}$$

La neurona oculta 1 se puede representar gráficamente como se muestra en la Figura 14.

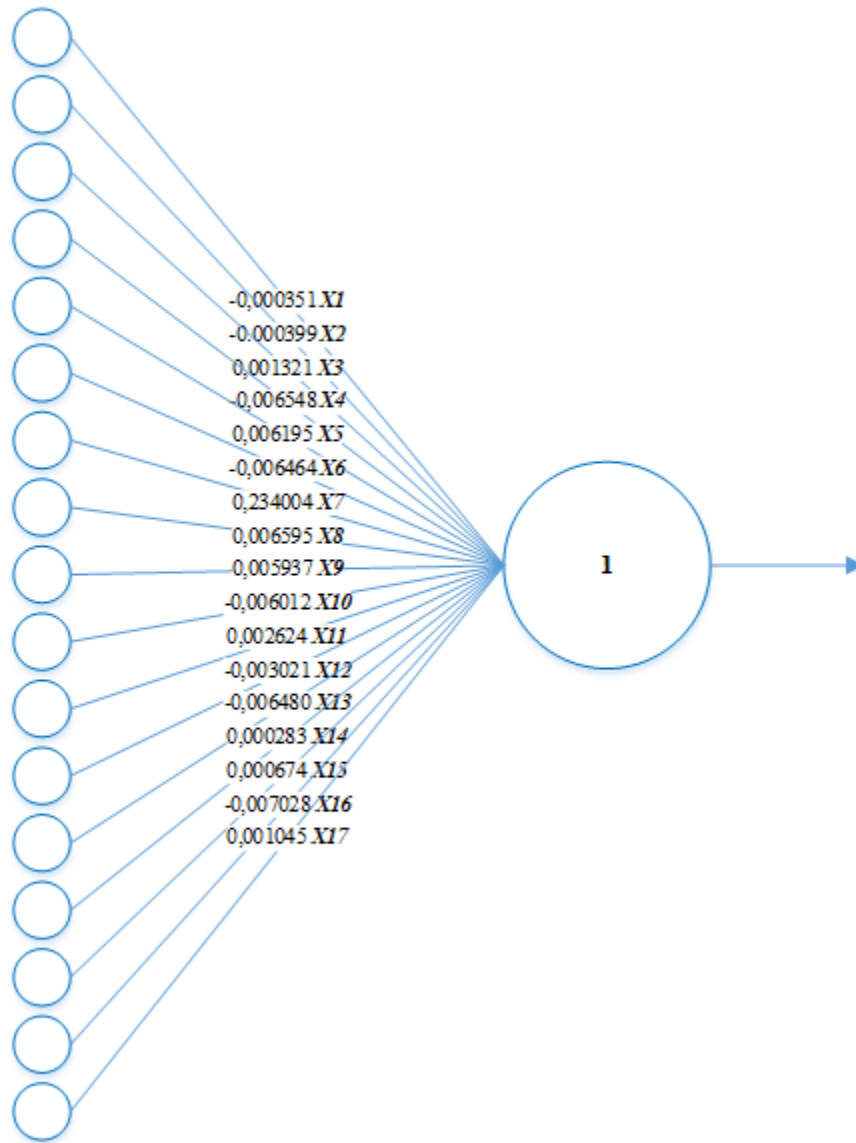


Figura 14. Pesos asociados a la neurona oculta 1.

De forma similar, la neurona de salida tiene unos pesos asociados y un umbral de activación (b_k) que se presentan en la siguiente Tabla 4. Donde W_{kj} es el peso asociado a la neurona de salida k ($k=1$) y la neurona oculta j .

Tabla 4.

Pesos y umbral asociado a la neurona de salida

Coefficiente	Valor
$W_{1,1}$	-0.002789
$W_{1,2}$	-0.269897
$W_{1,3}$	0.166310
$W_{1,4}$	-0.411917
$W_{1,5}$	-0.203027
$W_{1,6}$	0.292928
$W_{1,7}$	-0.267516
$W_{1,8}$	0.078721
$W_{1,9}$	-0.058333
$W_{1,10}$	0.459034
b_1	-0.151227

Entonces, la salida y (Predicción del rendimiento) asociada a la neurona de salida se presenta en la Ecuación 18, donde g corresponde a la función de activación Relu.

$$\begin{aligned}
 y = g & \left(-0.002789z_1 - 0.269897z_2 + 0.166310z_3 - 0.411917z_4 - 0.203027z_5 \right. \\
 & \left. + 0.292928z_6 - 0.267516z_7 + 0.078721z_8 - 0.058333z_9 \right. \\
 & \left. + 0.459034z_{10} - 0.151227 \right)
 \end{aligned} \tag{18}$$

La representación gráfica de la neurona de salida se puede representar gráficamente como se muestra en la Figura 15.

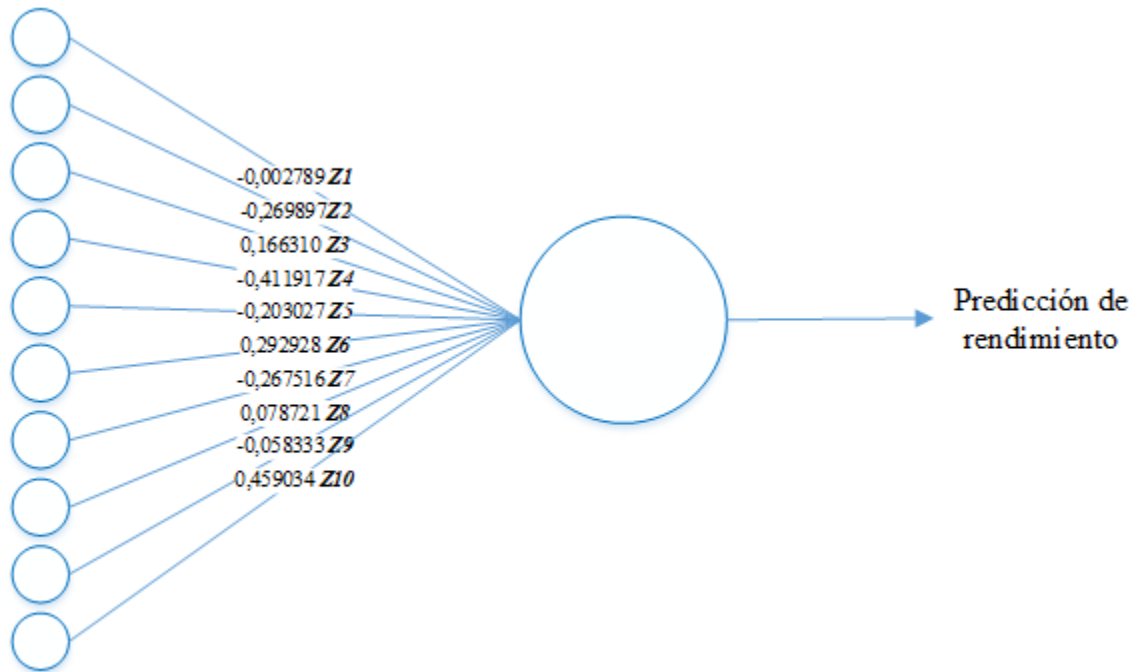


Figura 15. Pesos asociados a la neurona de salida.

6.3. Aplicación de métodos de selección e ingeniería de características

A continuación, se presentan los cuatro métodos empleados para la selección e ingeniería de características (variables de entrada) del modelo predictivo. Los cuatro métodos seleccionados fueron, en primera instancia un análisis de correlaciones entre las variables, en un segundo momento un análisis univariante por función de regresión, después un análisis por componentes principales (PCA, por sus siglas en inglés) y finalmente se implementa una selección paso a paso. Para cada uno de los subconjuntos de variables seleccionadas (4 subconjuntos, uno por cada método de selección) se elaboró la red neuronal propuesta en la Sección 6.2.1, y se aplicaron las métricas Coeficiente de determinación, Error cuadrático medio, Raíz del error cuadrático medio y

Error absoluto medio (definidas en la sección 5.11) con el fin de medir el impacto que tiene la selección e ingeniería de características.

6.3.1. Análisis de correlaciones. Para el análisis de correlaciones se elaboró una matriz de correlaciones con las 17 variables y su respectiva salida para saber el grado de asociación que existe entre ellos (tabla 5). De las cuales se seleccionaron las 6 variables que tenían mayor relación con la salida (Y).

Tabla 5.

Correlaciones de variables independientes con la variable dependiente y

Variable	Valor de correlación con Y
Altura planta	0.144431
Área de copa	0.085360
Diámetro tronco	0.191950
Área foliar no destructiva	0.038321
Fósforo (P)	-0.066231
Asimilación de CO ₂	-0.052186
Uso eficiente del agua	-0.088342
Transpiración	-0.006396
Conductividad	-0.016926
Potasio (K)	-0.033756
Materia Orgánica	0.044674
Radiación	0.202018
Temperatura	-0.199650
Conductividad eléctrica	0.154516

Variable	Valor de correlación con Y
Lluvias acumuladas	-0.011522
Humedad	-0.083938
Número de días con lluvia	0.197703

Adicionalmente, se analizaron las correlaciones entre las 6 variables, que dio como resultado la siguiente matriz (Tabla 6), en donde se puede observar que las variables: Altura de la planta y Diámetro de tronco; Conductividad eléctrica y Número de días con lluvia son variables que presentan un alto grado de correlación. Por lo tanto, se excluyeron Altura de la planta y Conductividad eléctrica, variables que presentaban menor grado de correlación con la salida, para finalmente escoger 4 variables por este método de selección (Radiación, Temperatura, Número de días con lluvia y Diámetro tronco).

Tabla 6.

Correlaciones entre las 6 primeras variables

	Altura planta	Diámetro tronco	Radiación	Temperatura	Conductividad eléctrica	Número de días con lluvia
Altura planta	1	0.66323	-0.26483	-0.18232	0.06399	-0.18125
Diámetro tronco	0.66323	1	-0.12656	-0.17598	-0.32402	-0.51903
Radiación	-0.26483	-0.12656	1	-0.28083	0.41238	0.42765
Temperatura	-0.18232	-0.17598	-0.28083	1	-0.50465	-0.12493
Conductividad eléctrica	0.06399	-0.32402	0.41238	-0.50465	1	0.71548
Número de días con lluvia	-0.18125	-0.51903	0.42765	-0.12493	0.71548	1

6.3.2. Análisis univariante por función de regresión. Para el análisis por función de regresión, se analiza el efecto individual que tiene cada variable sobre la salida (Y). Después de evaluar el efecto individual, la función asigna una puntuación a cada variable, en donde las de mayor puntuación, son las que mayor efecto tienen sobre la variable salida Y (Rendimiento).

Este análisis se realizó utilizando el paquete *f_regression* de la librería *sklearn.feature_selection*. Los valores p obtenidos de este análisis se presentan en la siguiente

Tabla 7.

Valores de puntuación de función de regresión lineal

Variable	Puntuación
Altura planta	0.557570
Área de copa	0.241485
Diámetro tronco	0.912540
Área foliar no destructiva	0.079986
Fósforo (P)	0.166729
Asimilación de CO ₂	0.119872
Uso eficiente del agua	0.254307
Transpiración	0.010483
Conductividad	0.030144
Potasio (K)	0.068188
Materia Orgánica	0.097494
Radiación	1
Temperatura	0.979029
Conductividad eléctrica	0.625121

Variable	Puntuación
Lluvias acumuladas	0.019673
Humedad	0.235484
Número de días con lluvia	0.961977

En este caso, se seleccionaron las 6 variables de puntuación superior a 0.5: Radiación, Temperatura, Número de días con lluvia, Diámetro del tronco, Conductividad eléctrica y Altura de la planta.

6.3.3. Análisis por componentes principales. El análisis por componentes principales fue utilizado para reducir las dimensiones del conjunto de datos de entrada seleccionando las características más importantes que capturan la máxima información sobre el conjunto de datos. Para poder obtener ésta información el conjunto original de variables es transformado en otro conjunto de nuevas variables no correlacionadas (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales.

Este análisis se realizó con la librería *PCA* de la librería *sklearn.decomposition*. De este análisis por componentes principales, los componentes principales y su respectiva varianza explicada se muestran en la siguiente Tabla 8.

Tabla 8.

Varianza explicada por componente principal

Componente	Varianza Explicada	Varianza Explicada Acumulada
1	41.39922	41.39922
2	16.88779	58.28701

Componente	Varianza Explicada	Varianza Explicada Acumulada
3	13.81995	72.10696
4	10.86917	82.97612
5	5.03267	88.00880
6	3.32849	91.33728
7	2.75909	94.09638
8	2.23143	96.32781
9	1.28232	97.61013
10	0.73906	98.34919
11	0.71105	99.06024
12	0.35982	99.42006
13	0.23957	99.65963
14	0.13661	99.79624
15	0.09715	99.89339
16	0.10661	100
17	0	100

Del análisis de PCA se excluyeron los últimos 9 componentes y se escogieron los 8 primeros componentes que acumulaban una varianza explicada superior al 95%.

6.3.4. Selección paso a paso. La selección paso a paso se escogió como último método de selección, dado que, al ser un proceso de selección envolvente, a diferencia de los mencionados anteriormente no es un proceso independiente que tenga lugar antes de la construcción del modelo, sino que, por el contrario, combina el proceso de selección de características con el modelo de red neuronal evaluando el conjunto de características para detectar el rendimiento del modelo entre las características y, posteriormente, selecciona una a una (en orden) las variables del modelo.

Para la selección paso a paso se fijó 10 como el número de iteraciones, es decir, el proceso de seleccionar las variables una a una (en orden) se realizó diez veces y los resultados de dichas iteraciones se presenten en la Tabla 9.

Tabla 9.

Posiciones variables en 10 iteraciones

Iteración				Variable														
1	17	15	11	7	9	1	8	6	4	5	13	16	3	2	10	14	12	
2	3	5	8	13	6	10	17	2	7	4	11	16	15	1	14	12	9	
3	12	3	13	9	14	6	5	16	11	17	1	10	7	15	2	8	4	
4	17	15	12	3	2	5	8	7	11	13	10	4	9	16	14	6	1	
5	17	15	12	13	14	4	7	5	3	10	8	16	2	11	6	1	9	
6	17	3	4	13	14	7	9	12	5	2	6	11	15	16	10	8	1	
7	17	3	4	1	6	10	15	14	12	13	16	2	9	5	7	11	8	
8	17	15	12	13	16	10	11	5	8	9	3	4	7	14	1	2	6	
9	3	12	17	15	8	14	11	4	9	7	2	16	13	5	6	1	10	
10	17	3	4	5	12	7	15	1	16	14	6	8	9	10	2	13	11	

En el caso de la primera iteración, el modelo escogió en primer lugar a la variable 17 (Número de días con lluvia), en segundo lugar, a la variable 15 (Lluvias acumuladas), en tercer lugar a la variable 11 (Materia orgánica) y así sucesivamente. Para poder interpretar la tabla anterior es necesario mencionar que cada variable presentados en la siguiente Tabla 10.

Tabla 10.

Valores de variables

Valor	Variable
1	Altura planta
2	Área de copa
3	Diámetro tronco
4	Área foliar no destructiva
5	Fósforo (P)
6	Asimilación de CO ₂
7	Uso eficiente del agua
8	Transpiración
9	Conductividad
10	Potasio (K)
11	MO
12	Radiación
13	Temperatura
14	Conductividad eléctrica
15	Lluvias acumuladas
16	Humedad

Para escoger las variables resultado de la selección paso a paso, se analizó la frecuencia en la que cada variable fue seleccionada en determinada posición, por ejemplo, la variable 16 (Número de días con lluvia) fue escogida 7 veces en la primera posición. La tabla de frecuencias de la selección paso a paso se presenta en el Apéndice D. Del análisis de la tabla de frecuencias, se escogieron finalmente 6 variables: Número de días con lluvia, Diámetro del tronco, Radiación, Temperatura, Fósforo (P) y Área foliar no destructiva.

Finalmente, el modelo de red neuronal propuesto en la Sección 6.2.1 fue evaluado con el cada uno de los subconjuntos de variables seleccionadas en cada uno de los métodos de selección, de los cuales se observa que el análisis de correlaciones, el análisis de función de regresión lineal y la selección paso a paso coinciden en escoger como variables importantes: Radiación, Temperatura, Número de días con lluvia y el Diámetro del tronco. Los resultados de esta evaluación se presentan en el Capítulo de Resultados.

6.4. Optimización del modelo predictivo

En este capítulo se aplicaran dos algoritmos de búsqueda para la optimización del modelo predictivo

6.4.1. Desarrollo del Algoritmo Búsqueda Exhaustiva. Con el propósito de analizar un modelo de optimización adicional al genético y poder identificar las diferencias en el funcionamiento de dos algoritmos, se elaboró una búsqueda exhaustiva de los parámetros de la red neuronal. Para la realización de dicha búsqueda se utilizó el paquete *GridSearchCV* de la librería *sklearn.model_selection*, la cual implementa *Cross Validation* para la búsqueda de los parámetros, por lo tanto se establecieron 5 particiones para realizar la validación cruzada. Además, a criterio de los investigadores, las variables de entrada del modelo de red neuronal, son las seleccionadas por paso a paso por ser el único método de selección envolvente.

Los posibles valores definidos para cada hiperparámetro se definieron como se muestra en la Tabla 11 para conformar un espacio de búsqueda de 1260 combinaciones de red.

Tabla 11.

Posibles valores para Búsqueda Exhaustiva.

Hiperparámetro	Posibles Valores							
Solver	L-BFGS		Gradiente descendiente estocástico				ADAM	
Función de activación	Relu		Tangente Hiperbólica				Logística	
Número de neuronas en capa oculta 1	4	5	6	7	8	9	10	
Número de neuronas en capa oculta 2	1		2	3	4		5	
Momentum		0.09			0.9			
Tasa de aprendizaje		0.025			0.05			

Además, el criterio de búsqueda utilizado fue el MAE, es decir, el algoritmo seleccionó la arquitectura que minimizaba este error. La arquitectura optima seleccionada por el algoritmo fue la presentada en la Tabla 12.

Tabla 12.

Solución por Búsqueda Exhaustiva.

Hiperparámetro	Valor
Solver	L-BFGS
Función de activación	Relu
Número de neuronas en capa oculta 1	6
Número de neuronas en capa oculta 2	3
Tasa de aprendizaje	0.025
Momentum	0.09

Las métricas de bondad de ajuste para la arquitectura de red seleccionada por búsqueda exhaustiva se presentan en el Capítulo de Resultados.

6.4.2. Desarrollo del algoritmo genético. Para el desarrollo del algoritmo genético se fijaron parámetros propios de este algoritmo como: la tasa de retención, selección y mutación, el tipo de cruce, el tamaño de la población y el número de generaciones los cuales se presentan a continuación, cuyos valores se escogieron a partir de la literatura consultada (Dorado & Gestal, 2010).

Tabla 13.

Parámetros de búsqueda Algoritmo Genético

Parámetros	Valor
Tasa de retención	0.2
Tasa de selección	0.1

Parámetros	Valor
Tasa de mutación	0.01
Tamaño de la población	20
Número de generaciones	20
Cruce	Máscara Aleatoria
Función Fitness	MAE

Aclarando que la tasa de retención corresponde al porcentaje de individuos más aptos que pasarán a la siguiente generación, la tasa de selección se aplica a los individuos menos aptos como una probabilidad para seleccionarlos y pasar a la siguiente generación y la tasa de mutación como una probabilidad de mutar a los individuos seleccionados como padres de una generación, todas las anteriores con el propósito de promover la diversidad.

Adicionalmente, el espacio de búsqueda fijado para este algoritmo es el siguiente:

Tabla 14.

Posibles valores para Algoritmo Genético

Hiperparámetro	Posibles Valores
Solver	L-BFGS Gradiente descendiente estocástico ADAM
Función de activación	Relu Tangente Hiperbólica Logística Identidad
Número de neuronas en capa oculta 1	1 a 15
Número de neuronas en capa oculta 2	1 a 10
Momentum	Entre 0.09 y 0.9
Tasa de aprendizaje	Entre 0.025 y 0.05

De los parámetros del algoritmo genético presentados en la Tabla 14, la arquitectura óptima seleccionada se describe a continuación y sus respectivas métricas se presentan en el Capítulo de resultados.

Tabla 15.

Solución por Algoritmo Genético

Hiperparámetro	Valor
Solver	L-BFGS
Función de activación	Logística
Número de neuronas en capa oculta 1	9
Número de neuronas en capa oculta 2	10
Tasa de aprendizaje	0.0342
Momentum	0.4371

7. Resultados

En este capítulo se presentan los resultados obtenidos a partir de las actividades descritas anteriormente.

7.1. Métricas de ajuste para el primer modelo de red neuronal

Luego de realizar la programación para la construcción del primer modelo de red neuronal con el total de variables (las 17 descritas en el Apéndice B), los hiperparámetros seleccionados y la técnica de validación *Cross Validation*, presentados en la Sección 6.2, se obtienen los siguientes valores de las métricas de ajuste como media de las 5 iteraciones realizadas por la técnica de validación.

Tabla 16.

Métricas de ajuste Primer modelo de red neuronal

Métrica de ajuste	Valor
R²	0,061375
MSE	0,021939
RMSE	0,147537
MAE	0,113636

7.2. Métricas de ajuste aplicando selección e ingeniería de características

A continuación, se presentan las métricas de ajuste de los cuatro métodos de selección e ingeniería de características empleados en el estudio (análisis de correlaciones, análisis de función de regresión, PCA y paso a paso) con su respectivo número de variables de entrada seleccionadas (X). Cabe resaltar que estos resultados se obtienen a partir de la media de las 5 iteraciones realizadas mediante la técnica de validación cruzada. Además, la arquitectura de la red usada en la aplicación de los métodos es la descrita en la Sección 6.2, es decir, la misma del primer modelo de red neuronal.

Tabla 17.

Métricas de ajuste según método de selección

Métrica de ajuste	Método de selección			
	Análisis de correlaciones	Análisis de función de regresión	PCA	Paso a paso
	X= 4	X= 6	X= 8	X= 6
R²	0,106835	0,139419	0,087410	0,120449
MSE	0,021096	0,020224	0,021272	0,020774

RMSE	0,144392	0,141507	0,145308	0,143157
MAE	0,109714	0,105878	0,111148	0,109928

De la

tabla anterior, se observa que el coeficiente de determinación R^2 de la red aplicando cualquiera de los cuatro métodos de selección e ingeniería de características arroja un valor mayor comparado con el coeficiente de determinación R^2 del primer modelo de red presentado en la sección 7.1, además de arrojar valores menores para MSE, RMSE y MAE.

Recordando que el R^2 del primer modelo de red neuronal fue de 0.061375, los métodos, análisis de función de regresión y paso a paso son los métodos que arrojan mejor resultado. Sin embargo, para estudios posteriores se selecciona el método paso a paso ya que es el único método que unifica el proceso de selección de características con el modelo de red neuronal, es decir, para la optimización de los hiperparámetros (Grid Search y Algoritmo Genético) se trabajará con las variables seleccionadas por este método (Diámetro tronco, A.F no destructiva, P, radiación, temperatura y número de días con lluvia).

7.3 Métricas de ajuste para los modelos de red optimizados

En esta sección se presentan los valores de las métricas de ajuste de la construcción de dos modelos de red neuronal, cada uno con los respectivos parámetros que arrojó el algoritmo de búsqueda exhaustiva (Grid Search) y el algoritmo genético, mencionados en la Sección 6.4. Cabe resaltar que sólo se modificara la arquitectura de la red, es decir, las variables de entrada (diámetro tronco, A.F no destructiva, P, radiación, temperatura y número de días con lluvia) y la técnica de validación (Cross Validation) será la misma para los modelos.

7.3.1. Algoritmo Búsqueda Exhaustiva. La siguiente tabla muestra las métricas de ajuste para un modelo de red neuronal con los hiperparametros que arrojó la búsqueda exhaustiva, descritos en la Sección 6.4.1. El resultado es la media de las iteraciones de la técnica Cross Validation.

Tabla 18.

Métricas de ajuste solución por Búsqueda Exhaustiva

Métrica de ajuste	Valor
R^2	0,172388
MSE	0,019498
RMSE	0,139336
MAE	0,104017

7.3.2. Algoritmo Genético. En la presente subsección, se muestran los valores de las métricas de ajuste del modelo de red neuronal con la arquitectura óptima que selecciono el algoritmo genético especificada en la Sección 6.4.2 y las 6 variables seleccionadas por Paso a Paso.

Tabla 19.

Métricas de ajuste solución por Algoritmo Genético

Métrica de ajuste	Valor
R^2	0,175335
MSE	0,019252
RMSE	0,138165
MAE	0,103197

7.3.3. Comparación de los algoritmos de búsqueda. La comparación de los algoritmos de búsqueda se basa en el tiempo computacional (segundos) que requirió cada algoritmo para emitir un resultado y en el ajuste que tiene el modelo. En la siguiente tabla se presenta la información.

Tabla 20.

Tiempos computacionales

Algoritmo	R^2	Tiempo computacional (seg)
Algoritmo de búsqueda exhaustiva	0,172388	88,3987
Algoritmo genético	0,175335	184,0253

De la información en la tabla anterior se puede mencionar que el algoritmo genético tiene un ajuste 1,7095 % mayor que el algoritmo de búsqueda exhaustiva (Grid Search) y requiere 95,6266 segundos más en tiempo computacional. Sin embargo, se selecciona el algoritmo genético como el mejor algoritmo de búsqueda para el caso de esta investigación ya que tiene mejor ajuste y, además, no tiene tantas limitaciones en el espacio de búsqueda como si lo tiene el Grid Search.

7.4. Comparación de los modelos de red neuronal

A continuación, se presentan los valores de las métricas de ajuste para cada uno de los modelos de red mencionados anteriormente. El modelo de red sin optimizar (17 variables de entrada y arquitectura a prueba y error), el modelo de red con el método de selección paso a paso (6 variables de entrada y arquitectura a prueba y error) y por último, red algoritmo genético (17 variables de entrada y arquitectura que selecciono el algoritmo). Esto con el fin de conocer que tiene más efecto sobre el ajuste, si la selección e ingeniería de características o la construcción de la red con la arquitectura óptima.

Tabla 21.

Comparaciones métricas de ajuste

Métrica	Red	Red	Red
	Sin optimizar	Paso a Paso	Algoritmo Genético
R^2	0,061375	0,120449	0,111537
MSE	0,021939	0,020774	0,020705
RMSE	0,147537	0,143157	0,143387
MAE	0,113636	0,109928	0,108493

En la siguiente tabla se relaciona el incremento porcentual en el coeficiente de determinación R^2 de los modelos (con selección de variables paso a paso y optimización de parámetros del algoritmo genético con 17 variables) con respecto al modelo de red neuronal sin optimizar.

Tabla 22.

Incremento porcentual R^2

Métrica	Referencia	Incremento	
	Red	Red	Red
R^2	0,061375	96,2509%	81,7303%

En la tabla anterior se observa que, lo que más influye en el ajuste del modelo predictivo es la correcta selección e ingeniería de características.

7.5. Comparación del modelo red neuronal antes y después de optimización

Como se mencionó anteriormente la optimización del modelo de red neuronal se hizo mediante métodos de selección e ingeniería de características y la aplicación del algoritmo genético para

establecer su arquitectura. En esta sección se realiza la comparación de métricas del modelo de red neuronal sin optimizar con el modelo de red neuronal optimizado (6 variables de entrada y la arquitectura óptima del algoritmo genético) a través de la validación de las métricas de ajuste. El incremento se calcula con respecto a la diferencia entre los valores de la red optimizada y los valores de la red sin optimizar.

Tabla 23.

Comparaciones métricas de ajuste primer modelo y modelo optimizado

Métrica de ajuste	Red sin optimizar	Red optimizada	Incremento (%)
R²	0,061375	0,175335	185,6782
MSE	0,021939	0,019252	12,24759
RMSE	0,147537	0,138165	6,35230
MAE	0,113636	0,103197	9,18634

De la información anterior, todas las métricas de ajuste, tanto los errores como el coeficiente de determinación R^2 fueron mejores en la red optimizada, es decir, el coeficiente de determinación aumentó mientras que MSE, RMSE y MAE disminuyeron. Y el incremento en el R^2 fue de 185,67%

8. Conclusiones

De la revisión de literatura consultada se evidenció la importancia que tienen las técnicas de Aprendizaje Automático y en particular las redes neuronales artificiales sobre la predicción de rendimiento de diferentes cultivos (maíz, cebada, tomate, trigo, soja, caña de azúcar y arroz). Y particularmente se evidenciaron casos de estudio en los cuales las redes neuronales permitían obtener un ajuste superior a otros modelos al no requerir que los datos sigan alguna distribución específica.

Aunque Colombia cuenta con riquezas naturales para potenciar su agricultura, y se ha incentivado el uso de nuevas tecnologías, sigue siendo menos competitiva frente a otros países debido a la falta de investigación que facilite la toma de decisiones basada en estudios no tradicionales y no en la experiencia propia del agricultor, dado que el rendimiento de los cultivos varía espacial y temporalmente con un comportamiento no lineal.

Dado que la toma de datos es un proceso complejo y sujeto a errores, la limpieza y preprocesamiento de la base de datos es fundamental para llenar posibles vacíos. Además, aunque las redes neuronales artificiales no exigen datos normalizados para su funcionamiento, se pudo observar en el caso de estudio con múltiples variables y diferentes rangos para cada una, que la normalización de la base de datos contribuía en un aumento del coeficiente de determinación.

Pese a que la validación de los modelos mediante la técnica Holdout en ocasiones arroja mejores resultados que Crossvalidation, la revisión de literatura permitió reconocer que para estudios con pocos datos se debe descartar Holdout por ser una técnica demasiado optimista que no utiliza todos los datos para prueba y entrenamiento en múltiples iteraciones, si no que por el contrario puede

sesgar los datos, al solo realizar una única iteración con un conjunto fijo de datos para entrenamiento y prueba.

Aunque muchas investigaciones enfocadas a la predicción utilizan redes neuronales, la elección de sus hiperparámetros (solver, función de activación, número de neuronas ocultas, tasa de aprendizaje, momentum, etc.) se basa en un juicio del investigador. Por lo tanto, la aplicación de un algoritmo genético para la búsqueda de los mismos permite que este proceso sea de forma objetiva y elimina la subjetividad de hacerlo mediante prueba y error.

Al realizar cuatro modelos de selección e ingeniería de características, se concluye que el análisis de correlaciones y regresión aun dejan gran parte de la tarea al investigador, además de que son procesos independientes del modelo (ocurren antes del mismo), como también sucede en el análisis de componentes principales, cuya principal desventaja es la perdida de información de las variables de entrada al crear los componentes. Por otro lado, la selección paso a paso destaca sobre los métodos anteriores al ser un método envolvente que prueba el desempeño de la red al mismo tiempo que prueba diferentes combinaciones de las variables y cuya aplicación mejoró el desempeño de la red un 96,25%. Además, al análisis de correlaciones, el análisis de función de regresión lineal y la selección paso a paso coinciden en escoger como variables importantes: Radiación, Temperatura, Número de días con lluvia y el Diámetro del tronco.

De la búsqueda exhaustiva (o Grid) se identificaron limitaciones en cuanto al espacio de búsqueda, ya que solo permitió una cuadrícula de 1260 posibles combinaciones de red. Por ejemplo, para la función de activación solo se pudo considerar tres de cuatro funciones posibles, así como solo dos opciones para momentum y tasa de aprendizaje. También es importante mencionar que la programación de la búsqueda Grid fue sencilla ya que Python posee una biblioteca definida específicamente para la realización de la misma y requirió de un tiempo

computacional de 88,39 segundos. Por otro lado, la programación del algoritmo genético requiere de más astucia y habilidades del programador para su elaboración y consumió mayor tiempo computacional, pero logró encontrar una mejor solución que el Grid.

Cuando se midió y comparó el impacto que tenía la selección de variables por paso a paso y la optimización de los hiperparámetros con algoritmo genético (dejando las 17 variables inicialmente propuestas) se observó que, de las dos técnicas, la que mayor influencia tiene sobre el aumento del coeficiente de determinación es la selección de las variables de entrada del modelo predictivo, ya que su incremento fue de 96,25%, 14,52% más que el algoritmo genético.

Finalmente, la implementación de ambas técnicas de optimización sobre el modelo de red neuronal propuesto inicialmente, es decir, la selección de variables por paso a paso y la búsqueda de hiperparámetros con algoritmo genético, incrementó el desempeño de la red R^2 un 185,67%. Sin embargo, un coeficiente del valor de 0,1753 sigue siendo muy bajo, debido a que se cuenta únicamente con 428 tomas para cada una de las variables predictoras.

Al comparar los resultados obtenidos en el presente proyecto con los del trabajo de grado titulado “Modelo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado” (Gamboa & Cáceres, 2019) el modelo de red neuronal optimizado tuvo un mejor desempeño que las máquinas de soporte vectorial en un 65,68% y el modelo lineal generalizado en 32,90%.

9. Recomendaciones

Se reconoce una debilidad a nivel nacional en investigación y agricultura, por lo tanto, es necesario incentivar trabajos de investigación por parte de la Escuela de Estudios Industriales y Empresariales y de igual forma de la Universidad Industrial de Santander que impulsen el desarrollo agrícola de la región desde diferentes áreas del conocimiento. De igual forma, se recomienda promover en los futuros egresados de la escuela de Ingeniería Industrial conocimientos y habilidades competitivas propias del aprendizaje automático.

Se recomienda al sector cacaotero tener en cuenta el resultado de esta investigación para analizar y controlar las características que tienen mayor impacto en la predicción del rendimiento del cultivo de estudio (Diámetro del tronco, Radiación, Temperatura y Número de días con Lluvia), y a su vez crear estrategias que aumenten la productividad y la calidad del mismo.

Se sugiere aumentar la toma de datos para el cultivo experimental de cacao (con el fin de aumentar la variabilidad en los datos) y también iniciar estudios a fines con otros cultivos característicos del departamento para fortalecer la toma de decisiones apoyada en la investigación en las diferentes federaciones de agricultores.

Se propone comparar los resultados de este proyecto con otras investigaciones que involucren otros modelos de aprendizaje automático aplicados al caso de estudio del cultivo experimental de cacao, con el fin de identificar diferencias entre los modelos y encontrar alguno que se ajuste en mayor medida a las predicciones.

Se plantea al grupo de investigación Opalo la idea de compilar los códigos de programación de este, pasados y futuros proyectos de investigación en un único Github, de modo que sirvan de

herramienta de apoyo a estudiantes interesados en realizar proyecto de grado en modalidad trabajo de investigación.

Referencias bibliográficas

- Acosta Bonilla, C. S., & Villarraga Rubio, M. V. (2006). *DIAGNOSTICO TECNOLÓGICO DEL CULTIVO DE CACAO (Theobroma cacao) EN EL MUNICIPIO DE EL DORADO, DEPARTAMENTO DEL META*. Retrieved from <http://bibliotecadigital.agronet.gov.co/bitstream/11348/6570/1/237.pdf>
- Amat, J. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. Retrieved March 27, 2019, from https://rpubs.com/Joaquin_AR/287787
- Ballesteros-ricaurte, J. A. (2015). Identificación del estado de madurez de las frutas con redes neuronales artificiales, un a revisión, *13*(1), 117–132.
- Bocca, F. F., & Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and Electronics in Agriculture*, *128*, 67–76. <https://doi.org/10.1016/j.compag.2016.08.015>
- Brownlee, J. (2014). An Introduction to Feature Selection. Retrieved February 10, 2019, from <https://machinelearningmastery.com/an-introduction-to-feature-selection/>
- Bustos M., J. R. (2005). Inteligencia Artificial en el sector agropecuario. *Seminario de Investigación I*, *2*, 1–8.
- Charoen-Ung, P., & Mittrapiyanuruk, P. (2011). Sugarcane Yield Grade Prediction using Random Forest with Forward Feature Selection and Hyper-parameter Tuning. *British Columbia Medical Journal*, *53*(6), 268–273. <https://doi.org/10.1007/978-3-319-93692-5>
- DataRobot. (2018). Model Tuning | DataRobot Artificial Intelligence Wiki. Retrieved March 7,

2019, from <https://www.datarobot.com/wiki/tuning/>

Departamento Nacional de Planeación. (2014). *Plan nacional de desarrollo 2014-2018*. Retrieved from [https://colaboracion.dnp.gov.co/CDT/PND/PND 2014-2018 Tomo 1 internet.pdf](https://colaboracion.dnp.gov.co/CDT/PND/PND%202014-2018%20Tomo%201%20internet.pdf)

Dorado, J., & Gestal, M. (2010). *Introduccion a Los Algoritmos Geneticos. Digitalia*.

Emamgholizadeh, S., Parsaeian, M., & Baradaran, M. (2015). Seed yield prediction of sesame using artificial neural network. *European Journal of Agronomy*, 68, 89–96. <https://doi.org/10.1016/j.eja.2015.04.010>

Enesco, M., & Keshav, D. (2015). Feature Engineering: Techniques, examples and case studies | CommonLounge. Retrieved April 7, 2019, from <https://www.commonlounge.com/discussion/039369851e32450db4ae2a65c4913b9a>

Fiszelew, A., & García-Martínez, R. (2012). GENERACIÓN AUTOMÁTICA DE REDES NEURONALES CON AJUSTE DE PARÁMETROS BASADO EN ALGORITMOS GENÉTICOS.

Gholipoor, M., Rohani, A., & Torani, S. (2013). Optimization of traits to increasing barley grain yield using an artificial neural network. *International Journal of Plant Production*, 7(1), 1–18. <https://doi.org/10.2135/cropsci2001.4151541x>

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, 12(2), 313. <https://doi.org/10.5424/sjar/2014122-4439>

Hilera González, J. R., & Martínez Hernando, V. J. (1995). Redes neuronales artificiales: fundamentos, modelos y aplicaciones / J.R. Hilera González, V.J. Martínez Hernando.

- Retrieved March 7, 2019, from https://www.researchgate.net/publication/31763291_Redes_neuronales_artificiales_fundamentos_modelos_y_aplicaciones_JR_Hilera_Gonzalez_VJ_Martinez_Hernando
- Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*, 31(2), 231–240. <https://doi.org/10.1016/j.eswa.2005.09.024>
- Irmak, A., Jones, J. W., Batchelor, W. D., Irmak, S., Boote, K. J., & Paz, J. O. (2006). ARTIFICIAL NEURAL NETWORK MODEL AS A DATA ANALYSIS TOOL IN PRECISION FARMING, 49(6), 2027–2038.
- Ji, B., Sun, Y., Yang, S., & Wan, J. (2007). Artificial neural networks for rice yield prediction in mountainous regions. *Journal of Agricultural Science*, 145(3), 249–261. <https://doi.org/10.1017/S0021859606006691>
- Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, 85(1), 1–18. <https://doi.org/10.1016/j.agry.2004.07.009>
- Khairunniza-Bejo, S., Mustaffa, S., & Ishak, W. I. (2014). (PDF) Application of Artificial Neural Network in Predicting Crop Yield: A Review. Retrieved March 7, 2019, from https://www.researchgate.net/publication/283570924_Application_of_Artificial_Neural_Network_in_Predicting_Crop_Yield_A_Review
- Lykhovyd, P. V. (2018). Prediction of sweet corn yield depending on cultivation technology parameters by using linear regression and artificial neural network methods. *Biosystems Diversity*, 26(1), 11–15. <https://doi.org/10.15421/011802>

- Mahanta, J. (2017). Introduction to Neural Networks, Advantages and Applications. Retrieved March 7, 2019, from <https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207>
- Mayo, M. (2018). Step Forward Feature Selection: A Practical Example in Python. Retrieved March 27, 2019, from <https://www.kdnuggets.com/2018/06/step-forward-feature-selection-python.html>
- Mejía, L. F. (2018). *Política de explotación de datos big data*. Retrieved from [https://colaboracion.dnp.gov.co/CDT/Prensa/Presentación Big Data Política explotación datos.pdf](https://colaboracion.dnp.gov.co/CDT/Prensa/Presentación%20Big%20Data%20Política%20explotación%20datos.pdf)
- Miller, G., Todd, P., & Hegde, S. (1989). Designing neural networks using genetic algorithms. *Proceedings of the 3rd International Conference on Genetic Algorithms*. <https://doi.org/10.1109/COGANN.1992.273946>
- Ministerio de Agricultura y Desarrollo Rural. (2018). El sector agropecuario mantiene un crecimiento positivo con una variación de 2% en el primer trimestre de 2018. Retrieved February 10, 2019, from <https://www.minagricultura.gov.co/noticias/Paginas/El-sector-agropecuario-mantiene-un-crecimiento-positivo-con-una-variación-de-2-en-el-primer-trimestre-de-2018.aspx>
- Montana, D. J., & Davis, L. (1989). Training feedforward neural networks using genetic algorithms. *Proceedings of the International Joint Conference on Artificial Intelligence*, 762–767. [https://doi.org/10.1016/S0749-6419\(99\)00031-5](https://doi.org/10.1016/S0749-6419(99)00031-5)
- Montoya Restrepo, I. A., Montoya-Restrepo, L. A., & Lowy Ceron, P. D. (2015). Oportunidades para la actividad cacaotera en el municipio de Tumaco, Nariño, Colombia * Petter David

- Lowy-Ceron. *Unilibre Cali*, 11(1). <https://doi.org/10.18041/entramado.2015v11n1.21107>
- Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, 05(03), 946–952. <https://doi.org/10.21917/ijsc.2015.0133>
- Naciones Unidas de Colombia. (2015). Objetivo 8: Promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos | Objetivos de Desarrollo sostenible. Retrieved April 5, 2019, from <https://nacionesunidas.org.co/ods/objetivo-8-promover-el-crecimiento-economico-sostenido-inclusivo-y-sostenible-el-empleo-pleno-y-productivo-y-el-trabajo-decente-para-todos/>
- Organización de las Naciones Unidas para la Alimentación y la Agricultura. (2002). *Agricultura mundial: hacia los años 2015/2030*. Retrieved from <https://agua.org.mx/wp-content/uploads/2017/11/Agricultura-mundial-hacia-los-años-20152030-Informe-resumido.pdf>
- Osisanwo, F., Akimsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, 48(3), 128–138. <https://doi.org/10.14445/22312803/IJCTT-V48P126>
- Palma, J., & Marín, R. (2008). *Inteligencia artificial: técnicas, métodos y aplicaciones*.
- Ramón, G. (2015). *Correlación entre variables*. Retrieved from http://viref.udea.edu.co/contenido/menu_alterno/apuntes/ac36-correlacion-variables.pdf

- Safa, M., Samarasinghe, S., & Nejat, M. (2015). Prediction of wheat production using artificial neural networks and investigating indirect factors affecting it: Case study in canterbury province, New Zealand. *Journal of Agricultural Science and Technology*, 17(4), 791–803.
- Shaikh, R. (2018). Feature Selection Techniques in Machine Learning with Python. Retrieved February 10, 2019, from <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
- Shekhar, A. (2018). What Is Feature Engineering for Machine Learning? – MindOrks – Medium. Retrieved February 10, 2019, from <https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a>
- Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
- Su, Y. (2011). Neural Network Hyperparameter Optimization with Genetic Algorithm on Handwritten Digit Problem , and the Observation on Neuron network by Distinctiveness.
- Whitley, D., Starkweather, T., & Bogart, C. (1990). Genetic algorithms and neural networks: optimizing connections and connectivity. *Parallel Computing*, 14(3), 347–361. [https://doi.org/10.1016/0167-8191\(90\)90086-O](https://doi.org/10.1016/0167-8191(90)90086-O)
- Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. *Australian Institute of Health and Welfare*. <https://doi.org/10.1109/5254.671091>

