v. 2, n. 17, 2022

Journal of
Engineering
Research

ALGORITMO DE CLASIFICACIÓN MEDIANTE UN ENFOQUE DE MACHINE LEARNING Y SU APLICACIÓN AL ESTUDIO METEOROLÓGICO

Pedro Elizardo Donis del Cid
https://orcid.org/0000-0003-0844-9796

Resumen: La clasificación de datos propone una polaridad tanto en el algoritmo de árbol como en el teorema de las causas. Se necesaria hacer uso de conjuntos de datos, en este estudio son variables meteorológicas, para construir un modelo predictivo basado en entrenamiento. Los datos son almacenados y procesados haciendo uso de almacenes en grandes volúmenes de datos, estos proponen un sistema basado en entidades con independencia sustancial, de objetos de datos, a partir de una técnica llamada ciencia de datos o Data Science. Técnica que permite adquirir información de valor de los datos, en este caso equipos y sensores de IoT. Inicialmente se capturan en formato no procesado o Raw Data. El objetivo del estudio es hacer uso de las plataformas tecnológicas disponibles para trabajar el pronóstico de datos de meteorología basados en algoritmos de ML (Machine Learning) y datos históricos para al altiplano central, costa del pacífico y valles de oriente de Guatemala. Los métodos utilizados son estadísticos embebidos en algoritmos de tipo predictivos, además de una matriz de confusión para evaluar los resultados obtenidos en el estudio de estas herramientas y recursos, con un enfoque cuantitativo e instrumentos de IoT, ML y Big Data. El tipo de estudio general es correlacional-predictivo-histórico del lado de la variable y comparativo-explicativo para el algoritmo. Las variables tienen una dimensión anual, mensual y diaria ente 2000 y 2018. En conclusión, se logra el pronosticó del nivel de humedad para las muestras seleccionadas en base a las variables meteorológicas tomadas con dispositivos electrónicos y procesadas por medios informáticos entre dos algoritmos y un dieciocho por ciento de diferencia al comparar la eficiencia.

Palabras clave: Clasificación de datos, Internet de las Cosas, Industria 4.0, Modelo Supervisado.

INTRODUCCIÓN

La automatización con los datos tiene su aplicación en diferentes industrias y campos de estudio. Estas herramientas permiten no solo la adquisición de datos sino también, un sistema de procesamiento estadístico, lingüística computacional y en un nivel más profundo; aprendizaje de máquina.

La industria 4.0, como se conoce a esta nueva revolución, posibilita con IoT y sistemas ciberfísicos (CPS) conseguir que máquinas y otros sistemas electrónicos se comuniquen, como en una red social. Sin embargo, lo que habilita esta nueva forma de industria, es la posibilidad de comunicación en tiempo real.

En la Industria 4.0, la clasificación de datos y el aprendizaje de la máquina por medio de datos, es decir, la generación código que puede convertirse en acciones de un agente de IA sin necesidad de ser previamente programado (Ain, et al., 2015; Nicholson, et al., 2019; Réda, et al.,2020; Turesson et al.,2016; Vu et al.,2018; Zhang, et al., 2017). Es la base de esta revolución industrial (Galvão, et al., 2022; Osmana, Ghirana, 2019). Además, del aumento de las capacidades en las centrales de datos por medio de la tecnología Big Data y la posibilidad de conectividad y recolección de datos automatizado por medio de IoT (Qaffas et al., 2021). Incluye sensores y otras tecnologías especializadas.

Big Data es una tecnología que se ha desarrollado debido exponenciales crecimientos en volumen de datos a nivel mundial y las limitaciones de los microprocesadores actuales. Están técnicas habilitan el cómputo para trabajar una tarea compleja en múltiples nodos, con varios núcleos y memoria agrupada (Franke et al., 2016; Qaffas et al.,2021; Tang et al.,2019; van Evert et al.,2017). Por su parte, IoT recolecta la información, es un conjunto de componentes electrónicos que tienen la

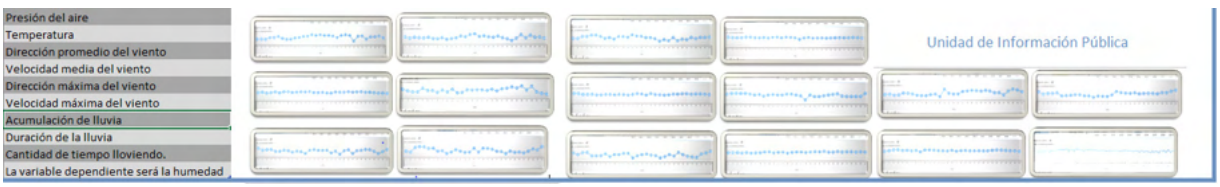
capacidad de conectarse y comunicar datos (Qaffas, et al., 2021).

Para este estudio se utilizaron sensores que son implementados y administrados por el Instituto Nacional de Sismología, Vulcanología, Meteorología e Hidrología de la República de Guatemala y es dependencia del Ministerio de Comunicaciones y Obras Públicas. Según acuerdo gubernativo de 26 de marzo de 1976. Estos sensores tienen la capacidad de comunicación con la base de almacenamiento de datos, es decir, en la parte operativa de recolección de los datos de la investigación no es necesario la presencia de entrevistas o encuestas para capturar datos. Pero si es necesario configurar los equipos y parametrizar los sistemas.

En el método utilizado se hizo un resumen de las variables agrupándolas por día, mes y año. Y las variables operadas de manera continua de temperatura, lluvia, velocidad del viento, radiación, nubosidad.

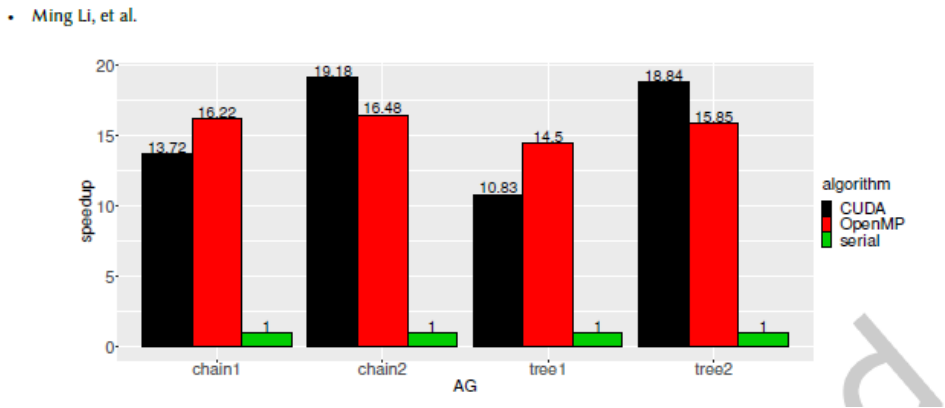
Por su parte, Galvão et al (2022) utilizando la plataforma Cloudera acompañado de Spark con Python logró presentar un modelo con datos de un proceso industrial. Se usaron funciones de agregación, reportes por medio de tableros. Donde, se determinó un déficit de tornillos apretados del 14.3% para ello se utilizaron DataFrame a partir de objetos JVM y utilizando métodos de mapas y reductores por medio de API's muy semejante a la utilización del lenguaje R. Después de realizar la adquisición de datos, procesarlos y almacenarlos, se generan todos los sistemas de visualización de datos para los temas administrativos correspondientes donde la comunicación entre máquinas es muy importante, porque los datos no son digitalizados por personas.

Li et al. (2022) con el experimento se realiza comparaciones entre los algoritmos



Variables de entrenamiento.

Elaboración propia. Origen de datos: Unidad de Información Pública, INSIVUMEH.



Eficiencia de otros algoritmos.

Fuente: Li et al (2022).

de grafo de ataque utilizando computación de alto rendimiento para algoritmos OpenMP y CUDA AG en un sistema de procesamiento distribuido. El experimento ha demostrado la eficiencia de la estrategia del procesamiento en paralelo. Los resultados fueron utilizados para hacer parche de las vulnerabilidades.

REPOSITORIOS DE DATOS

Los almacenes de datos han estado presentes en la industria desde hace algunos años, sin embargo, gracias a esos datos históricos que se guardan en estos almacenes es posible predecir. Esto combinado con la variedad de datos que se tienen con las nuevas tecnologías de Big Data y IoT trasciende en tecnologías digitales que permiten mejorar procesos industriales con impactos económicos importantes. La investigación de Galvão el al (2022) hace referencia al uso de estas tecnologías emergentes que se aplican para el tratamiento de la información, tanto la extracción como, el tratamiento y almacenamiento**.**

En los resultados se puede apreciar una medición a partir de un proceso industrial el cual denota información que se guarda en Big Data, se procesa y se envía para el monitoreo por medio de tableros. Las herramientas estadísticas e informáticas toman un papel relevante. Gracias a la estadística es posible predecir o inferir los sucesos próximos para anticiparse a los escenarios que se presentan. La informática hace posible recolectar, almacenar y procesar la información que llega desde máquinas como es el caso de IoT. Tecnología que permite, por medio de internet, transferir datos de sensores y otros elementos de medición.

Muchos de los gráficos que muestra este caso de estudio está orientado al uso de software especializado para este diseño de información. Donde, se hace presente nuevos métodos para visualizar los datos. Datos que son importantes para la toma de decisiones.

Mohamed et al. (2019) realizó un estudio del COVID-19 empleando Big Data con retrospectiva y descriptiva avanzada se generan indicadores y tendencias por medio de bases de datos estructuradas y no estructuradas se generan correlaciones de datos para poder inferir con mayor objetividad.

De Lecuona (2018) hace énfasis en que se deben de incorporar expertos en ciencias de datos para poder determinar problemas técnicos o problemas éticos aplicados al big data. Esto implica un esfuerzo conjunto con las direcciones de sistemas, servicios jurídicos y áreas de gestión e innovación para proyectos de investigación con datos masivos o Big Data. En el caso de estudio de la información clínica es necesario crear los mecanismos legales, técnicos y tecnológicos para que no se vulnere ningún derecho. Las tecnologías 4.0 presentan grandes oportunidades tanto de investigación como para la ocupación y reactivación económica. La salud es una industria que necesita de implementación de sistemas basados en grandes volúmenes de datos (big data).

INTERNET DE LAS COSAS, HERRAMIENTAS PARA LOS DATOS

Qaffas et al. (2019) muestra un caso de estudio donde existen variables de entrada las cuales son definidas como conjunto de datos de enfermedad de hipertensión y variables de salida; específicamente, comprende una clasificación SVM, máquina de vectores de soporte, este método; carga un conjunto de datos, agrega de una manera aleatoria un conjunto de entrenamiento del 20% y un conjunto de datos de prueba del 80%, en el caso de estudio, se incluye una generación de clasificadores en base al conjunto de datos de entrenamiento, posteriormente, se entrena el clasificador y se construye una predicción aplicada a la muestra de datos por medio

de este algoritmo clasificador, entrenado previamente.

El caso de estudio incluye una evaluación de los clasificadores basados en los parámetros, así como una selección de sus características según sus pesos; para poder aplicar este tipo de biotecnología en la resolución de problemas asistidos por ordenadores en la clasificación de los parámetros o factores que inciden en la hipertensión, es necesario hacer uso de IoT. Al final de la investigación se determinó que los datos basados en IoT son muy efectivos y prometedores, sin embargo, es necesario realizar alguna optimización; dentro de los resultados relevantes del estudio se tiene que las personas mayores con diabetes tienen más probabilidad de desarrollar hipertensión. El tabaquismo juega un papel menos importante que la diabetes y las personas mayores deben consumir menos sal. SVM produce mejores resultados que el algoritmo C4.5.

Herrera et al. (2018) demostró que en el Big Data existen muchos desafíos retos y oportunidades dentro de las ventajas tenemos que permite describir las necesidades de la compañía, mejora en la toma de decisiones, extracción de información valiosa, evaluación de productos, segmentación de clientes y fluidez de la información. Sin embargo, entre las principales desventajas se tiene que la privacidad y el control de la información, los grandes costos de hardware y software, la capacitación del personal.

Kalbandi y Anuradha (2015) muestran algunas herramientas utilizadas para el Big Data, entre las cuales destacan Spark, Flume, Tez, Cassandra, Avro, ZooKeeper, Hive, Pig, Hbase, Hadoop Yarn/Map Reduce, Hadoop Distributed File System (HDFS); y entre los principales resultados obtenidos se halló que Hadoop es una es una herramienta libre que sirve para la integración de datos, el procesamiento de datos, monitoreo y la programación del flujo de trabajo.

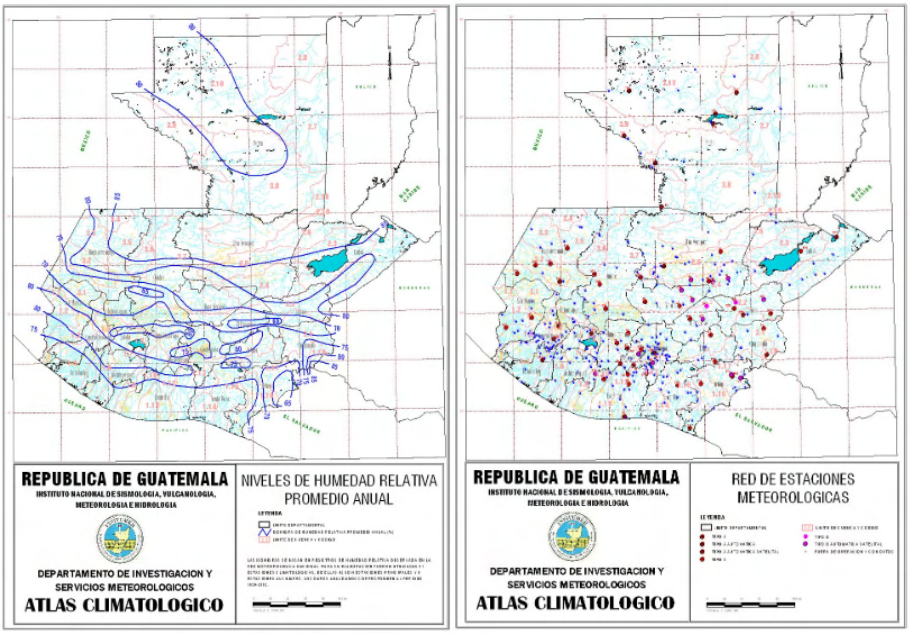
HUMEDAD RELATIVA Y CARACTERÍSTICAS DEL SUELO

Van Evert et al (2017) realizó un estudio de control de malezas haciendo uso de sistemas de Big Data y dentro de los métodos utilizados se hizo uso de ambientes NoSQL (no solo Structured Query Language) porque los sistemas tradicionales resultan difíciles de manejar cuando se distribuye el sistema en varias máquinas. Se hizo uso de técnicas SVMs y el conjunto de datos respaldados por Big Data para el control de amenazas incluye información espacial, la posición del paisaje, las características del suelo y del clima.

El resultado se determinó que se puede controlar mediante los factores de tiempo, severidad y ubicación, además, se determinó que en las temporadas de crecimiento de la maleza es cuando se deben de capturar estos datos en tiempo real para obtener avances en la ciencia agrícola en TICs para la captura, almacenamiento de datos, análisis y los aportes en la misma.

Dentro de la metrología, la humedad relativa viene dada por una relación entre el vapor de agua que tiene el aire y la cantidad que debería tener para saturarse. Este valor viene dado en porcentaje. La fórmula quima del agua es H2O dos moléculas de hidrógeno y uno de oxígeno, este es entonces un compuesto químico. La atmósfera terrestre son mezclas de compuestos en estado gaseoso sobre la superficie de la tierra. Para este estudio se tuvo a bien estudiar variables meteorológicas cerca de la tierra, siendo estas las que se encuentra en interacción directa con las personas, existen otras partes de la atmosfera terrestre que interactúan con la troposfera y que alteran la misma en primer lugar la estratosfera y otras capas que interactúan con esta última. A continuación, se presenta un mapa general.

La metrología y la meteorología, ambas interactúan con esta variable climática y tiene que ver en gran medida con el sistema



Mapa de Guatemala

Fuente: https://insivumeh.gob.gt/hidrologia/ATLAS_HIDROMETEOROLOGICO/Atlas_Climatologico/esmeteo.jpg

socioeconómico en la naturaleza. Dada la naturaleza de ciencias exactas, ocupan diferentes instrumentos para la experiencia y comprensión de los sucesos. Haciendo uso de mapas de información geológica.

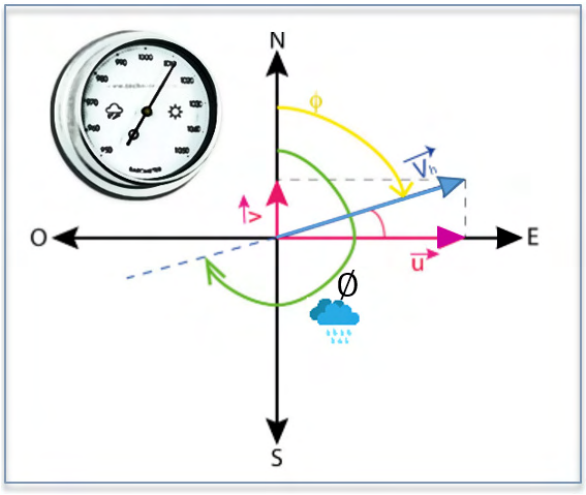


Imagen ilustrativa

Para este estudio se ocupó el concepto de Teorema de Bayes (teorema de las causas) para el computo de resultados (Mesa et al.,2021; Tang et al., 2019) enunciado por Thomas Bayes, donde se denota una probabilidad condicional en términos de distribución de probabilidad: P (A|B) = P (B∩A) / P(A).

$$P\ (A_{i,}\ /\ B)\ =\ \frac{P\ (A_i)\ P\ (B\ /\ A_i)}{\sum_{i=1}^{n} P\ (A_i)\ P\ (B\ /\ A_i)}$$

Este teorema se caracteriza por representar los resultados en términos de la probabilidad de que un evento ocurra dado que ha sucedido otro. Esta reacción se puede hacer en cadena para poder unir variables. Lo que se hizo en este estudio, pero con la diferencia que se utilizaron herramientas avanzadas de informática. Esta última, se ocupa del tratamiento automático de la información. En el teorema existe un acercamiento a la probabilidad, a 100% o a 0%.

Hussein et al (2021) realizó un estudio utilizando un enfoque escalable de detección de intrusos haciendo uso de un Framework de

Big Data utilizando métodos de clasificación como K-Means, RUSBoost y DT. Esto demuestra la factibilidad de transferencia de datos por un medio físico, almacenamiento y procesamiento del mismo en el contexto tecnológico actual.

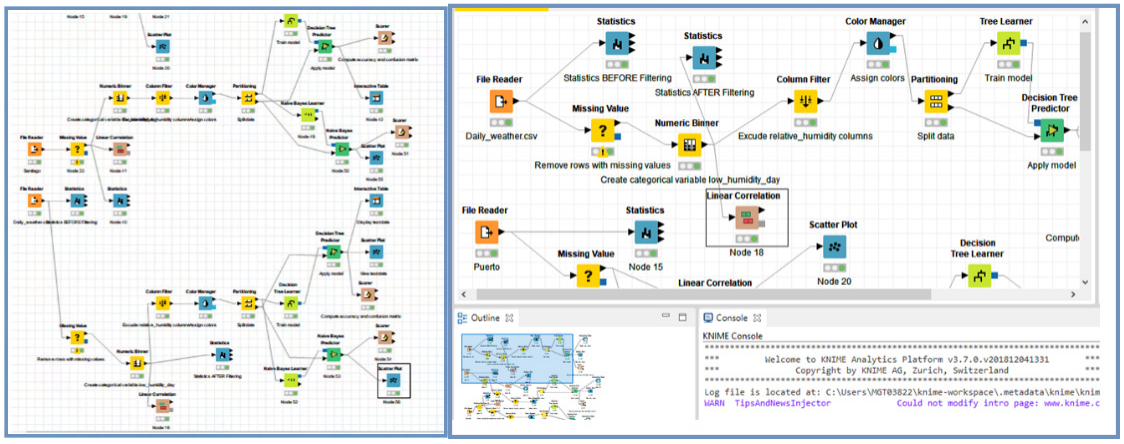
METODOLOGÍA

El estudio hace uso de las plataformas tecnológicas disponibles para trabajar el pronóstico de datos de meteorología basados en algoritmos de ML y datos históricos Big Data Warehouse para al altiplano central, costa del pacífico y valles de oriente para Guatemala.

El método utilizado fue la integración de datos por medio de sensores capaces de tomar las mediciones y trasmitir los datos. Se utilizó un modelo supervisado de aprendizaje de máquina con datos cuantitativos de las variables de metrología. Consistente en determinar una variable categórica de baja humedad o alta humedad en base al conjunto de variables independientes del modelo. Además, se utilizó un algoritmo de árbol de decisión y otro método de inferencia bayesiana, muy parecida a la regresión logística que ocupa una variable categórica dicotómica con resultado final. Esto se da por aproximación a un valor ya sea 0 o 1.

El enfoque de la investigación tiene como propósito la indagación sobre estos algoritmos especializados y orientados a datos para que pueda ser aplicado en otras industrias por los profesionales de las ciencias de la computación. Dentro de las herramientas y tareas se llevó acabo la configuración de los equipos para la transferencia de comunicación o transferencia de datos y, una herramienta para la configuración del flujo para el job del procesamiento en ML.

Las variables fueron agrupadas por la dimensión de día, mes y año a partir del año 2000 con 3,690 observaciones completas sin valores atípicos para el altiplano central, muestra de: Santiago Atitlán, Cubulco; 43,464 observación para el pacífico, muestra para Puerto de San Jose; 6,478 observaciones después de la limpieza para valles de oriente, muestra Esquipulas y La Unión. Las variables utilizadas fueron de razón, de intervalo y una variable predictiva con valor dual (dicotómica). Se utilizaron dos algoritmos como métodos predictivos, además de un matriz de confusión para evaluar los resultados obtenidos en el estudio de las herramientas y recursos, con corte longitudinal, enfoque cuantitativo con instrumentos de IoT, ML y Big Data, y un tipo de estudio predictivo desde el punto de



Diseño Conceptual del Proceso.

Elaboración propia.

vista de la variable analizada y descriptivo para los algoritmos.

Las mediciones fueron continuas con variables de meteorología: Presión del aire, temperatura dirección promedio del viento, velocidad media del viento, dirección máxima del viento, velocidad máxima del viento, acumulación de lluvia, duración de la lluvia y cantidad de tiempo lloviendo, y procesadas ser registrarse en almacenes de datos de forma periódica en ciclos: diarios, mensuales y anuales para el rango de fechas del año 2000 al 2018. El almacenamiento de estos datos y la calidad juega un papel muy importante para poder generar conjuntos de datos resumidos de las variables.

RESULTADOS

Guatemala es un país agrícola que depende mucho de las situaciones climatológicas para los medios de producción, a continuación, se presenta una imagen de una ciudad (casco urbano de municipio o pueblo) promedio y su situación socioeconómica:

Fuente: Google.

Se caracteriza por dos épocas clásicas, una lluviosa y otra seca, la primera va de mayo a octubre y la última de noviembre a abril. La época seca se da por un incremento en la presión atmosférica y movimientos de aire frío. La época lluviosa por lo general da inicios en el mes de mayo. Las temporadas de siembra y cosecha, dependen muchas veces de estas épocas. Las cuales pueden variar en condiciones, lo cual puede afectar las cosechas o mejorar el rendimiento de la producción agrícola.

Se le aplicó un test por medio de la matriz de confusión. En la primera celda se presenta, 76 observaciones efectivas pronosticando la humedad baja. Y en la última casilla, 95 observaciones fueron efectivas al pronosticar la humedad alta (no baja). El resto de valores presentados corresponden a la tasa de error del modelo cada uno de ellos tiene un tipo: Error Tipo I y Error Tipo II. Afirmar que es baja cuando no lo es, y, por otra parte, que es no baja cuando si lo es. El algoritmo de árbol de decisión ha mostrado una exactitud del 80% frente a un 62% de la regresión bayesiana. Así mismo al comparar el % de error se aprecia una diferencia del 18%.

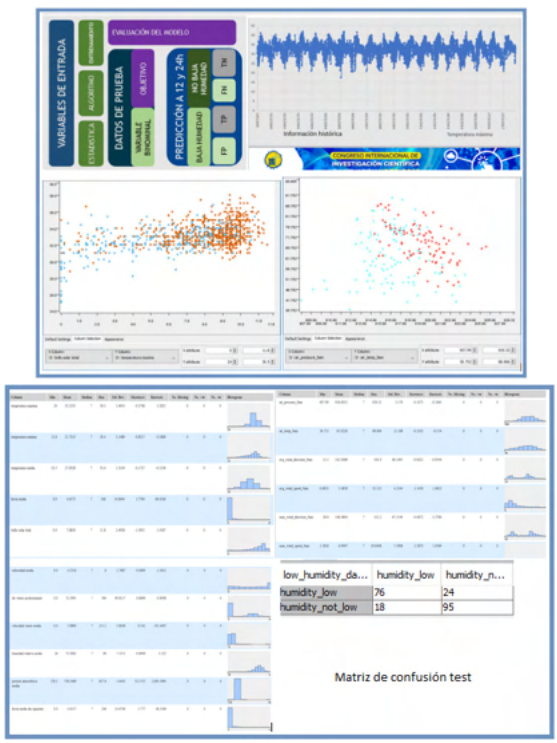


Gráfico de Variables.

En la parte superior izquierda podemos apreciar el esquema del modelo, el cual comprende un conjunto de datos de entrada: Presión del aire, temperatura dirección promedio del viento, velocidad media del viento, dirección máxima del viento, velocidad máxima del viento, acumulación de lluvia, duración de la lluvia y cantidad de tiempo lloviendo. La parte estadística que consiste en las operaciones con datos, el algoritmo, son instrucciones de máquina, para procesar las operaciones. Estas se llevan a cabo de forma ordenada y el entrenamiento.

Posteriormente sigue el modelo, este consiste en tomar la información del entrenamiento y proyectar con datos de prueba, es acá donde se tiene un objetivo que se traduce en una variable. Dentro del modelo también existe una parte de predicción y su parte medular es la matriz de confusión. La cual indica los FP Falsos Positivos TP Verdaderos Positivos FN Falsos Negativos TN Verdaderos Positivos. Cada uno de estos tiene una interpretación en la matriz. Y mide la efectividad del modelo. Para este estudio la variable dependiente es la humedad. La cual tiene un valor dicotómico, alta o baja.

En la parte inferior de la gráfica se puede ver la dispersión entre variables en este caso tenemos dos, el brío solar total y la temperatura máxima, esto puede cambiar dependiendo de cuales variables se comparen. Esto es importante para poder comparar entre sí las variables independientes. A continuación, se presenta una tabla con una muestra de datos, tanto las variables de entrada, la humead real y la predictiva.

Entonces, al comparar los dos algoritmos la diferencia fundamental radica en la interpretación del modelo y de los resultados, donde el árbol de decisión es más fácil de ejemplificar y reconstruir con valores concretos. Por ejemplo, veamos la siguiente imagen del árbol de decisión.

Podemos apreciar en las hojas del árbol; si la dirección del viento es mayor al parámetro indicado en el diagrama se clasifican las observaciones: Caso primero 15 y 14 respectivamente. Sin embargo, esto puede cambiar según el valor o parámetro del nodo padre anterior en el árbol, que es la variable de temperatura del aire: Si la temperatura del aire es menos o igual que 72.3578 se dirigen las muestras respectivamente, en el

Row ID	D temper...	D temper...	D temper...	D lluvia m...	D evapor...	I nubosid...	I dir vien...	D velocid...	S humedad di...	S Prediction (...
38111	24.8	15.5	21.8	0	4.6	6	0	15	no_baja_humedad	no_baja_humedad
38113	25.5	14.5	19.8	0	3.1	5	0	17	baja_humedad	baja_humedad
38117	25.5	13.5	19.6	0	4.4	6	0	10	no_baja_humedad	no_baja_humedad
38122	25.3	12.8	18.3	0	1.9	6	0	15	no_baja_humedad	no_baja_humedad
38128	24.3	15.5	20.4	0	3.4	6	0	10	no_baja_humedad	no_baja_humedad
38135	23.8	14.3	19.1	1.4	8.3	5	0	9	no_baja_humedad	no_baja_humedad
38177	22	16	19.4	8.1	1.9	6	0	17	no_baja_humedad	no_baja_humedad
38182	24.5	15.3	19.7	0	3.7	6	0	17	no_baja_humedad	baja_humedad
38191	24.8	14.3	19.4	0	3.7	6	0	8	no_baja_humedad	no_baja_humedad
38193	26	15.5	19.8	84.2	3.3	6	0	214	no_baja_humedad	no_baja_humedad
38194	26.8	14	20.5	0	22.9	4	0	10	no_baja_humedad	baja_humedad
38204	24.5	15	20.3	0	1.3	3	0	15	no_baja_humedad	baja_humedad
38206	25	13	19.6	0	4.9	5	0	15	no_baja_humedad	no_baja_humedad
38210	25.5	15.3	20	0	4.9	6	0	20	no_baja_humedad	no_baja_humedad
38211	25.3	13.3	19	0	3.3	4	0	28	no_baja_humedad	no_baja_humedad
38219	27.5	17	20.7	0	5.2	3	0	28	no_baja_humedad	no_baja_humedad
38232	26	15	19.8	0	5.6	6	0	17	no_baja_humedad	baja_humedad
38238	23.5	13.5	17.6	6.3	4.1	6	0	13	no_baja_humedad	no_baja_humedad
38241	24.5	14.5	18.5	0	3	6	0	10	no_baja_humedad	no_baja_humedad
38242	23.8	13.8	17.8	7.2	4.2	5	0	10	no_baja_humedad	no_baja_humedad
38256	24	15.5	18.5	0.6	3.4	6	0	7	no_baja_humedad	no_baja_humedad
38261	25.5	15.5	19.8	0	2.8	4	0	16	no_baja_humedad	no_baja_humedad
38262	26.3	14.5	20.6	0	5.1	6	0	10	no_baja_humedad	no_baja_humedad
38263	24.5	17	19.4	0.1	1.1	5	0	7	no_baja_humedad	no_baja_humedad
38270	24	14.3	17.5	32.2	3.3	6	0	10	no_baja_humedad	no_baja_humedad

Tabla de Muestra.

Elaboración propia.

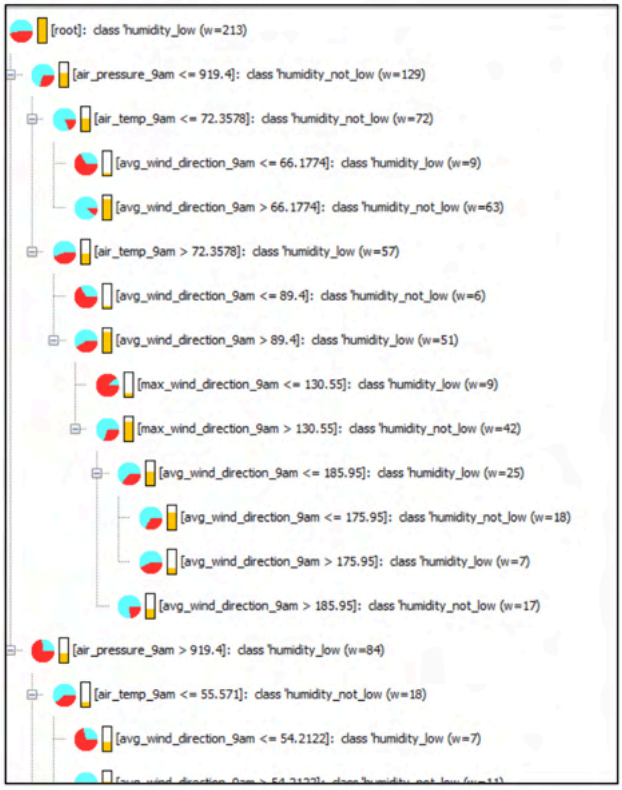
Árbol de decisión.

Elaboración propia.

primer caso 39 con humedad baja y 240 con humedad alta. Y así el otro nodo hermano en el árbol, si la temperatura del aire es mayor que 72.3578; las clasificaciones son diferentes 86 y 86 respectivamente para la baja y la no baja (alta). Todos estos valores que indican hacia donde se dirige la observación o dato, depende del proceso que genera el algoritmo clasificador o modelo.

Cada mes o segmento de datos, tiene diferentes características climatológicas, meses fríos, secos y lluviosos. Así, se observa que la temperatura para Santiago Atitlán es uniforme en contraste con los valles de oriente.

El resultado muestra una eficiencia en el algoritmo del árbol de decisión, el cual se resume a continuación:



Árbol de decisión para la clasificación de datos.

Elaboración propia.

Cada parte del árbol clasificador tiene un gráfico el cual indica la proporción del valor de la variable de estudio que es clasificada en cada nodo. Estos se van agrupando desde las hojas y los nodos padres hasta llegar a la raíz. Es más fácil visualizar las decisiones que tomará el modelo y llevarlo a la programación, pero esto no es el fin del modelo. Ya que se cuenta con las cajas en las herramientas, estas ya disponen de componentes para nuevos conjuntos de datos a pronosticar y se puede colocar de manera automatizada, sin necesidad de desarrollar código en algún lenguaje en particular. En fin, el modelo y las condiciones del árbol pueden cambiar con otro entrenamiento del modelo y esto es uno de los principales conceptos computacionales aplicados en este artículo. Lo cual nos lleva a concluir que él árbol de decisión es muy efectivo para el estudio de datos meteorológicos y las mediciones realizadas por medio de equipos ha mostrado ser capaz de recolectar, almacenar y procesar efectivamente estos datos y a pesar de que existen valores faltantes, el modelo es capaz de predecir un nivel de aceptación alto.

En la raíz del árbol inicia los clasificadores como baja humedad con 213 casos. Los siguientes valores paramétricos en orden: 919.4, 72.3578, 66.1774, 89.4, 130.55, 185.95, 175.95; correspondientes a presión del aire, temperatura, dirección del viento promedio, dirección del viento media, respectivamente. Estos valores dependen directamente del modelo de entrenamiento de datos, el cual se puede ir afinando con más datos, es decir más variables o una granularidad menor en la medición cambiando de diario a horaria. Esto para poder pronosticar la humedad correcta.

Con cada configuración de valores de variables se toma una decisión si clasificar la muestra obtenida como baja o alta humedad. Esta configuración es de manera recursiva, es decir existe un orden en las variables y cada nodo tiene dos nodos hijos a los cuales debe de recurrir para determinar la clasificación final.

Solamente en las hojas del árbol se puede concluir el resultado gracias a la ruta que se recorrió en el árbol.

CONCLUSIONES

El algoritmo de árbol de decisión ha mostrado una exactitud del 80% frete a un 62% de la regresión bayesiana. Así mismo, se pronosticó el nivel de humedad para las muestras seleccionadas en base a las variables meteorológicas tomadas con dispositivos electrónicos y procesadas por medios informáticos (Filiberto, et al., 2011; Mohamed, et al., 2017). El almacenamiento de estos datos y la calidad juega un papel muy importante, principalmente, en este tipo de investigación (Roriz, et al., 2019; Spark.apache.org, 2019); ya que se tiene información de varios años, la cual debe de organizarse primero para su procesamiento. La grafica siguiente muestra una variable, una muestra (Santiago Atitlán) y la siguiente configuración de dimensión temporal: año 2006, meses de enero a diciembre, días, todos los días a excepción de los valores ausentes.

Cuando se compara porcentaje de error se aprecia una diferencia del 18%. Para finalizar, es importante hacer notar que ML (aprendizaje automático o aprendizaje de máquina) es una sub disciplina de la IA, parte fundamental de la informática que se ocupa del tratamiento automático de la información y se apoya de la estadística (Amaya, et al., 2017; Contreras, et al., 2017; Ji, et al., 2018; Tabales, et al., 2017; Wang, et al., 2018). Este estudio se apoyó además de la metrología y meteorología, las cuales fueron descritas en este estudio, en consecuencia, se comprobó la efectividad de un algoritmo basado en árbol de decisión (Contreras, et al., 2017) respecto a la regresión bayesiana (Mesa et al.,2021;

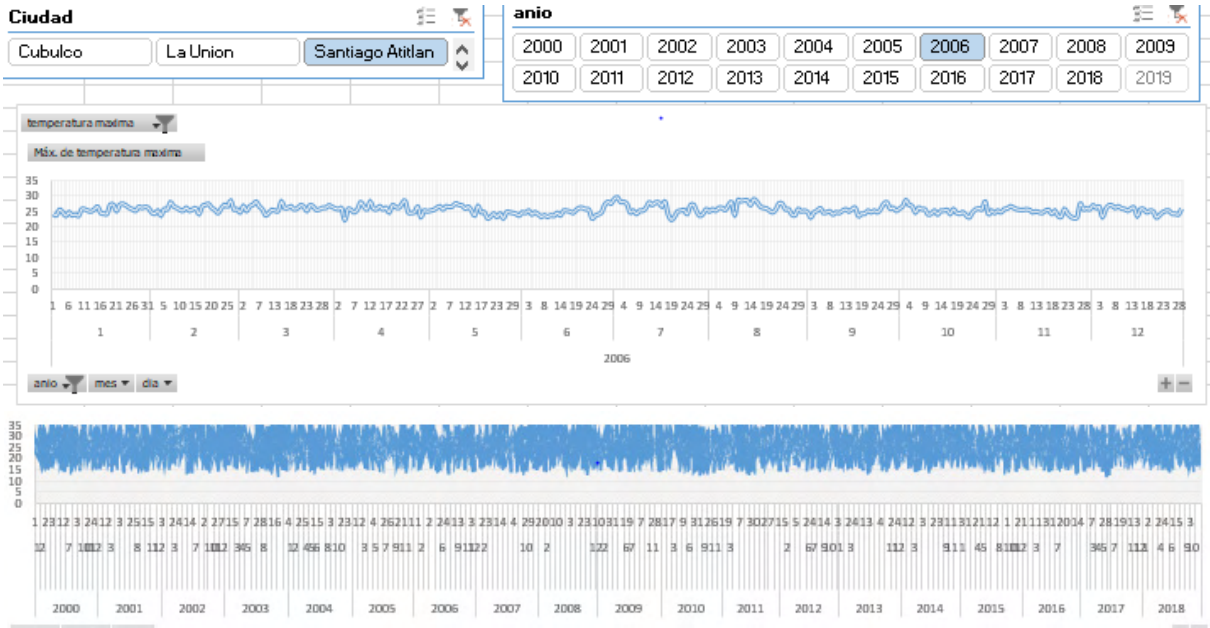


Grafico histórico.

Tang et al., 2019) y aplicado a un análisis de correlación causal de variables del clima, de tres regiones en Guatemala.

Para el modelo, cada mes o segmento de datos, tiene diferentes características climatológicas, por ejemplo, existen meses más fríos que otros, dependiendo de la región, todo esto aplica para el resto de variables como dirección del viento, radiación solar, lluvias y temporadas secas. Además, en el caso de la temperatura para Santiago Atitlán no cambia demasiado en comparación con La Unión o Esquipulas.

RECOMENDACIONES

Es importante dar mantenimiento a los equipos de medición para evitar valores atípicos, muchos de estos *missing values* o valores atípicos se deben a ausencia de información, por un incidente en el equipo, en la transmisión de datos o en el almacenamiento.

Utilizar más variables para poder mejorar el modelo que se ha presentado en esta investigación.

REFERENCIAS

Abd A.; Tawalbeh L.; Maleh Y.; Saldamli G. (2021). Big Scientific Data and Machine Learning in Science and Engineering

Ain, Q.; Aleksandrova, A.; Roessler, F.; Ballester, P. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdisciplinary Reviews: Computational Molecular Science, 5(6), 405–424. https://doi.org/10.1002/wcms.1225

Amaya, C.; Magaña, P.; Ochoa, I. (2017). Evaluación de destinos turísticos mediante la tecnología de la ciencia de datos. Universidad de Colima – México.

Contreras, E.; Ferreira, F.; Valle, M. (2017). Diseño de un Modelo Predictivo de Fuga de Clientes Utilizando Árboles de Decisión. Revista Ingenieria Industrtial, 16(1), 07–23. https://doi.org/10.22320/s07179103/2017.01

Filiberto, Y.; Caballero, Y.; Bello, R.; Frías, M. (2011). Algorithm to learn clasification rules based on the extended rough set theory. DYNA (Colombia), 78(169).

Franke, B.; Plante, J.; Roscher, R.; Lee, E.; Smyth, C.; Hatefi, A.; Chen, F.; Gil, E.; Schwing, A.; Selvitella, A.; Hoffman, M.; Grosse, R.; Hendricks, D.; Reid, N. (2016). Statistical Inference, Learning and Models in Big Data. International Statistical Review, 84(3), 371–389. https://doi.org/10.1111/insr.12176

Galvão J.; Ribeiro, D.; Machado, I.; Ferreira, F.; Gonçalves, J.; Faria, R.; Moreira, G.; Costa, C.; Cortez, P.; Santos, M. (2022) Bosch's Industry 4.0 Advanced Data Analytics: Historical and Predictive Data Integration for Decision Support.

Herrera, F. (2014). Big Data with Cloud Computing: An insight on the computing environment, MapReduce, and programming frameworks.

Hussein, A.; Arun Kumar, N.; Burbano-Fernandez, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; de Albuquerque, V. H. C. (2018). An automated remote cloud-based heart rate variability monitoring system.

Jara, M.; Cibertec, I. (2017). Introducción Machine Learning. 1–10.

Ji, Y.; Kim, S.; Kim, Y.; Lee, K. (2018). Human-like sign-language learning method using deep learning.

Kalbandi, I.; Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology.

Mesa L.; Rivera M.; Romero J. (2021). Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión.

Mohamed, A.; Berg, W.; Peng, H.; Luo, Y.; Jankowitz, R.; Wu, Sh. (2017). A deep learning method for classifying mammographic breast density categories.

Mohamed, A.; Nahafabadi, M.; Wah, Y.; Zaman, E; Maskat, R. (2019). The state of the art and taxonomy of big data analytics: View from the new big data framework.

Nicholson, K.; Richardson, R.; van Roden, E.; Quinton, R.; Anzilotti, K.; Richards, J. (2019). Machine learning algorithms for predicting scapular kinematics. Medical Engineering and Physics, 65, 39–45. https://doi.org/10.1016/j.medengphy.2019.01.005

Li, M.; Hawrylak, P.; Hale, J. (2022). Strategies for Practical Hybrid Atack Graph Generation and Analysis.

Osmana, C.; Ghirana, A. (2019). When Industry 4.0 meets Process Mining.

Qaffas A.; Hoque R.; Almazmomi N. (2021). The Internet of Things and Big Data Analytics for Chronic Disease Monitoring in Saudi Arabia

Réda, C.; Kaufmann, E.; Delahaye, D. (2020). Machine Learning applications in drug development. Computational and Structural Biotechnology Journal,18. 241-252. https://doi.org/10.1016/j.csbj.2019.12.006.

Roriz, M.; Magalhães, F; Guedes, Á.; Colcher, S.; Endler, M. (2019). An introduction to data stream processing: a complex event processing approach.

Spark.apache.org: Spark SQL and DataFrames - Spark 1.5.2 Documentation, https://spark.apache.org/docs/latest/sql-programming-guide.html, last accessed 2021/07/19.

Tabales, J M Núñez, Carmona, F J Rey, Caridad, J M (2017). Redes neuronales (RN) aplicadas a la valoración de locales comerciales.

Tang, Y.; Wang, J.; Nguyen, M.; Altintas, I. (2019). PEnBayes: A multi-layered ensemble approach for learning bayesian network structure from big data. Sensors (Switzerland), 19(20). https://doi.org/10.3390/s19204400

Turesson, H.; Ribeiro, S.; Pereira, D.; Papa, J.; de Albuquerque, V. (2016). Machine learning algorithms for automatic classification of marmoset vocalizations. PLoS ONE, 11(9), 1–14. https://doi.org/10.1371/journal.pone.0163041

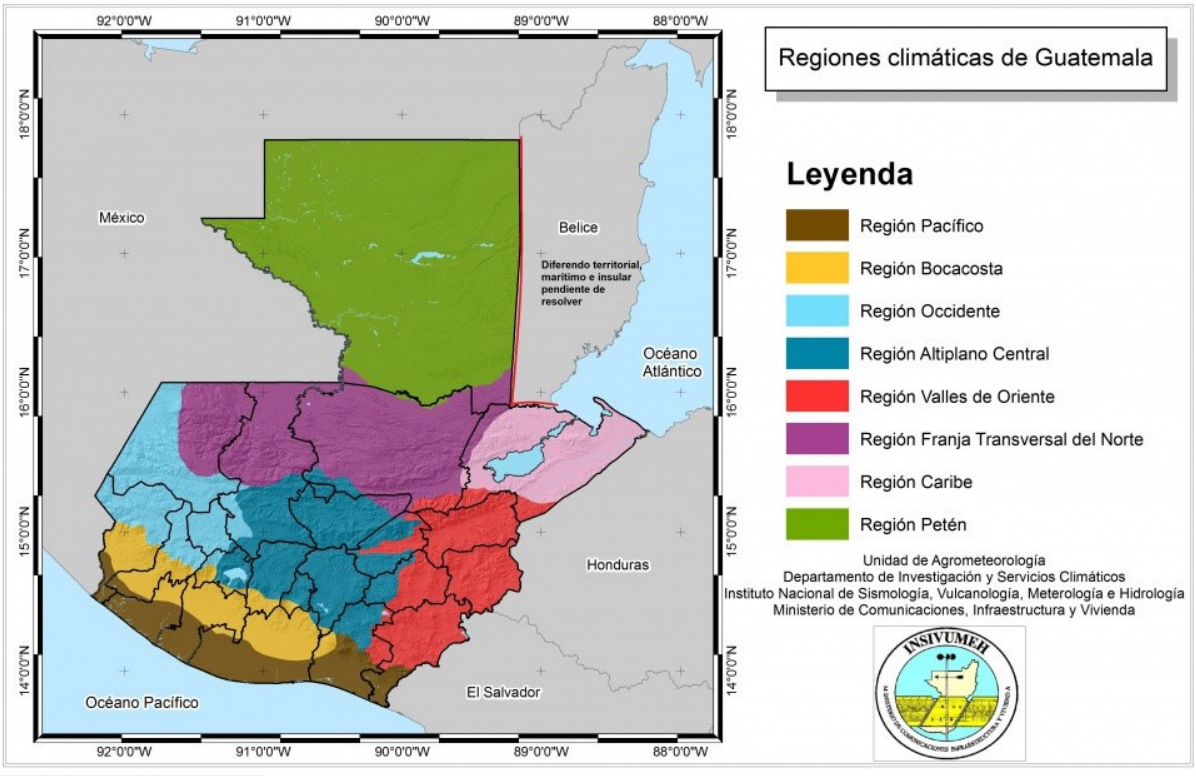
van Evert, F.; Fountas, S.; Jakovetic, D.; Crnojevic, V.; Travlos, I.; Kempenaar, C. (2017). Big Data for weed control and crop protection

Vu, Ch.; Kim, J. (2018). Human Motion Recognition by Textile Sensors Based on Machine Learning Algorithms

Wang, Sh.; Sun, S.; Xu, J. (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12

Zhang, Q; Wang, L. Xu, Z. (2017). Open source machine-learning algorithms for the prediction of optimal cancer drug therapies.

ANEXOS



Regiones climáticas de Guatemala.

Fuente: Unidad de Agrometeorología, INSIVUMEH.