Algoritmos de Minería de Datos para la Predicción del Comportamiento de Indicadores Ambientales en el Observatorio Ambiental de Matanzas.

Ing. Manuel Alejandro Naranjo Rey[1], Ing. Adrián Marcos Quintana[2], Ing. Eduardo Berrio Turiño[3]

1. Universidad de Matanzas – Sede "Camilo Cienfuegos", Vía Blanca Km.3, Matanzas, Cuba. manuel.naranjo@umcc.cu

2. Universidad de Matanzas – Sede "Camilo Cienfuegos", Vía Blanca Km.3, Matanzas, Cuba.

3. Universidad de Matanzas – Sede "Camilo Cienfuegos", Vía Blanca Km.3, Matanzas, Cuba.

Monografías 2020

ISBN: 978-959-16-4472-5

Resumen

La necesidad de adoptar adecuados enfoques empresariales para lograr un mejor desempeño ambiental constituye una tarea inminente. Desarrollar modelos predictivos para indicadores ambientales constituye el objetivo principal de esta investigación. Una herramienta informática que apoye la toma de decisiones permitirá una eficiente gestión ambiental empresarial, de forma tal que se eviten errores que se comenten en la actualidad. Para ello se emplearon tecnologías que demostraron ser competentes para el logro del objetivo de la investigación. La aplicación de técnicas de minería de datos permitió capturar los patrones pasados y replicarlos, además de realizar estimaciones con datos nuevos o fuera de muestra, así como inferir comportamientos y resultados futuros, en aras de anticipar posibles situaciones de deterioro que comprometan la sostenibilidad ambiental. Los experimentos diseñados para comparar los resultados en la clasificación al emplear los modelos predictivos, demuestran que el porcentaje de error oscila entre el 2% y el 3%, lo que demuestra un grado muy bueno (alto), de acuerdo con las escalas de comprobación. Por lo que podemos afirmar que la implementación de herramientas predictivas constituye un paso significativo hacia al objetivo estratégico de convertir información en conocimiento, evidentemente dicho conocimiento será trascendental para un mejor desempeño futuro.

***Palabras claves**: observatorios ambientales; predicción; inteligencia artificial; minería de datos.*

ISBN:

Introducción

El análisis de indicadores ambientales permite no solo el análisis desde el punto de vista geográfico, sino el profundo estudio del comportamiento de estos en el tiempo, para analizar las tendencias y posibles valores futuros. Realizar una correcta predicción de los datos, posibilitará una valoración futura de las consecuencias de las acciones realizadas en la actualidad, colaborando con la toma de decisiones.

En el camino hacia la búsqueda de soluciones y la prevención de la crisis en el futuro, algunos investigadores como (Selpa & Espinosa, 2009) afirman que en el mundo actual se hace imprescindible el uso de herramientas y procesos que ayuden al correcto desenvolvimiento de las entidades empresariales en lo que a su gestión ambiental se refiere. Naturalmente, también en otras áreas de la dirección organizacional, pues la empresa es una solamente, que funciona como un sistema integrado de elementos y procesos interconectados.

De ahí la importancia de contar con herramientas que ayuden en el análisis de la información, como son los sistemas de apoyo para la toma de decisiones cuyo propósito es ayudar a la administración para que marque tendencias, señale problemas y tome decisiones inteligentes. Su función básica es recolectar datos operacionales del negocio y reducirlos a una forma que pueda ser usada para analizar el comportamiento del mismo.

A lo anterior se agrega que, en la mayoría de los casos, lo que constituye el detonante de una decisión es el tiempo límite en el que se debe tomar. Lo cual hace que el verdadero objetivo de un sistema de apoyo a las decisiones sea proporcionar la mayor cantidad de información relevante en el menor tiempo posible, con el fin de decidir lo más adecuado.

En cualquier caso, la gestión ambiental cubana ha estado caracterizada por cambios bruscos e inesperados en direcciones muchas veces contrapuestas, que nos han llevado, en los últimos años, a replantearnos el empleo de las técnicas normalmente utilizadas para el tratamiento de una realidad que de tan cambiante se ha convertido en incierta.

La constante mutabilidad a la que se ven sometidos los fenómenos ambientales no permite, en la mayor parte de los casos, tomar en consideración datos del pasado para poder establecer inferencias del futuro. Es por ello que la preparación de una decisión, simple o compleja, se convierte en una actividad organizativa del pensamiento en la que inevitablemente se combina intuición y lógica.

Adicionalmente, son muchas las ocasiones en que está latente la ocurrencia de problemas relacionados con la obtención de resultados estadísticamente significativos, derivados del uso inapropiado de técnicas estadísticas y econométricas, además de la recurrente ausencia de datos o en otros casos y la duplicidad de los mismos.

Luego, se está en presencia de modelos econométricos muy útiles, desde el punto de vista de las relaciones de causalidad que describen, pero que pierden su fuerza al mostrar resultados pobres en sus parámetros estadísticos formales o al utilizarse para plazos largos, cuestión no concebida para algunos de estos modelos.

Es en este sentido, que los modelos predictivos de minería de datos son apropiados para tratar problemas que tienen evidentes relaciones no lineales a largo plazo y donde se requieren pronósticos con un elevado nivel de fiabilidad formal (bondad del ajuste) y confiabilidad (apropiada selección de variables a relacionar).

La inexistencia de una herramienta informática que asista la toma de decisiones y viabilice la actividad humana en la gestión ambiental, constituye un obstáculo en el objetivo estratégico de convertir la información en conocimiento; dicho conocimiento será trascendental para lograr un desarrollo sostenible. Por lo que, para la gestión ambiental cubana, contar con herramientas que apoyan la toma de decisiones es crucial para alcanzar el tan ansiado desarrollo sostenible.

Principales definiciones asociadas a la minería de datos

La minería de datos es un proceso ampliamente utilizado en la investigación de diferentes bigdatas. Su uso va en ascenso exponencial dado que con el pasar de los años se acumulan cada vez más datos. Diversos trabajos similares a esta investigación se han realizado en diferentes campos de aplicación.

- Modelos ARIMA univariante de series temporales para la producción y demanda de agua en el distrito de Lambayeque, periodo 2002 – 2017. El presente estudio tuvo como objetivo principal determinar el modelo univariante que permita predecir el comportamiento de la producción y demanda de agua en el distrito de Lambayeque, periodo 2002 al 2017. El análisis de los datos de la serie, se realizó mediante la metodología de Box – Jenkins, para identificar el modelo que mejor se adecue a los datos observados. En función a ello se determinó: El mejor modelo que explica el comportamiento de la producción de agua en el periodo indicado es el modelo SARMA (1,0,2) (1,0,0), con un Error Cuadrático Medio = 16 932.67, con Error Absoluto Medio = 13 254.61, el Error Porcentual Absoluto Medio = 3.98% y con coeficientes estimados AR (1) = 0.608, MA (1) = 0.299, MA (2) = -0.248, SAR

(1) = 0.215. Mientras que el modelo ARIMA (0,1,1), con un Error Cuadrático Medio = 3 062.74, con Desviación Absoluta de la media = 2 303.09, el Porcentaje de Error Medio Cuadrado Absoluto = 1.44% y con coeficientes estimados MA (1) = 0.377 es el que explica mejor el comportamiento de la demanda de agua en el periodo analizado. (López Jiménez & Villanueva Vásquez, 2020)

- Desarrollo e implementación de modelos de Machine Learning para aplicaciones de gestión y eficiencia energética. El propósito del proyecto es el desarrollo de un sistema de Aprendizaje Automático que permita realizar predicciones de consumo de suministros eléctricos. También debe permitir la detección de errores de datos de series temporales de consumo para su posterior análisis y corrección. Todo ello tomando como datos de entrenamiento el histórico de series temporales de cada uno de los suministros eléctricos proporcionados por las comercializadoras. (Ruiz Brückel, 2020)

- Análisis de datos en entornos inteligentes basados en el Internet de las cosas. La investigación e innovación contribuyen de manera decisiva a la lucha contra el cambio climático. Las TIC pueden reducir un 20 % de las emisiones mundiales de CO2 de aquí a 2030. El Aprendizaje Automático es útil para detectar ineficiencias de las ciudades modernas que contribuyen a la inestabilidad climática. Análisis iniciales sugieren que la conversión a edificios inteligentes gracias a la sensorización del Internet de las Cosas (IdC), junto con el análisis de datos, podría ser una opción para abordar estos problemas. Para ello, hemos identificado las siguientes necesidades de los edificios: mejorar las decisiones de gasto, reducir el consumo de energía, mejorar la eficiencia operativa y capacitar a sus usuarios con conocimientos energéticos; y las siguientes necesidades relativas a los análisis: cumplir los requisitos Big Data (volumen, velocidad, variedad, etc.), proporcionar mecanismos de fusión de datos, identificación de patrones de movilidad humana, reducción de información redundante en tiempo real, mejora de la predicción de series temporales mediante la selección de características y la gobernanza de datos. (González Vidal, 2020)

- Análisis sobre el uso de la red social Facebook en el proceso de enseñanza-aprendizaje por medio de la ciencia de datos. Hoy en día, los avances tecnológicos están provocando el desarrollo de nuevas estrategias de enseñanza-aprendizaje. De hecho, las redes sociales están adquiriendo gran relevancia durante la planeación de las actividades escolares. En particular, esta investigación mixta analiza el uso de Facebook como medio de difusión, comunicación, aprendizaje, interacción y colaboración durante la realización de las prácticas de laboratorio en la asignatura “Desarrollo de aplicaciones para los negocios”. La minería de datos permite establecer los modelos predictivos sobre el impacto de Facebook durante el diseño de la interfaz web considerando las técnicas bayesiana y árbol de decisión (ciencia de datos). La muestra está compuesta por 69 estudiantes de la Licenciatura en Gestión de Negocios y Tecnologías de Información. Por medio del enfoque cuantitativo y cualitativo, este estudio analiza el empleo de esta red social en el proceso de enseñanza-aprendizaje relacionado con

> el diseño de la interfaz web, las instrucciones HTML, el lenguaje de programación PHP, la aplicación WampServer y la base de datos MYSQL. Asimismo, el método ANOVA evalúa el rendimiento académico de los grupos experimental y control por medio de la calificación en el proyecto práctico. Los resultados obtenidos permiten afirmar que Facebook representa una alternativa tecnológica para mejorar la organización e implementación de las experiencias educativas en el siglo XXI. (Salas Rueda & Salas Rueda, 2019)

La minería de datos es un campo de la estadística y las ciencias de la computación referido al proceso de detectar la información procesable de los conjuntos grandes de datos. El término es un concepto de moda, y es frecuentemente mal utilizado para referirse a cualquier forma de datos a gran escala o procesamiento de la información. En el uso de la palabra, el término clave es el descubrimiento, comúnmente se define como "la detección de algo nuevo", para esto utiliza el análisis matemático para deducir los patrones y tendencias que existen. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional porque las relaciones son demasiado complejas o porque hay demasiados datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en conocimiento.

Ciertamente, la minería de datos bebe de la estadística, de la que toma las siguientes técnicas (Maimon & Lior, 2010):

- Análisis de varianza, mediante el cual se evalúa la existencia de diferencias significativas entre las medias de una o más variables continuas en poblaciones distintas.

- Regresión: define la relación entre una o más variables y un conjunto de variables predictoras de las primeras.

- Análisis de agrupamiento o clustering: permite la clasificación de una población de individuos caracterizados por múltiples atributos (binarios, cualitativos o cuantitativos) en un número determinado de grupos, con base en las semejanzas o diferencias de los individuos.

- Análisis discriminante: permite la clasificación de individuos en grupos que previamente se han establecido, permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto una mejor identificación de cuáles son las variables que definan la pertenencia al grupo.

- Series de tiempo: permite el estudio de la evolución de una variable a través del tiempo para poder realizar predicciones, a partir de ese conocimiento y bajo el supuesto de que no van a producirse cambios estructurales.

Las técnicas más representativas son:

- Redes neuronales. Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

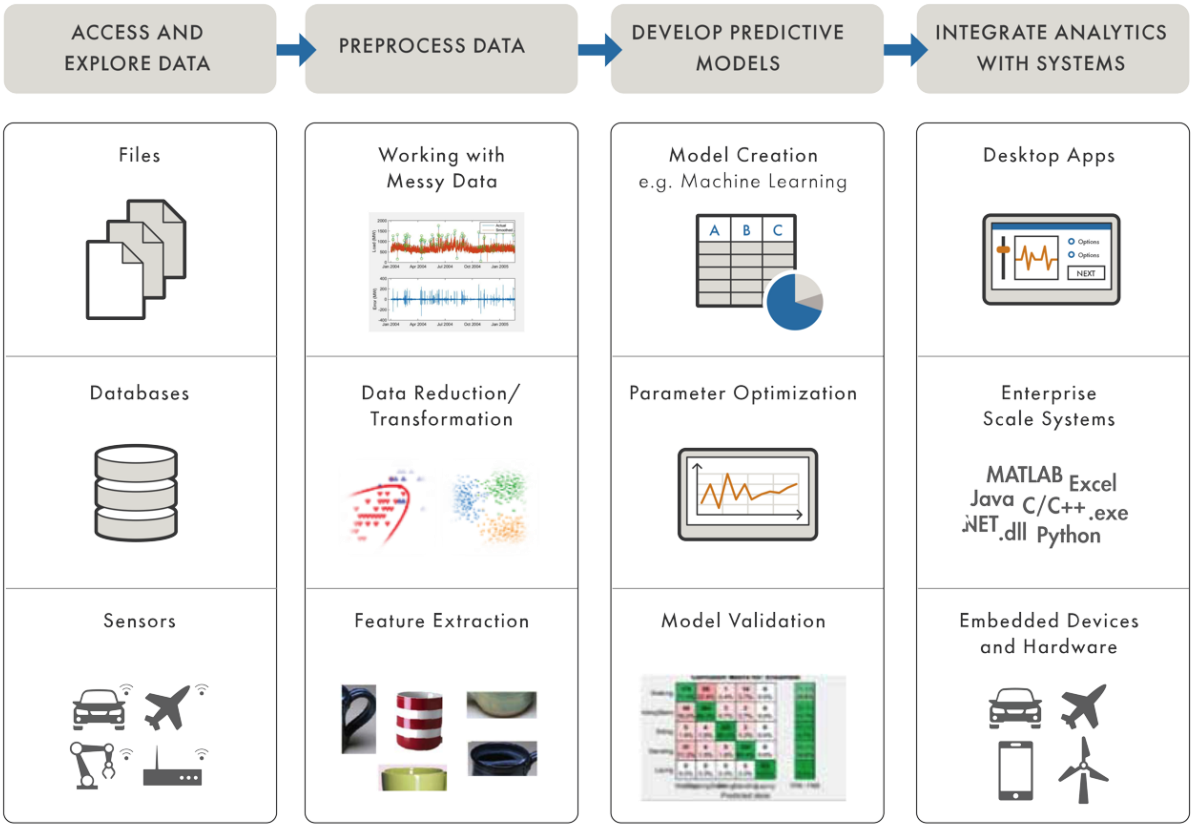
- Regresión lineal. Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

- Árboles de decisión. Es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

- Reglas de asociación. Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

- Series temporales. Son utilizadas para el análisis de la relación causal entre diversas variables que cambian con el tiempo y se influyen entre sí.

El análisis predictivo es un área de la minería del dato que pretende extraer conocimiento que le permita predecir tendencias y patrones de comportamiento. A menudo una circunstancia desconocida de interés se va a producir en el futuro, pero el análisis predictivo se puede aplicar igualmente a lo desconocido tanto en el pasado, el presente o el futuro. Por ejemplo, identificar sospechosos después de haberse producido un crimen o un fraude con tarjeta de crédito. Lo fundamental del análisis predictivo está en identificar relaciones entre las variables explicativas y las variables predictivas del pasado de forma que se pueda escalar a lo que está por ocurrir. Es importante advertir, en cualquier caso, que la fiabilidad y usabilidad de los resultados dependerán mucho del nivel de análisis del dato y la calidad de las hipótesis. (G., 2019)



Ilustracion1 Flujo de trabajo de un análisis predictivo. Fuente: https://es.mathworks.com/discovery/predictive-analytics.html#c%C3%B3mo-funciona

Los modelos predictivos son modelos de la relación entre el rendimiento específico de un sujeto en una muestra y uno o más atributos o características del mismo sujeto. El objetivo del modelo es evaluar la probabilidad de que un sujeto similar tenga el mismo rendimiento en una muestra diferente. Esta categoría engloba modelos en muchas áreas como el marketing, donde se buscan patrones de datos ocultos que respondan preguntas sobre el comportamiento de los clientes o modelos de detección de fraude. Los modelos predictivos a menudo ejecutan cálculos durante las transacciones en curso, por ejemplo, para evaluar el riesgo o la oportunidad de un cliente o transacción en particular, de forma que aporte conocimiento a la hora de tomar una decisión. Gracias

a los avances de ingeniería en el análisis de grandes volúmenes de datos estos modelos son capaces de simular el comportamiento humano frente a estímulos o situaciones específicas.

Desarrollo del modelo

Los datos que se utilizan en esta investigación son extraídos de la web Banco Mundial de Datos. Esta web ofrece datos de acceso abierto y gratuito sobre el desarrollo en el mundo. Fueron seleccionados para el entrenamiento de los modelos los datos Temperatura Mensual (˚C) – Cuba del Centro de Análisis de Información, División de Ciencias Ambientales del Laboratorio Nacional de Oak Ridge (Tennessee, Estados Unidos).

Esta base de datos posee la temperatura en Cuba desde 1901 hasta 2016, reporta un total de 1392 observaciones puesto que tiene una frecuencia mensual.

Una serie de tiempo es una lista de unidades de tiempo ordenadas tales como fechas, semestres o trimestres, cada una de las cuales se asocia a un valor. Las series de tiempo son un modo estructurado de representar datos. Visualmente, es una curva que evoluciona a lo largo del tiempo. Por ejemplo, las ventas diarias de un producto pueden representarse como una serie de tiempo. El pronóstico de las series de tiempo significa que extendemos los valores históricos al futuro, donde aún no hay mediciones disponibles. Existen dos variables estructurales principales que definen un pronóstico de serie de tiempo, el período, que representa la frecuencia con la que se miden los datos y el horizonte, que representa la cantidad de períodos por adelantado que deben ser pronosticados.

Las series temporales se pueden definir como un caso particular de los procesos estocásticos, ya que un proceso estocástico es una secuencia de variables aleatorias, ordenadas y equidistantes cronológicamente referidas a una característica observable en diferentes momentos. El análisis de series temporales explica el hecho de que los puntos de datos tomados a lo largo del tiempo pueden tener una estructura interna (como la autocorrelación, la tendencia o la variación estacional) que debe tenerse en cuenta (RStudio, 2017).

Si todas las variables aleatorias que componen el proceso están idénticamente distribuidas, independientemente del momento del tiempo en que se estudie el proceso, entonces la serie es estacionaria.

Es decir, la función de distribución de probabilidad de cualquier conjunto de k variables (siendo k un número finito) del proceso debe mantenerse estable (inalterable) al desplazar las variables s

períodos de tiempo tal que, si P(Yt+1,Yt+2,…,Yt+k) es la función de distribución acumulada de probabilidad, entonces: (Parra, 2019)

$$P(Yt+1,Yt+2,\ldots,Yt+k)=P(Yt+1+s,Yt+2+s,\ldots,Yt+k+s),\ \forall t,k,s$$

Sin embargo, la versión estricta de la estacionariedad de un proceso suele ser excesivamente restrictiva para las necesidades prácticas de un economista. Es por ello que generalmente nos conformaremos con un concepto menos exigente, el de estacionariedad en sentido débil o de segundo orden, la cual se da cuando la media del proceso es constante e independiente del tiempo, la varianza es finita y constante, y el valor de la covarianza entre dos periodos depende únicamente de la distancia o desfase entre ellos, sin importar el momento del tiempo en el cual se calculan. (Parra, 2019)

Una serie puede ser no estacionaria por una variación en la media, una variación en la varianza o por la presencia de estacionalidad. Esto significa que si existe alguno de estos casos es necesario aplicar transformaciones en la serie. A simple vista podemos observar que la serie no es estacionaria en media.

En esta serie temporal no se analiza la estacionalidad puesto que tiene como frecuencia 1, lo que significa que los datos se miden anualmente y no tiene sentido realizar análisis estacionales en estos casos. Muchas series presentan cierta periodicidad o, dicho de otro modo, variación de cierto periodo(semestral, mensual …). Por ejemplo, las temperaturas aumentan en verano y disminuyen en invierno. Estos tipos de efectos son fáciles de entender y se puede medir explícitamente o incluso se pueden eliminar del conjunto de los datos, desestacionalizando la serie original. Estos efectos suelen aparecer en variables medioambientales por los cambios que producen las estaciones en estas.

Los métodos de descomposición estacional son eminentemente descriptivos. Tratan de separar la serie en subseries correspondientes a la tendencia, la estacionalidad y el ruido (componente aleatorio).

En ocasiones tendencia y estacionalidad se enmascaran, a veces una tendencia marcada puede no dejarnos ver la estacionalidad, y viceversa. Los métodos de descomposición estacional separan tendencia, estacionalidad y ruido, pero no predicen. Para predecir es necesario combinarlos con métodos de ajuste de tendencia. De esta forma realizamos un ajuste de tendencia con el fin de obtener un modelo extrapolable, y le añadimos la estacionalidad.

El primer paso a seguir a la hora de descomponer una serie es determinar cómo se combinan sus componentes. Las combinaciones aditiva y multiplicativa son las más habituales. Decimos que estamos en presencia de una aditiva cuando a pesar del crecimiento de la tendencia, la varianza y la media se mantienen estáticas, en cambio las multiplicativas son cuando la varianza y la media varias en consecuencia de la tendencia. En una serie temporal Xt es una función que depende de cuatro componentes:

Componentes aditivas: $Xt = Ct + Tt + St + Et$

Componentes multiplicativas: $Xt = Ct \times Tt \times St \times Et$

Donde:

Tendencia (Tt),

Ciclo (Ct),

Componente estacional (St),

Componente irregular o ruido (Et).

Para aplicar un modelo ARIMA ajustado es necesaria la transformación de la serie temporal en otra que sea aproximadamente estacionaria. Para esto se emplean técnicas como la diferenciación y logaritmos en dependencia de la no estacionariedad.

Según las fuentes consultadas, las pruebas más utilizadas para este fin son aplicadas a nuestra serie para el análisis de raíz unitaria.

Prueba de Dickey-Fuller aumentada (ADF) es una versión aumentada de la prueba Dickey-Fuller para un conjunto más amplio y más complejo de modelos de series de tiempo. La estadística Dickey-Fuller Aumentada (ADF), utilizada en la prueba, es un número negativo. Cuanto más negativo es, más fuerte es el rechazo de la hipótesis nula de que existe una raíz unitaria para un cierto nivel de confianza.

Esta prueba da como resultado un p-value< 0.01, tomando un nivel de significancia del 95%, se rechaza la hipótesis nula por ser 0.01<0.05. Por tanto, la serie es estacionaria.

La siguiente prueba que se realizará será Kwiatkowski-Phillips-Schmidt-Shin. Su hipótesis nula es que no posee raíz unitaria.

Esta prueba da como resultado un p-value< 0.01, tomando un nivel de significancia del 95%, se rechaza la hipótesis nula por ser 0.01<0.05. Por tanto la serie no es estacionaria.

Posteriormente se aplica la prueba Phillips-Perron cuya hipótesis nula es que posee raíz unitaria. Se basa en la prueba de Dickey-Fuller. Al igual que la prueba de Dickey-Fuller aumentada, la prueba de Phillips-Perron aborda la cuestión de que el proceso de generación de datos podría tener un orden superior de autocorrelación que es admitido en la ecuación de prueba. Mientras que la prueba de Dickey-Fuller aumentada aborda esta cuestión mediante la introducción de retardos de como variables independientes en la ecuación de la prueba, la prueba de Phillips-Perron hace una corrección no paramétrica a la estadística t-test. El ensayo es robusto con respecto a la autocorrelación y heterocedasticidad en el proceso de alteración de la ecuación de prueba

.

Este test da como resultado p-value< 0.01 lo cual implica la existencia de raíz unitaria y la serie es estacionaria.

Una manera simple de ver una diferencia única (o "de primer orden") es verla como x (t) - x (t-k) donde k es el número de rezagos para volver. Las diferencias de orden superior son simplemente la reaplicación de una diferencia a cada resultado anterior. Si hay alguna duda sobre diferenciar o no, o sobre cuantas veces hacerlos, se calcula la varianza de la serie original y de la serie sometida a diferentes diferenciaciones, tomando como diferenciación adecuada aquella para la que la varianza es mínima.

Una vez obtenemos la serie diferenciada aplicamos nuevamente las pruebas de raíz unitaria pero esta vez sobre la serie diferenciada.

Prueba	P-Value
adf.test	<0.01
kpss.test	>0.1
pp.test	<0.01

Tabla1: Elaboración propia

Comparando los resultados obtenidos con sus respectivas hipótesis obtenemos que la serie ya es estacionaria.

En otros ejemplos observamos que en ocasiones las pruebas de estacionariedad no coinciden en sus hipótesis, en estos casos nos afecta generalmente una ruptura estructural. Cuando esto sucede el test de Dickey Fuller Aumentado dará falsos reportes de series no estacionarias. Para saber si los datos poseen un cambio en la estructura se utilizará la librería urca. Esta librería contiene la prueba de Zivot and Andrews que permite además de conocer si una serie tiene raíz unitaria, saber si tiene ruptura estructural y en qué punto de esta existe ya sea en el intercepto, la tendencia lineal o en ambos.

Posteriormente se analiza la estacionalidad de la serie si esta tiene una frecuencia superior a una medición por año. Si hay presencia de esta, significa que las estaciones influyen directamente en el valor de la variable. En nuestro caso de antemano sabemos que influyen puesto que la temperatura aumenta en verano y disminuye en invierno. Para evaluar la estacionalidad en una serie se emplean las siguientes pruebas: Prueba de raíz unitaria de Osborn, Chui, Smith y Birchenhall; Prueba de raíz unitaria de Hylleberg, Engle, Granger y Yoo y la Prueba de raíz unitaria de Canova y Hansen.

Por los resultados obtenidos en la aplicación de estas pruebas es necesaria una diferenciación en la parte estacional.

Modelo ARIMA

En estadística y econometría, en particular en series temporales, un modelo autorregresivo integrado de promedio móvil o ARIMA (acrónimo del inglés autoregressive integrated moving average) es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Se trata de un modelo dinámico de series temporales, es decir, las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes. Fue desarrollado a finales de los sesenta del siglo XX. Box y Jenkins han desarrollado modelos estadísticos para series temporales que tienen en cuenta la dependencia existente entre los datos, esto es, cada observación en un momento dado es modelada a partir de los valores anteriores. Los análisis se basan en un modelo explícito.

El modelo ARIMA necesita identificar los coeficientes y número de regresiones que se utilizarán. Este modelo es muy sensible a la precisión con que se determinen sus coeficientes.

Se suele expresar como ARIMA(p,d,q) donde los parámetros p, d y q son números enteros no negativos que indican el orden de las distintas componentes del modelo — respectivamente, las componentes autorregresivas, integrada y de media móvil. Cuando alguno de los tres parámetros es cero, es común omitir las letras correspondientes del acrónimo — AR para la componente autorregresiva, I para la integrada y MA para la media móvil. Por ejemplo, ARIMA(0,1,0) se puede expresar como I(1) y ARIMA(0,0,1) como MA(1).

El modelo ARIMA puede generalizarse aún más para considerar el efecto de la estacionalidad. En ese caso, se habla de un modelo SARIMA (seasonal autoregressive integrated moving average).

El modelo ARIMA (p,d,q) se puede representar como:

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

en donde d corresponde a las d diferencias que son necesarias para convertir la serie original en estacionaria, ϕ 1 , … , ϕ p son los parámetros pertenecientes a la parte "autorregresiva" del modelo, θ 1 , … , θ q los parámetros pertenecientes a la parte "medias móviles" del modelo, ϕ 0 es una constante, y εt es el término de error.

Se debe tomar en cuenta que:

$$\Delta Y_t = Y_t - Y_{t-1}$$

Condiciones Necesarias para el Modelo ARIMA

Los datos deben ser estacionarios, esto significa que las propiedades de la serie no dependen del momento en que se capturan. Una serie de ruido blanco y series con comportamiento cíclico también pueden considerarse series estacionarias.

Los datos deben ser univariantes, ARIMA trabaja en una sola variable. La regresión automática tiene que ver con la regresión de los valores pasados.

El modelo ARIMA permite describir un valor como una función lineal de datos anteriores y errores debidos al azar, además, pueden incluir un componente cíclico o estacional. Es decir, debe contener todos los elementos necesarios para describir el fenómeno. Box y Jenkins recomiendan como mínimo 50 observaciones en la serie temporal.

Identificación práctica del modelo: Identificar un modelo significa utilizar los datos recogidos, así como cualquier información de cómo se genera la serie temporal objeto de estudio, para sugerir un conjunto reducido de posibles modelos, que tengan muchas posibilidades de ajustarse a los datos. Ante una serie temporal empírica, se deben encontrar los valores (p, d, q) más apropiados.

Como la serie temporal presentó una tendencia, lo primero fue aplicar una diferenciación, de orden d. Una vez diferenciada la serie, una buena estrategia consiste en comparar los correlogramas de la función de autocorrelación (ACF) y la función de autocorrelación parcial (ACFP), proceso que suele ofrecer una orientación para la formulación del modelo orientativo.

Los procesos autorregresivos presentan función de autocorrelación parcial (ACFP) con un número finito de valores distinto de cero. Un proceso AR(p) tiene los primeros p términos de la función de

autocorrelación parcial distintos de cero y los demás son nulos. En la práctica se considera que una muestra dada proviene de un proceso autorregresivo de orden p si los términos de la función de autocorrelación parcial son casi cero a partir del que ocupa el lugar p. Un valor se considera casi cero cuando su módulo es inferior a $2/\sqrt{T}$. Los programas de ordenador constituyen la franja ($-2/\sqrt{T}$; $2/\sqrt{T}$) y detectan los valores de la ACFP que caen fuera de ella.

Los procesos de medias móviles presentan función de autocorrelación con un número finita de valores distintos de cero. Un proceso MA(q) tiene los primeros q términos de la función de autocorrelación distintos de cero y los demás son nulos. Las dos propiedades descritas son muy importantes con vistas a la identificación de un proceso mediante el análisis de las funciones de autocorrelación y autocorrelación parcial.

Como el modelo ARIMA tendrá una parte estacional se analizarán los primeros retardos para la parte estacionaria y los retardos en cada periodo para el análisis de la parte estacional.

A partir de este análisis se derivan las siguientes variantes para el modelo ARIMA(p,d,q),(P,D,Q):

ARIMA(2,1,2)(1,1,1)

ARIMA(2,1,2)(1,1,2)

ARIMA(2,1,2)(2,1,1)

ARIMA(2,1,2)(2,1,2)

Método de suavizado exponencial

El suavizado exponencial puede utilizarse para predecir a corto plazo en las series temporales, son técnicas de tipo predictivo. Proporcionan previsiones razonables para horizontes de predicción inmediatos. Además, los resultados que se obtienen con ellos son satisfactorios, incluso cuando no se dispone de un gran número de datos históricos. A diferencia de los métodos de descomposición estacional, para aplicar los de suavizado no es necesario convertirla en una serie aproximadamente estacionaria. Dentro de estos últimos existen modelos para series no afectadas por tendencia ni estacionalidad, para series con tendencia y para series con tendencia y estacionalidad.

El modelo Holt-Winters incorpora un conjunto de procedimientos que conforman el núcleo de la familia de series temporales de suavizado exponencial. A diferencia de muchas otras técnicas, este modelo puede adaptarse fácilmente a cambios y tendencias, así como a patrones estacionales. En comparación con otras técnicas, como ARIMA, el tiempo necesario para calcular el pronóstico es considerablemente más rápido. Su aplicación en entornos de negocio es muy común, se utiliza habitualmente por muchas compañías para pronosticar la demanda a corto plazo cuando los datos de venta contienen tendencias y patrones estacionales de un modo subyacente.

El método de Holt-Winters utiliza las siguientes fórmulas (Vásquez Mejía & Chavez Gonzales, 2019):

La fórmula para la estimación del nivel es:

$Nt = \alpha * + (1- \alpha) * (Nt-1 + Tt-1)$ (Fórmula 1)

Donde:

Nt es el valor suavizado del nivel del periodo t

α es la constante de suavizamiento del nivel

Xt es la observación histórica en el periodo t

El subíndice s representa la longitud de la estacionalidad, normalmente un año)

St-s es el valor suavizado de la estacionalidad del periodo t-s

Nt-1 es el valor suavizado del nivel del periodo t-1

Tt-1 es la estimación de la tendencia en el periodo t-1

La fórmula para estimación de la tendencia es:

$Tt = \beta (Nt - Nt\text{-}1) + (1- \beta) Tt\text{-}1$ (Fórmula 2)

Donde:

Tt es la estimación de la tendencia en el periodo t

β es la constante de suavizamiento de la tendencia

Nt y Nt-1 son los valores suavizados del nivel del periodo t y t-1 respectivamente

Tt-1 es la estimación de la tendencia en el periodo t-1

La fórmula para la estimación de la estacionalidad es:

$Et = \gamma * + (1- \gamma) * Et\text{-}s$ (Fórmula 3)

Donde:

Et es la estimación de la estacionalidad en el periodo t

γ (letra griega gama) es la constante de suavizamiento de la estacionalidad

Xt es la observación histórica en el periodo t

Nt es el valor suavizado del nivel del periodo t

El subíndice s representa la longitud de la estacionalidad

Et-s es la estimación de la estacionalidad en el periodo t – s

Y finalmente la fórmula para la previsión es:

$F_{t+1} = (N_t + T_t) * E_{t-s+1}$ (Fórmula 4)

Donde:

F_{t+1} es la previsión del periodo t

N_t es el valor suavizado del nivel del periodo t

T_t es la estimación de la tendencia en el periodo t

Es subíndice s representa la longitud de la estacionalidad

E_{t-s+1} es la estimación de la estacionalidad en el periodo $t – s + 1$

Para el método de suavizado exponencial el modelo HoltWinters(beta = FALSE) resultó ser el más adecuado para estos datos.

Análisis de los modelos

Con la propuesta de modelos ya terminada solo queda analizarla con el fin de corroborar los resultados que se obtienen. En las diferentes pruebas realizadas se utilizaron los datos Temperatura Mensual (˚C) – Cuba del Centro de Análisis de Información, División de Ciencias Ambientales del Laboratorio Nacional de Oak Ridge (Tennessee, Estados Unidos), con el objetivo de saber si los resultados ofrecidos por el sistema son los acertados. Se realizaron diferentes pruebas, con el objetivo de mejorar los resultados con cada una de ellas.

La prueba de Ljung-Box es un tipo de prueba estadística de si un grupo cualquiera de autocorrelaciones de una serie de tiempo son diferentes de cero. En lugar de probar la aleatoriedad en cada retardo distinto, esta prueba la aleatoriedad "en general" basado en un número de retardos.

Esta prueba también es conocida como la prueba Q de Ljung-Box, y está estrechamente relacionada con la prueba de Box-Pierce. Esta es una versión simplificada de la estadística de Ljung-Box para los cuales los estudios de simulación posteriores han demostrado un rendimiento deficiente.

La prueba de Ljung-Box se puede definir de la siguiente manera.

H0: Los datos se distribuyen de forma independiente (es decir, las correlaciones en la población de la que se toma la muestra son 0, de modo que cualquier correlación observada en los datos es el resultado de la aleatoriedad del proceso de muestreo).

Ha: Los datos no se distribuyen de forma independiente.

La estadística de prueba es:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{p_k^2}{n-k}$$

donde n es el tamaño de la muestra, p_kes la autocorrelación de la muestra en el retraso k y h es el número de retardos que se están probando. Por nivel de significación α, la región crítica para el rechazo de la hipótesis de aleatoriedad es

$$Q > x^2_{1-\alpha,h}$$

donde $x^2_{1-\alpha,h}$ es la α- cuantil de la distribución chi-cuadrado con m grados de libertad.

Para que el modelo seleccionado sea validado tiene que tener los residuales estacionarios, normalizados e independientes. Para esto se realiza la prueba de ruido blanco o Ljung-Box. Un ruido blanco es una serie tal que su media es cero, la varianza es constante y no se puede correlacionar.

E(at)=0

Var(at)=σ2a

cov(at,at+h)=0

Se trata de un proceso en el que todas sus variables son independientes.

Modelo ARIMA:

ARIMA(2,1,2)(1,1,1)

0.9674

ARIMA(2,1,2)(1,1,2)

0.9736

ARIMA(2,1,2)(2,1,1)

0.9864

ARIMA(2,1,2)(2,1,2)

0.9376

Los resultados de este test aceptan en todos los modelos la hipótesis nula. Esto significa que los residuales se distribuyen como un ruido blanco. Por tanto, estos presentan estacionariedad, normalidad e independencia, lo cual implica que se está en presencia de modelos adecuados para la predicción.

Modelo HoltWinters:

El modelo es validado de la misma forma el modelo ARIMA, solamente que en este caso la prueba se le debe hacer a los residuales de las predicciones.

HoltWinters(beta=FALSE)

p-value< 2.2e-16

El resultado de este test es p-value< 2.2e-16 rechazando la hipótesis nula. Esto significa que los residuales no se distribuyen como un ruido blanco. Por tanto, estamos en presencia de un modelo que no es adecuado para la predicción.

Para seleccionar el modelo que mejor ajuste posee nos apoyaremos en AIC. El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo. AIC maneja un sacrificio entre la bondad de ajuste del modelo y la complejidad del modelo. Se basa en la entropía de información: se ofrece una estimación relativa de la información pérdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos.

AIC no proporciona una prueba de un modelo en el sentido de probar una hipótesis nula, es decir AIC puede decir nada acerca de la calidad del modelo en un sentido absoluto.

El modelo que tenga menor AIC será el más ajustado a los datos, porque será menor la información perdida con dicho modelo.

ARIMA(2,1,2)(2,1,2)

aic = 3044.929

ARIMA(2,1,2)(2,1,1)

aic = 3043.881

ARIMA(2,1,2)(1,1,2)

aic = 3044.07

ARIMA(2,1,2)(1,1,1)

aic = 3042.96

Según los valores devueltos por el software por el criterio de selección aplicado el modelo 4 es el más ajustado a los datos estudiados.

Una vez seleccionado el modelo que mejor se ajusta se realizaran las predicciones para el próximo año.

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2017	23.71792	22.80461	24.63123	22.32113	25.11471
Feb 2017	23.70863	22.74742	24.66984	22.23858	25.17868
Mar 2017	24.71125	23.74230	25.68020	23.22936	26.19313
Apr 2017	25.67066	24.70073	26.64058	24.18728	27.15403
May 2017	26.89284	25.92226	27.86343	25.40846	28.37722
Jun 2017	27.84484	26.87410	28.81557	26.36022	29.32945
Jul 2017	28.26223	27.29116	29.23329	26.77711	29.74735
Aug 2017	28.33761	27.36642	29.30880	26.85230	29.82291

```
Sep 2017       27.92950    26.95803 28.90096 26.44377 29.41523

Oct 2017       27.04149    26.06989 28.01309 25.55555 28.52743

Nov 2017       25.34470    24.37284 26.31656 23.85837 26.83102

Dec 2017       24.14931    23.17731 25.12132 22.66276 25.63587
```

Aquí obtuvimos las predicciones para el próximo año, que como tiene una frecuencia mensual, fueron necesarias 12 predicciones. Las columnas siguientes son los intervalos de confianza para un 80 y un 95 por ciento de confianza respectivamente. Por lo que si queremos conocer cuál es la temperatura predicha para diciembre de 2017 será de 24.14 grados Celsius y con un intervalo de confianza de un 95% la temperatura en este mes estará entre 22.66 y 25.63 grados Celsius.

Para más claridad en la predicción es posible graficarla siguiendo la línea de los datos estudiados.

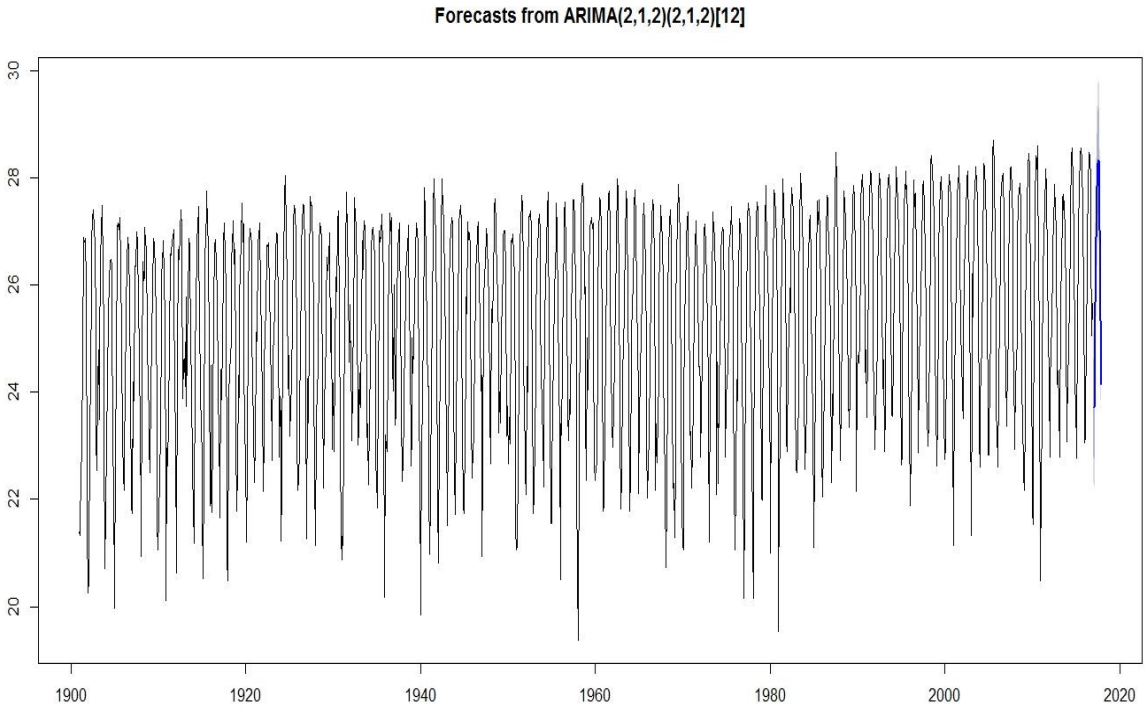


Ilustración2 Valores predichos por ARIMA Fuente: Elaboración propia

Para evaluar el modelo planteado para la solución será usado el indicador de Porcentaje de Error Medio Absoluto (MAPE) por su fácil interpretación, el cual se calcula con la siguiente ecuación:

$$\boldsymbol{MAPE} = \frac{\sum_{t=1}^{n} \left| \frac{(y_t - \hat{y_t})}{y_t} (100) \right|}{n}$$

Dónde:

$\boldsymbol{y_t}$: Es el valor observado (valor real del indicador)

$\boldsymbol{\hat{y_t}}$: Es el valor pronosticado (predicción del indicador)

n: Es la cantidad de observaciones

Al aplicar el MAPE se obtuvo en el modelo aproximadamente 2.108366% de error absoluto que al comparar con el esquema de clasificación se determina como alta precisión, por lo que es un modelo altamente confiable para la predicción de futuros indicadores.

% de Error MAPE	Clasificación del pronóstico
Menor de 10%	Alta precisión
10% -20%	Buena precisión
20% - 50%	Precisión razonable
Mayor del 50%	Poco fiable

Tabla2: Criterio de validación de los modelos de pronóstico. Fuente: (Lewis, 1982)

CONCLUSIONES

El estudio realizado sobre los antecedentes, el estado actual de la temática, la bibliografía y documentos relacionados con el objeto de estudio, permitió aportar los elementos necesarios para dar solución a la problemática planteada ya que los antecedentes encontrados, vinculados al tema no le dan solución al problema planteado por lo que no es factible su utilización. Con este trabajo se logró la implementación de modelos predictivos que permitirán procesar los datos y realizar inferencias futuras, demostrando la necesidad de un histórico de datos amplio para lograr valores de predicción óptimos. Para los métodos analizados el modelo ARIMA(2,1,2)(1,1,1) resultó ser el que mejores resultados arroja y los mismos demostraron que los modelos implementados son fiables y que sus pronósticos tienen un ¨alto grado de precisión¨. Estos son derivados de la simulación y no de la subjetividad de los investigadores, lo cual provee de solidez y rigor en la toma de decisiones, abriendo un mayor espectro para su uso a partir de sus propiedades estadísticas posibilitando que las predicciones del software sean muy cercanas a la realidad, permitiendo así emitir criterios acertados para evaluar una situación en un espacio de tiempo determinado.

ISBN:

Bibliografía

PATRICIO ARENADA, G. (11 de 9 de 2019). *RPubs.* Obtenido de https://www.rpubs.com/paraneda/predictivo

GONZÁLEZ VIDAL, A. (2020). *Análisis de datos en entornos inteligentes basados en el Internet de las cosas.* Murcia: Universidad de Murcia.

LÓPEZ JIMÉNEZ, K. J., & VILLANUEVA VÁSQUEZ, W. J. (2020). *Modelos Arima univariante de series temporales para la producción y demanda de agua en el distrito de Lambayeque, periodo 2002 – 2017.* Lambayeque.

ODED MAIMON, O., & ROKACH LIOR, R. (2010). *Data Mining and Knowledge Discovery Handbook.* New York.

PARRA RODRÍGUEZ, F. (2019). *Estadística y Machine Learning con R.* Bookdown. Obtenido de https://bookdown.org/content/2274/series-temporales.html

RUIZ BRÜCKEL, T. (2020). *Desarrollo e implementación de modelos de Machine Learning para aplicaciones de gestión y eficiencia energética.* Catalunya.

RSTUDIO. (26 de 11 de 2017). *RPubs*. Obtenido de https://rpubs.com/palominoM/series

SALAS RUEDA, R. A., & SALAS RUEDA, R. D. (2019). Análisis sobre el uso de la red social Facebook en el proceso de enseñanza-aprendizaje por medio de la ciencia de datos. *Revista de Comunicación de la SEECI*, 50.

SELPA NAVARRO, A. Y., & ESPINOSA CHONGOA, D. (2009). La gestión del capital de trabajo como proceso de la gestión financiera operativa. *Gestión Joven*, 24-33.

ISBN:

VÁSQUEZ MEJÍA, E. J., & CHAVEZ GONZALES, S. (2019). Trabajo Teórico Experimental Predicción del consumo de energía eléctrica residencial de la Región Cajamarca mediante modelos Holt-Winters. *Ingeniería Energética*, *40*(3), 181-191.

ISBN: