UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al Servicio del Pueblo

UNIDAD ACADÉMICA DE TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

CARRERA DE INGENIERIA

DESARROLLO DE UN PROTOTIPO PARA LA PREDICCIÓN DE NUEVOS CASOS DE COVID-19 EN EL ECUADOR MEDIANTE EL USO DE INTELIGENCIA ARTIFICIAL

TRABAJO DE TITULACIÓN O PROYECTO DE INTEGRACIÓN CURRICULAR PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS

AUTOR: JUAN ANDRES PAGUAY HURTADO

DIRECTOR: ANDRES SEBASTIAN QUEVEDO SACOTO

Azogues - Ecuador

2021

UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al Servicio del Pueblo

UNIDAD ACADÉMICA DE TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

CARRERA DE INGENIERÍA

DESARROLLO DE UN PROTOTIPO PARA LA PREDICCIÓN DE NUEVOS CASOS DE COVID-19 EN EL ECUADOR MEDIANTE EL USO DE INTELIGENCIA ARTIFICIAL

TRABAJO DE TITULACIÓN O PROYECTO DE INTEGRACIÓN CURRICULAR PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS

AUTOR: JUAN ANDRES PAGUAY HURTADO

DIRECTOR: ANDRES SEBASTIAN QUEVEDO SACOTO

AZOGUES - ECUADOR

2021

Azogues, 14 de julio de 2021

Asunto: Informe final de Tutoría de Trabajo de Titulación

Señora Economista
Nancy Peralta Idrovo
Auxiliar de Secretaría de la Unidad Académica de Tecnologías de la Información y Comunicación.
UNIVERSIDAD CATÓLICA DE CUENCA SEDE AZOGUES.
Su despacho. –

De mi consideración:

Por medio del presente me permito saludarle y a la vez indicarle que una vez culminada la revisión del trabajo de titulación del alumno Juan Andrés Paguay Hurtado, es mi deber conferir la nota de 50/50 al trabajo de titulación en mención, cumpliendo de esta manera con los parámetros establecidos por parte de nuestra Alma Mater como Tutor de la misma, dentro de la Unidad de Titulación.

Dicho proyecto lleva por nombre **"**DESARROLLO DE UN PROTOTIPO PARA LA PREDICCIÓN DE NUEVOS CASOS DE COVID-19 EN EL ECUADOR MEDIANTE EL USO DE INTELIGENCIA ARTIFICIAL**",** Previo a la obtención del título de Ingeniero de Sistemas, de la Unidad Académica de Tecnologías de la información y Comunicación.

Es menester informar que el presente trabajo de titulación tiene coincidencia del 0% de similitud de contenidos con otras fuentes, según reporte del sistema anti-plagio (Turnitin) de nuestra Universidad, reporte que se adjunta.

Por la atención que dé a la presente suscribo de Ud. Agradecido.

Atentamente.

Ing. Sebastián Quevedo Sacoto

TUTOR

Ciencia Digital
www.cienciadigital

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

Desarrollo de un prototipo para la predicción de nuevos casos de covid-19 en el ecuador mediante el uso de inteligencia artificial.

Development of a prototype for the prediction of new cases of covid-19 in ecuador through the use of artificial intelligence.

Juan Andrés Paguay Hurtado[1*],
Ingeniería de Sistemas[1]
Universidad Católica de Cuenca[1d]
*japaguayh26@est.ucacue.edu.ec
juanandres1435@gmail.com
https://orcid.org/0000-0001-5375-6475

Recibido: 10-05-2021 / Revisado: 15-06-2021 /Aceptado: 04-07-2019/ Publicado: 25-08-2021

DOI: https://doi.org/10.33262/cienciadigital.v5i3

Abstract

Introduction. Cases of coronavirus (Covid-19) around the world are increasing. The uncertainty of a figure close to reality generates anguish in the population. Objective. Propose the use of Artificial Intelligence (AI) to determine the increase in Covis-19 cases in Ecuador, applying this model will provide approximate information on coronavirus cases. Helping to keep the entire population informed about the spread of this virus. **Methodology.** The design of this research was quantitative, the population that was taken was 17,268,000 and the sample was the data of the infections of Covid-19 from the month of April to the month of December of the year 2020. For this, it was taken as data source the information published daily on the official website of the National Risk and Emergency Management Service. Using predictive models as support, these data were stored in a data set, to later be consolidated and later entered into an algorithm, which using time series will make predictions based on historical data using the weka software. The following article presents a model capable of predicting the close-to-reality number of coronavirus cases, achieving 80% effectiveness. So it can be stated that this model is very useful for making predictions within a given period. **Results.** After applying the prediction model, the most frequent results are the increases in Covid-19 infections with an increase of (1%) for each day that has elapsed. **Conclusion.** It was concluded that the cases will continue to increase over time since the majority of the population does not take the respective precautions and disrespects social distancing.

Keywords: Data Mining, Machine Learning, Time Series, Weka.

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

Resumen.

Introducción. Los casos de coronavirus (Covid-19) en el mundo entero, van cada vez en aumento. La incertidumbre de una cifra cercana a la realidad, genera angustia en la población. **Objetivo.** Plantear el uso de la Inteligencia Artificial (IA) para determinar el incremento de casos de Covis-19 en el ecuador, al aplicar este modelo se tendrá una información aproximada de los casos de coronavirus. Ayudando a tener informada a toda la población sobre la propagación de este virus. **Metodología.** El diseño de esta investigación fue cuantitativa, la población que se tomó fue 17.268.000 y la muestra fueron los datos de los contagios de Covid-19 desde el mes de abril hasta el mes de diciembre del año 2020. Para ello, se tomó como fuente de datos la información publicada diariamente en la página oficial del Servicio Nacional de Gestión de Riesgos y Emergencias. Utilizando como apoyo los modelos predictivos, se almacenaron estos datos en un data set, para luego ser consolidados y posteriormente introducirlos en un algoritmo, el cual utilizando series de tiempo realizará las predicciones en base a datos históricos mediante el software weka. El siguiente artículo, presenta un modelo capaz de predecir la cifra cercana a la realidad de casos de coronavirus, consiguiendo un 80% de efectividad. Por lo que se puede manifestar que este modelo resulta muy útil para realizar predicciones dentro de un periodo determinado. **Resultados.** Luego de aplicar el modelo de predicción los resultados mayor frecuencia son los incrementos de contagios de Covid-19 con un incremento del (1%) por cada día transcurrido. **Conclusión.** Se concluyó que los casos seguirán incrementando con el pasar del tiempo ya que la mayoría de la población no toma las precauciones respectivas e irrespeta el distanciamiento social.

[1]Universidad Católica de Cuenca, Facultad de Ingeniería de Sistemas. Azogues, Ecuador. japaguayh26@est.ucacue.edu.ec

[2]Universidad Católica de Cuenca, Facultad de Ingeniería de Sistemas. Azogues, Ecuador. juanandres1435@gmail.com

Palabras Clave: Data Mining, Machine Learning, Serie de Tiempo, Weka.

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

I. INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) declaró la enfermedad por coronavirus (Covid-19) una emergencia de salud pública de importancia internacional. (Organización Mundial de la Salud, 2020). Desde su inicio hasta el 25 de febrero del 2020, se documentó un total de 81.109 casos confirmados por laboratorios de todo el mundo (Wei-Jie, y otros, 2020).Recientes investigaciones realizadas a 425 casos confirmados demuestran que el Covid-19 es capaz de duplicar el número de personas afectadas cada 7 días, los mismos pueden propagar la infección a otras personas en un promedio de 2,2 (Jasper Fuk-Woo Chan*, 2020).

El Servicio Nacional de Gestión de Riesgos y Emergencias es el encargado de publicar información sobre los casos de Covid-19 en el país.

Para el análisis de datos se pretende desarrollar un prototipo mediante el software weka, dentro de este software se aplicarán técnicas para realizar el modelo predictivo el cual arrojara como resultado datos, los mismos son de gran ayuda para tomar medidas y con ello reducir el creciente incremento de casos de Covid-19.

Para el análisis de datos estos se presentarán dentro de gráficas para demostrar el resultado obtenido por el modelo predictivo y a su vez hacer el análisis comparativo pertinente entre los datos reales y los datos que son arrojados por el modelo. De este modo se tiene un producto final eficiente el cual ayudara a la toma de decisiones en base a los resultados presentados.

II. Marco Teórico

A. Reseña Histórica

El papel principal que desarrolla la inteligencia artificial es el tratamiento y análisis de datos.

En ocasiones, se desarrollan dos fases dentro de la IA; la primera fase es la fase de aprendizaje y una segunda es la fase de predicción.

En la primera fase (fase de aprendizaje) se ingresan los datos más representativos de ciertas situaciones que van a ser analizadas, de esta forma el sistema IA aprende las características más relevantes de los datos analizados de este modo es capaz de generalizar su estructura, esta estructura forma un modelo de datos mediante los cuales se pueden realizar una predicción acertada a partir de nuevas características. (Inteligencia Artificial Avanzada, 2014).

En el área de ingeniería la IA se utiliza para:

- La organización de la producción
- La optimización de procesos
- El cálculo de estructuras
- La planificación y logística
- El diagnóstico de fallos
- La toma de decisiones (Inteligencia Artificial Avanzada, 2014).

B. Weka

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

Es un software de aprendizaje automático de código abierto, el mismo nos permite trabajar por medio de una interfaz gráfica o mediante las aplicaciones de terminal estándar esto es posible a través de una API de Java. Este software tiene un sin número de herramientas integradas para realizar tareas estándar de aprendizaje automático.

Weka es actualmente una de las plataformas para la minería de datos más populares y cuenta con un paquete dedicado específicamente a la predicción de series temporales mediante técnicas de regresión (Time Series Analysis and Forecasting with Weka - Pentaho Data Mining - Pentaho Wiki, s.f.).

C. Serie de Tiempo

Una serie temporal se define como una secuencia de □ observaciones o datos □ □ ordenadas cronológicamente, sobre una característica (serie univariable) o sobre varias características (serie multivariable) de una unidad observable, tomadas en diferentes momentos.

Las series temporales se caracterizan fundamentalmente por la gran numerosidad de los datos que la conforman, la alta dimensionalidad y la necesidad de su constante actualización (Tak-chung, 2011).

Las series temporales se estudian principalmente con el objetivo de extraer información de algún fenómeno del pasado e intentar predecir el futuro, lo cual permite descubrir características en los datos y determinar su variación a largo plazo (Md & Alam, 2012).

D. Machine Learning

Es el aprendizaje automático que consiste en programar computadoras para optimizar un criterio utilizando datos de ejemplo o experiencia pasada. (Ethem, 2010)

E. Data Mining

La minería de datos consiste en descubrir nuevas correlaciones significativas, modelos y tendencias, filtrando grandes cantidades de datos almacenados en repositorios digitales, a través del uso de patrones de reconocimiento de modelos, así como de técnicas estadísticas y matemáticas (DANIEL & CHANTAL).

F. Base de Datos

Una base de datos consiste en una colección de datos almacenados dentro de un repositorio.

El software de base de datos proporciona mecanismos para definir la estructura que debe tener la misma y cómo debe ser el almacenamiento de datos; el software lo realiza para especificar y gestionar concurrentes, compartidos, o acceso a datos distribuidos; de este modo no solo se garantiza la coherencia de la información que se almacena dentro de la base de datos si no también seguridad de la información que se maneja (Han, Micheline, & Pei).

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

G. Data Set

Una data set son conjuntos de datos se componen de objetos de datos, los objetos de datos representan una entidad.

Los objetos de datos se describen por atributos, y estos atributos a su vez pueden ser nominales, binarios, ordinales o numéricos (Han, Micheline, & Pei).

H. Tipos de Datos

Los valores de tipo nominal (o categórico) son símbolos o nombres de cosas, donde cada valor representar una categoría, un código o estado (Han, Micheline, & Pei).

Los atributos binarios son atributos nominales con solo dos estados posibles (como 1 y 0 o verdadero y falso). Si los dos estados son igualmente importantes, el atributo es simétrico (Han, Micheline, & Pei).

Un atributo ordinal tiene como posibles valores un orden significativo o clasificarse entre ellos, pero se desconoce la magnitud entre valores sucesivos (Han, Micheline, & Pei).

Un atributo numérico es cuantitativo no es una cantidad medible, la misma se representa con valores enteros o reales (Han, Micheline, & Pei).

I. Modelo de Predicción

Es un modelo se utiliza para predecir una variable de clase de objetos para el o los valores que se desconoce (Han, Micheline, & Pei).

J. Holt-Winters

Holt-Winters es una clase que implementa el método de suavizado exponencial triple de para el pronóstico de series de tiempo. Diseñado para ser utilizado en el entorno de predicción de Weka [1].

Holt-Winters considera nivel, tendencia y estacional de una determinada serie de tiempo. Este método tiene dos principales modelos, dependiendo del tipo de estacionalidad: (Nancy, 2020)

El modelo multiplicativo estacional: Este modelo presupone que a medida que se incrementan los datos, también se incrementa el patrón estacional, la mayoría de las gráficas que se presentan mediante el uso de series de tiempo muestran este patrón. (Nancy, 2020).

El modelo aditivo estacional:

Es un modelo de datos en el que los efectos de los factores individuales se diferencian y se agrupan para modelar los datos. Un modelo aditivo es opcional para los procedimientos de descomposición y para el método de Winters (Nancy, 2020).

Existen tres fases de trabajo, las mismas trabajan con tres conjuntos de datos diferentes

1. El primer grupo de datos es para inicializar el modelo, dónde debemos determinar los indicadores de nivel, tendencia y estacionalidad (Nancy, 2020).
2. Es necesario un segundo conjunto de datos probar los índices de suavización Alfa, Beta y Gamma (Nancy, 2020).

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

3. Con el tercer grupo de datos para se realiza el pronóstico, evaluación y el funcionamiento del modelo propuesto. es la siguiente:

La fórmula que se utiliza para el pronóstico es la siguiente: D t, t+1 = (at +T.bt) + F t +T-P Dónde: **D** = Es la variable a estimar o pronosticar; **a** = Nivel promedio de casos; **b** = Tendencia; **F** = Factor de estacionalidad; **t** = Período actual; **T** = Número de períodos que se desean avanzar (Nancy, 2020).

III. Metodología

La metodología utilizada para realizar este trabajo será la metodología de piloto experimental ya que el objetivo que se tiene es evaluar la efectividad del algoritmo y comprobar una predicción en base a información obtenida y almacenada de manera previa.

Los pasos a seguir para el desarrollo de esta investigación son los siguientes:

A. ***Obtener la información proveniente del total de contagiados dentro de las diferentes provincias del Ecuador:*** La información será obtenida de la página oficial del Registro nacional de Gestión de Riesgos y Emergencias para ser preparada y utilizada en el modelo de entrenamiento.

B. ***Preparar datos para el análisis:*** Se seleccionarán los atributos y características más relevantes, los cuales serán analizados en búsqueda del resultado esperado.

C. ***Selección de algoritmo:*** Se seleccionará un algoritmo en este caso se utilizará Holt-Winters el cual será entrenado con datos provenientes de la data set es decir de los casos confirmados dentro de las provincias del Ecuador.

D. ***Comparación y Resultados:*** Una vez obtenidos los resultados podremos medir la efectividad del algoritmo escogido.

IV. Desarrollo

Al empezar a recolectar los datos desde que se publicó por primera vez el informe situacional en cuanto a los contagiados de coronavirus (Covid-19) en el ecuador, desde ese momento se empezó a recolectar estos datos para posteriormente ser limpiados y almacenados en data set; en este caso el data set fue realizado en una hoja de cálculo de Excel, el tipo de dato utilizado es de tipo numérico.

Existen datos que se dejaron de publicar dentro de los informes situacionales tal es el caso de los datos que se almacenan dentro de la variable denominada “Posibles Casos”, estos datos fueron presentados dentro de los informes hasta el día cinco del mes de abril del año en curso.

Al tener alimentado el data set con la información a utilizar procedemos a ingresarla dentro del modelo de predicción, para realizar una comparativa posterior entre la información de la página web del Servicio Nacional de Gestión de Riesgos y los datos arrojados por el modelo.

El modelo está basado en el análisis de series en el tiempo ya que utiliza toda la información de la data set para entrenar y presentar la predicción para los días posteriores que se indiquen.

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

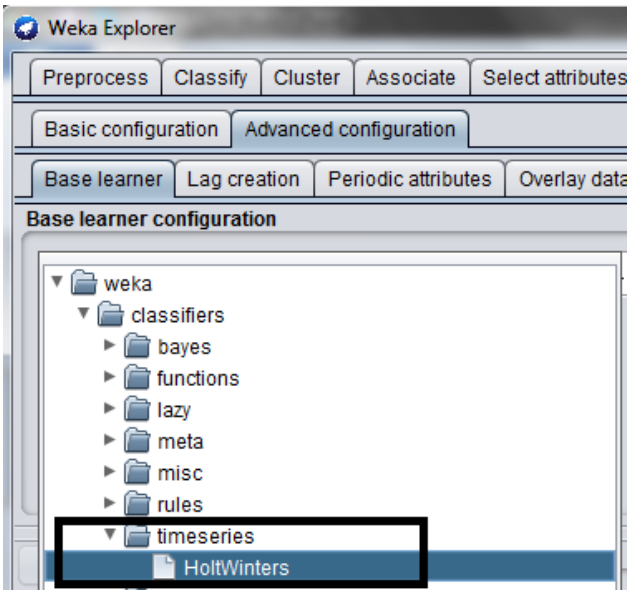


Fig. 1 Selección del Algoritmo
Fuente: Elaboración propia.

Como se observa en la Fig.1, nos presenta el algoritmo el cual va analizar los datos en este caso se analiza con HoltWinters el cual está directamente vinculado con las series de tiempo.

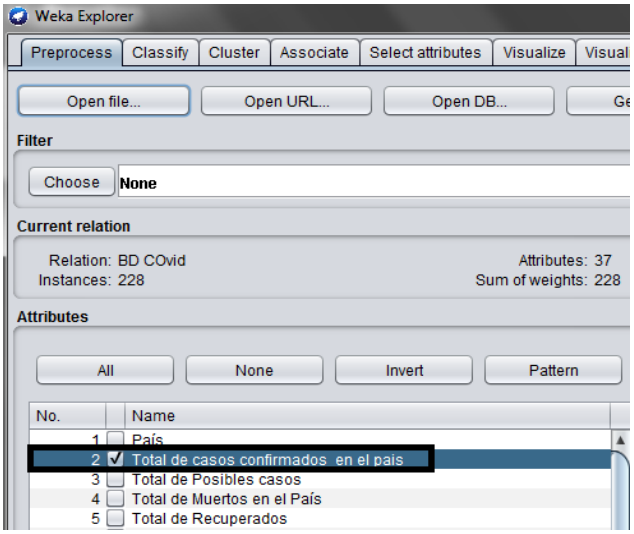


Fig. 2 Selección de la variable a predecir
Fuente: Elaboración propia.

En la Fig.2 se observa la variable que se va a predecir, este caso la variable analizada es el total de casos de coronavirus (Covid-19) que existen en el país, pero el modelo se puede emplear para predecir cada una de las ciudades dentro del país, en el apartado de resultados se presentara las predicciones para ciertas ciudades dentro del país.

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

V. Resultados

Al modelo se ingresaron datos desde el 23 de marzo del 2020 hasta el día 15 de diciembre de 2020 para predecir los datos de los días posteriores.

Cabe recalcar que las predicciones que se realizan tienen un 80% de efectividad.

Los resultados obtenidos son los siguientes:

Num Instancia	Valores Iniciales	Valor Predicción	Error
272	199228	210574	113466
273	200379	213150	127716
274	200765	215243	144785
275	201524	217341	158178
276	202120	218656	165367
277	202180	204723	254375
278	202356	204476	212094

Tabla 1 Resultados de la predicción
Fuente: Elaboración propia.

En la tabla anterior se presentan un extracto de los resultados de la predicción obtenida; con la totalidad de los datos procederemos a realizar las gráficas respectivas en donde se visualizará y se comprenderá de mejor manera el trabajo realizado.

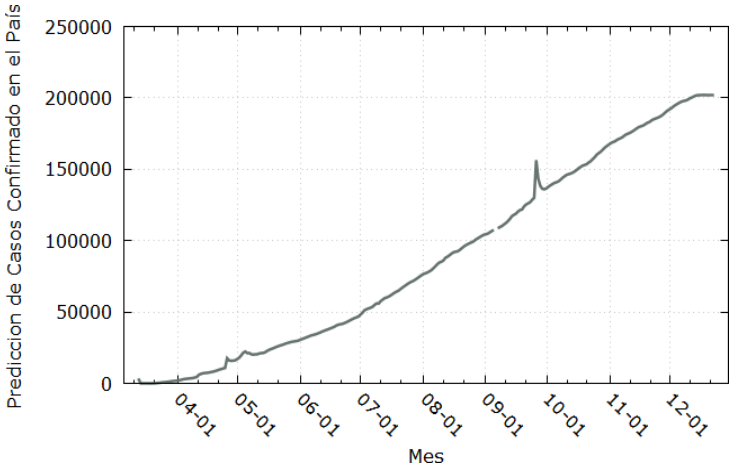


Fig. 3 Resultados de la predicción
Fuente: Elaboración propia.

En la figura anterior se muestra: El número de instancia que es el número de días que han transcurrido desde que se reportó el primer caso, los valores iniciales que son los valores que se ingresaron de manera diaria dentro del data set; estos valores diarios dejaron de ser ingresados para posteriormente realizar una comparativa entre valores iniciales y valores de predicción que arroja el modelo.

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

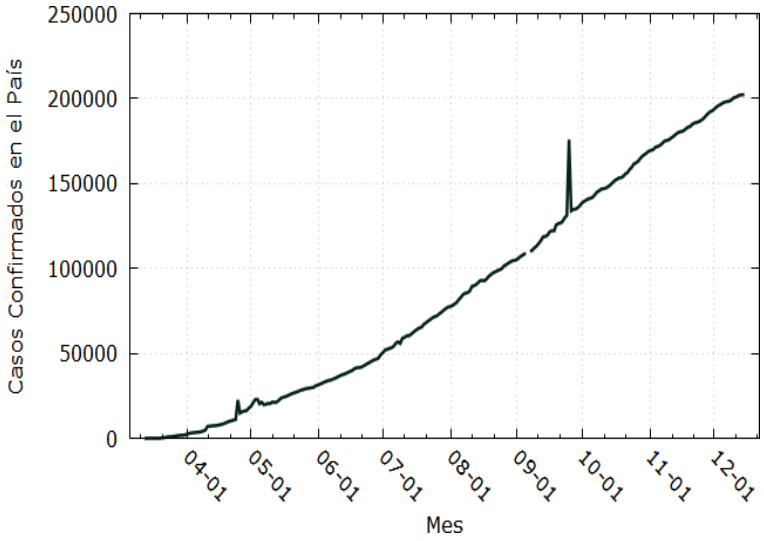


Fig. 4 Resultados del total de casos confirmados
Fuente: Elaboración propia.

En la figura anterior se muestra el total de los casos confirmados dentro del país está separado en semanas para una mayor comprensión.

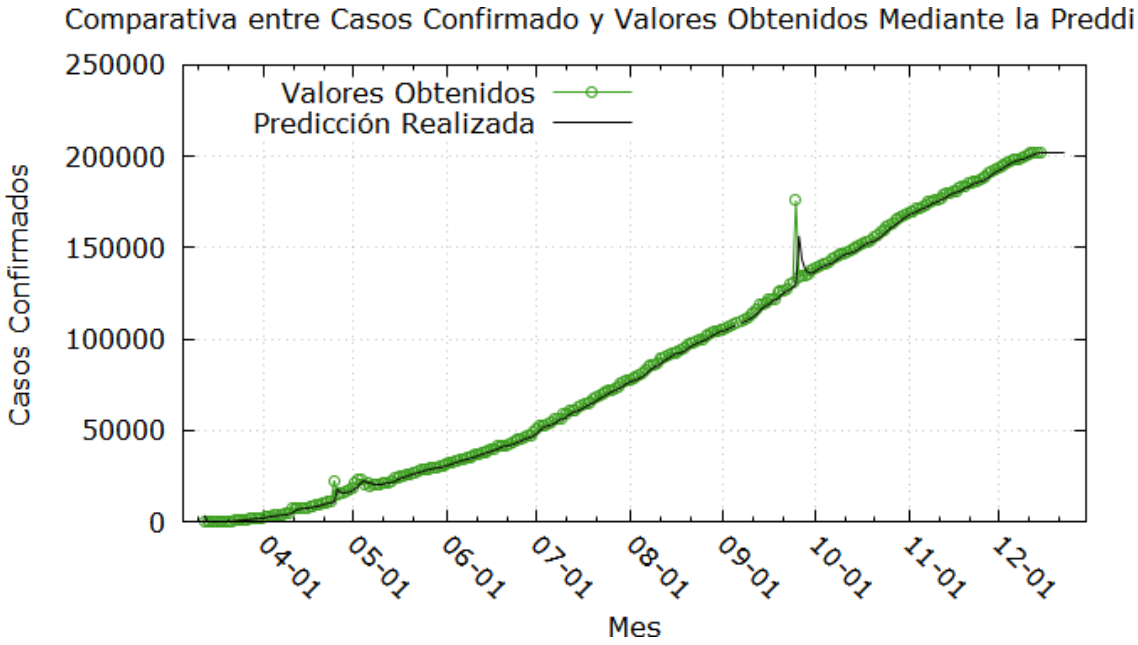


Fig. 5 Comparativa de resultados
Fuente: Elaboración propia.

En la figura se puede observar la gráfica entre los datos analizados y la predicción, las gráficas están interceptadas en ciertos puntos, pero los valores están relativamente cerca entre el valor obtenido y el valor predicho por el prototipo.

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

VI. Conclusión

El modelo utilizado resulta muy útil al momento de realizar predicciones, las mismas no están tan alejadas de la realidad ya que están basadas en datos históricos los cuales son analizados por una serie de tiempo hasta entregar un valor aproximado; con este valor aproximado se puede tomar una decisión temporal o definitiva.

Analizando los datos obtenidos en base a los resultados y comparando con la actualidad del país, resulta evidente la eficacia de este modelo predictivo, puesto que, si bien existe variación, es mínima y está dentro de los rangos admitidos como porcentaje de error.

El uso de la herramienta Weka resulta de gran apoyo al momento de realizar este tipo de modelos, debido a que nos permite automatizar ciertos procesos necesarios para los mismos.

VII Discusión

Si bien el contar con datos históricos de un intervalo de tiempo considerable (años, por ejemplo) es de gran apoyo al momento de realizar modelos predictivos, en este caso en particular con los datos recopilados con el paso de los meses ayudan para poder ofrecer un modelo lo suficientemente efectivo.

Se recomienda para trabajos a futuro sobre modelos predictivos el uso de Python, puesto que Weka es una herramienta que poco a poco está quedando en el pasado.

Referencias

(2014). En B. Raúl, E. Gerard, K. Samir, & M. RodóDavid, Inteligencia Artificial Avanzada (pág. 298). Barcelona, España: Universitat Oberta de Catalunya.

Antonio, E. M. (Noviembre de 2018). WEKA, ÁREAS DE APLICACIÓN Y SUS ALGORITMOS: UNA REVISIÓN SISTEMÁTICA. ECOCIENCIA. Obtenido de media.proquest.com/media/hms/PFT/1/xoCQ9?_s=UII2KLRQanh4Il7%2Fe49wkkQ%2B7c4%3D

DANIEL, T. L., & CHANTAL, D. L. (s.f.). En DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining (Segunda ed.). New Jersey, Esatados Unidos de Norte América. Obtenido de doc.lagout.org/Others/Data%20Mining/Discovering%20Knowledge%20in%20Data_%20An%20Introduction%20to%20Data%20Mining%20%282nd%20ed.%29%20%5BLarose%20%26%20Larose%202014-06-30%5D.pdf

Ethem, A. (2010). Introduction to Machine Learning (Segunda ed.). Londres, Inglaterra. Obtenido de kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf

ISSN: 2602-8085
Vol. 4, N°1, p. 78-97, enero-marzo, 2021

Han, J., Micheline, K., & Pei, J. (s.f.). Data Mining Concepts and Techniques (Tercera ed.). Obtenido de myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

HoltWinters. (2020). Recuperado el 25 de Febrero de 2021, de timeseriesForecasting 1.1.27 API: https://weka.sourceforge.io/doc.packages/timeseriesForecasting/weka/classifiers/timeseries/HoltWinters.html. [Accessed: 22-Dec-2020].

Jasper Fuk-Woo Chan*, S. Y.-H.-W. (January de 2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmition: a study of a family cluster. The Lancet, 395(10223), 514-523.

Md, S., & Alam, K. (2012). Cointegration and causal relationships between energy consumption and output: Assessing the evidence from Australia. Energy Economics, 34, 2182-2188. doi:10.1016 / j.eneco.2012.03.006

Nancy, V. R. (2020). Recuperado el Febrero de 2021, de “RPubs - Holt-Winters.”: https://rpubs.com/nanrosvil/283121

Organización Mundial de la Salud. (Noviembre de 2020). PAndemia de Enfermedad por coronavirus(Covid-19). OMS. Obtenido de www.who.int/emergencies/diseases/novel-coronavirus-2019

Tak-chung, F. (Septiembre de 2011). A review on time series data mining. ELSEIVER(24), 164-181.

Time Series Analysis and Forecasting with Weka - Pentaho Data Mining - Pentaho Wiki. (s.f.). Recuperado el 20 de Noviembre de 2020, de Wiki.pentaho.com: wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka

Wei-Jie, G., Zheng-Yi, N., Wen-Hua, L., Chun-quan, O., Shan, H., & Chun-liang, L. (Abril de 2020). Clinical Charasteristics of Coronavirus Disease 2019 in China. The New England Journal of Medicine, 382, 1708-1720. doi:10.1056/NEJMoa2002032

Juan Andrés Paguay Hurtado portador(a) de la cédula de ciudadanía Nº **0302165626**. En calidad de autor/a y titular de los derechos patrimoniales del trabajo de titulación **"Desarrollo de un prototipo para la predicción de nuevos casos de covid-19 en el ecuador mediante el uso de inteligencia artificial"** de conformidad a lo establecido en el artículo 114 Código Orgánico de la Economía Social de los Conocimientos, Creatividad e Innovación, reconozco a favor de la Universidad Católica de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos y no comerciales. Autorizo además a la Universidad Católica de Cuenca, para que realice la publicación de éste trabajo de titulación en el Repositorio Institucional de conformidad a lo dispuesto en el artículo 144 de la Ley Orgánica de Educación Superior.

Cuenca, **02 julio de 2021**

F: ……

Juan Andrés Paguay Hurtado

C.I. **0302165626**

Cuenca, 26 de julio de 2021

Asunto: Embargo Temporal del Trabajo de Titulación

Señor,
Dr. Leopoldo Pauta Ayabaca, PhD.,
Decano de la Unidad Académica de Tecnologías de la Información y Comunicación,
Cuenca.

De mi consideración:

Señor Decano, yo Juan Andrés Paguay Hurtado como autor del Trabajo de Titulación **"Desarrollo de un prototipo para la predicción de nuevos casos de covid-19 en el ecuador mediante el uso de inteligencia artificial"** y el Ing. Sebastián Quevedo como director de la misma, solicitamos a usted y por su digno intermedio a Biblioteca y al responsable del repositorio institucional, el EMBARGO TEMPORAL del mismo, por un lapso de seis meses, con la finalidad de evaluar su contenido con fines de: protección por propiedad intelectual; evaluación de artículo científico para publicación en revista Ciencia Digital, indexada y con catálogo 2.0; contener información reservada que será utilizada con fines de investigación. Entiendo que luego de vencido este período automáticamente la obra será puesta a disposición del público bajo las normas de gestión de la Universidad.

Por la atención que sepa dar al presente, nos suscribimos de usted muy agradecidos.

Atentamente
,

CI: 0302165626
Autor. Juan Andrés Paguay Hurtado

C.C.: Biblioteca.

Curiginal

Document Information

Analyzed document	07_Juan Andres Paguay_U_CatoCuenca.docx (D107418272)
Submitted	6/1/2021 8:24:00 PM
Submitted by	Efraín
Submitter email	luisefrainvelastegui@cienciadigital.org
Similarity	0%
Analysis address	efrainvelastegui.cde@analysis.urkund.com

Sources included in the report

Entire Document

Desarrollo de un prototipo para la predicción de nuevos casos de COVID-19 en el ecuador mediante el uso de inteligencia artificial.

Development of a prototype for the prediction of new cases of COVID-19 in Ecuador through the use of artificial intelligence.

Juan Andrés Paguay Hurtado1*, Universidad Católica de Cuenca1 *japaguayh26@est.ucacue.edu.ec juanandres1435@gmail.com

DOI: https://doi.org/10.26871/killkana_tecnica.v4i1.586 Resumen

Los casos de coronavirus (Covid-19) en el mundo entero, van cada vez en aumento. La incertidumbre de una cifra cercana a la realidad, genera angustia en la población. Razón por la cual, me he visto en la necesidad de, plantear el uso de la Inteligencia Artificial (IA). Mediante la cual, se tendrá una información aproximada de los casos de coronavirus. Ayudando a tener informada a toda la población sobre la propagación de este virus.

Para ello, se tomó como fuente de datos la información publicada diariamente en la página oficial del Servicio Nacional de Gestión de Riesgos y Emergencias. Utilizando como apoyo los modelos predictivos, se almacenaron estos datos en un data set, para luego ser consolidados y posteriormente introducirlos en un algoritmo, el cual utilizando series de tiempo realizará las predicciones en base a datos históricos mediante el software weka.

El siguiente artículo, presenta un modelo capaz de predecir la cifra cercana a la realidad de casos de coronavirus, consiguiendo un 80% de efectivad. Por lo que se puede manifestar que este modelo resulta muy útil para realizar predicciones dentro de un periodo determinado.

Por lo que, con esta investigación se prende ser una herramienta para el desarrollo de futuros modelos de proyección que utilicen otro software, con el fin de probar el nivel de eficacia comparando los parámetros de estudio.

Palabras Clave: Data Mining, Machine Learning, Serie de Tiempo, Weka. Abstract

Cases of coronavirus (Covid-19) around the world are increasing. The uncertainty of a figure close to reality generates anguish in the population. Which is why, I have seen the need to propose the use of Artificial Intelligence (AI). Through which, you will have approximate information on coronavirus cases. Helping to keep the entire population informed about the spread of this virus.

For this, the information published daily on the official website of the National Risk and Emergency Management Service was taken as a database. Using predictive models as support, these data were stored in a data set, to later be consolidated and later entered into an algorithm, which using time series will make predictions based on historical data using the weka software.

The following article presents a model capable of predicting the close-to-reality number of coronavirus cases, achieving 80% effectiveness. So, it can be stated that this model is very useful for making predictions within a given period.

Therefore, with this research it turns out to be a tool for the development of future projection models that use other software, in order to test the level of efficiency by comparing the study parameters.

Keywords: Data Mining, Machine Learning, Time Series, Weka.

I. INTRODUCCIÓN La Organización Mundial de la Salud (OMS) declaró la enfermedad por coronavirus (Covid-19) una emergencia de salud pública de importancia internacional.CITATION Org20 \l 12298 [1]. Desde su inicio hasta el 25 de febrero del 2020, se documentó un total de 81.109 casos confirmados por laboratorios de todo el mundo CITATION Wei20 \l 12298 [2].Recientes investigaciones realizadas a 425 casos confirmados demuestran que el Covid-19 es capaz de duplicar el número de personas afectadas cada 7 días, los mismos pueden propagar la infección a otras personas en un promedio de 2,2 CITATION Jas20 \l 12298 [3]. El Servicio Nacional de Gestión de Riesgos y Emergencias es el encargado de publicar información sobre los casos de Covid-19 en el país. Para el análisis de datos se pretende desarrollar un prototipo mediante el software weka, dentro de este software se aplicarán técnicas para realizar el modelo predictivo el cual arrojara como resultado datos, los mismos son de gran ayuda para tomar medidas y con ello reducir el creciente incremento de casos de Covid-19. Para el análisis de datos estos se presentarán dentro de gráficas para demostrar el

resultado obtenido por el modelo predictivo y a su vez hacer el análisis comparativo pertinente entre los datos reales y los datos que son arrojados por el modelo. De este modo se tiene un producto final eficiente el cual ayudara a la toma de decisiones en base a los resultados presentados.

II. Marco Teórico A. Reseña Histórica El papel principal que desarrolla la inteligencia artificial es el tratamiento y análisis de datos. En ocasiones, se desarrollan dos fases dentro de la IA; la primera fase es la fase de aprendizaje y una segunda es la fase de predicción. En la primera fase (fase de aprendizaje) se ingresan los datos más representativos de ciertas situaciones que van a ser analizadas, de esta forma el sistema IA aprende las características más relevantes de los datos analizados de este modo es capaz de generalizar su estructura, esta estructura forma un modelo de datos mediante los cuales se pueden realizar una predicción acertada a partir de nuevas características. CITATION Ben14 \l 12298 [4]. En el área de ingeniería la IA se utiliza para: • La organización de la producción • La optimización de procesos • El cálculo de estructuras • La planificación y logística • El diagnóstico de fallos • La toma de decisiones CITATION Ben14 \l 12298 [4].

B. Weka Es un software de aprendizaje automático de código abierto, el mismo nos permite trabajar por medio de una interfaz gráfica o mediante las aplicaciones de terminal estándar esto es posible a través de una API de Java. Este software tiene un sin número de herramientas integradas para realizar tareas estándar de aprendizaje automático. Weka es actualmente

una de las plataformas para la minería de datos más populares y cuenta con un paquete dedicado específicamente a la predicción de series temporales

mediante técnicas de regresión CITATION Tim20 \l 12298 [5].

C. Serie de Tiempo Una serie temporal se define como una secuencia de observaciones o datos ordenadas cronológicamente, sobre una característica (serie univariable) o sobre varias características (serie multivariable) de una unidad observable, tomadas en diferentes momentos. Las series temporales se caracterizan fundamentalmente por la gran numerosidad de los datos que la conforman, la alta dimensionalidad y la necesidad de su constante actualización CITATION FuT11 \l 12298 [6]. Las series temporales se estudian principalmente con el objetivo de extraer información de algún fenómeno del pasado e intentar predecir el futuro, lo cual permite descubrir características en los datos y determinar su variación a largo plazo CITATION MdS12 \l 12298 [7].

D. Machine Learning Es el aprendizaje automático que consiste en programar computadoras para optimizar un criterio utilizando datos de ejemplo o experiencia pasada. CITATION Eth10 \l 12298 [8]

E. Data Mining

La minería de datos consiste en descubrir nuevas correlaciones significativas, modelos y tendencias, filtrando grandes cantidades de datos almacenados en repositorios digitales, a través del uso de patrones de reconocimiento de modelos, así como de técnicas estadísticas y matemáticas

CITATION DAN \l 12298 [9].

F. Base de Datos Una base de datos consiste en una colección de datos almacenados dentro de un repositorio. El software de base de datos proporciona mecanismos para definir la estructura que debe tener la misma y cómo debe ser el almacenamiento de datos; el software lo realiza para especificar y gestionar concurrentes, compartidos, o acceso a datos distribuidos; de este modo no solo se garantiza la coherencia de la información que se almacena dentro de la base de datos si no también seguridad de la información que se maneja CITATION Jia \l 12298 [10].

G. Data Set Una data set son conjuntos de datos se componen de objetos de datos, los objetos de datos representan una entidad. Los objetos de datos se describen por atributos, y estos atributos a su vez pueden ser nominales, binarios, ordinales o numéricos CITATION Jia \l 12298 [10].

H. Tipos de Datos Los valores de tipo nominal (o categórico) son símbolos o nombres de cosas, donde cada valor representar una categoría, un código o estado CITATION Jia \l 12298 [10]. Los atributos binarios son atributos nominales con solo dos estados posibles (como 1 y 0 o verdadero y falso). Si los dos estados son igualmente importantes, el atributo es simétrico CITATION Jia \l 12298 [10]. Un atributo ordinal tiene como posibles valores un orden significativo o clasificarse entre ellos, pero se desconoce la magnitud entre valores sucesivos CITATION Jia \l 12298 [10]. Un atributo numérico es cuantitativo no es una cantidad medible, la misma se representa con valores enteros o reales CITATION Jia \l 12298 [10].

I. Modelo de Predicción Es un modelo se utiliza para predecir una variable de clase de objetos para el o los valores que se desconoce CITATION Jia \l 12298 [10].

J. Holt-Winters Holt-Winters es una clase que implementa el método de suavizado exponencial triple de para el pronóstico de series de tiempo. Diseñado para ser utilizado en el entorno de predicción de Weka [1]. Holt-Winters considera nivel, tendencia y estacional de una determinada serie de tiempo. Este método tiene dos principales modelos, dependiendo del tipo de estacionalidad: CITATION VRN20 \l 12298 [11] El modelo multiplicativo estacional: Este modelo presupone que a medida que se incrementan los datos, también se incrementa el patrón estacional, la mayoría de las gráficas que se presentan mediante el uso de series de tiempo muestran este patrón. CITATION VRN20 \l 12298 [11]. El modelo aditivo estacional: Es un modelo de datos en el que los efectos de los factores individuales se diferencian y se agrupan para modelar los datos. Un modelo aditivo es opcional para los procedimientos de descomposición y para el método de Winters CITATION VRN20 \l 12298 [11]. Existen tres fases de trabajo, las mismas trabajan con tres conjuntos de datos diferentes 1. El primer grupo de datos es para inicializar el modelo, dónde debemos determinar los indicadores de nivel, tendencia y estacionalidad CITATION VRN20 \l 12298 [11]. 2. Es necesario un segundo conjunto de datos probar los índices de suavización Alfa, Beta y Gamma CITATION VRN20 \l 12298 [11]. 3. Con el tercer grupo de datos para se realiza el pronóstico, evaluación y el funcionamiento del modelo propuesto. es la siguiente: La fórmula que se utiliza para el pronóstico es la siguiente: D t, t+1 = (at +T.bt) + F t +T-P Dónde: D = Es la variable a estimar o pronosticar; a = Nivel promedio de casos; b = Tendencia; F = Factor de estacionalidad; t = Período actual; T = Número de períodos que se desean avanzar CITATION VRN20 \l 12298 [11].

III. Metodología La metodología utilizada para realizar este trabajo será la metodología de piloto experimental ya que el objetivo que se tiene es evaluar la efectividad del algoritmo y comprobar una predicción en base a información obtenida y almacenada de manera previa. Los pasos a seguir para el desarrollo de esta investigación son los siguientes: A. Obtener la información proveniente del total de contagiados dentro de las diferentes provincias del Ecuador: La información será obtenida de la página oficial del Registro nacional de Gestión de Riesgos y Emergencias para ser preparada y utilizada en el modelo de entrenamiento. B. Preparar datos para el análisis: Se seleccionarán los atributos y características más relevantes, los cuales serán analizados en búsqueda del resultado esperado. C. Selección de algoritmo: Se seleccionará un algoritmo en este caso se utilizará Holt-Winters el cual será entrenado con datos provenientes de la data set es decir de los casos confirmados dentro de las provincias del Ecuador. D. Comparación y Resultados: Una vez obtenidos los resultados podremos medir la efectividad del algoritmo escogido.

IV. Desarrollo Al empezar a recolectar los datos desde que se publicó por primera vez el informe situacional en cuanto a los contagiados de coronavirus (Covid-19) en el ecuador, desde ese momento se empezó a recolectar estos datos para posteriormente ser limpiados y almacenados en data set; en este caso el data set fue realizado en una hoja de cálculo de Excel, el tipo de dato utilizado es de tipo numérico. Existen datos que se dejaron de publicar dentro de los informes situacionales tal es el caso de los datos que se almacenan dentro de la variable denominada “Posibles Casos”, estos datos fueron presentados dentro de los informes hasta el día cinco del mes de abril del año en curso. Al tener alimentado el data set con la información a utilizar procedemos a ingresarla dentro del modelo de predicción, para realizar una comparativa posterior entre la información de la página web del Servicio Nacional de Gestión de Riesgos y los datos arrojados por el modelo. El modelo está basado en el análisis de series en el tiempo ya que utiliza toda la información de la data set para entrenar y presentar la predicción para los días posteriores que se indiquen.

Fig. 11 Selección del Algoritmo

Fuente: Autor del Proyecto

Como se observa en la Fig.1, nos presenta el algoritmo el cual va analizar los datos en este caso se analiza con HoltWinters el cual está directamente vinculado con las series de tiempo.

Fig. 22 Selección de la variable a predecir

Fuente: Autor del Proyecto

En la Fig.2 se observa la variable que se va a predecir, este caso la variable analizada es el total de casos de coronavirus (Covid-19) que existen en el país, pero el modelo se puede emplear para predecir cada una de las ciudades dentro del país, en el apartado de resultados se presentara las predicciones para ciertas ciudades dentro del país.

V. Resultados Al modelo se ingresaron datos desde el 23 de marzo del 2020 hasta el día 15 de diciembre de 2020 para predecir los datos de los días posteriores. Cabe recalcar que las predicciones que se realizan tienen un 80% de

efectividad. Los resultados obtenidos son los siguientes:

Tabla 11 Resultados de la predicción

Fuente: Autor del Proyecto

En la tabla anterior se presentan un extracto de los resultados de la predicción obtenida; con la totalidad de los datos procederemos a realizar las gráficas respectivas en donde se visualizará y se comprenderá de mejor manera el trabajo realizado.

Fig. 3 Resultados de la predicción

Fuente: Autor del Proyecto En la figura anterior se muestra: El número de instancia que es el número de días que han transcurrido desde que se reportó el primer caso, los valores iniciales que son los valores que se ingresaron de manera diaria dentro del data set; estos valores diarios dejaron de ser ingresados para posteriormente realizar una comparativa entre valores iniciales y valores de predicción que arroja el modelo.

Fig. 4 Resultados del total de casos confirmados

Fuente: Autor del Proyecto

En la figura anterior se muestra el total de los casos confirmados dentro del país está separado en semanas para una mayor comprensión.

Fig. 5 Comparativa de resultados

Fuente: Autor del Proyecto

En la figura se puede observar la gráfica entre los datos analizados y la predicción, las gráficas están interceptadas en ciertos puntos, pero los valores están relativamente cerca entre el valor obtenido y el valor predicho por el prototipo. Dentro de los anexos se mostrarán la graficas obtenidas para la provincia del cañar.

VI. Conclusión El modelo utilizado resulta muy útil al momento de realizar predicciones, las mismas no están tan alejadas de la realidad ya que están basadas en datos históricos los cuales son analizados por una serie de tiempo hasta entregar un valor aproximado; con este valor aproximado se puede tomar una decisión temporal o definitiva. Analizando los datos obtenidos en base a los resultados y comparando con la actualidad del país, resulta evidente la eficacia de este modelo predictivo, puesto que, si bien existe variación, es mínima y está dentro de los rangos admitidos como porcentaje de error. El uso de la herramienta Weka resulta de gran apoyo al momento de realizar este tipo de modelos, debido a que nos permite automatizar ciertos procesos necesarios para los mismos.

VII Discusión Si bien el contar con datos históricos de un intervalo de tiempo considerable (años, por ejemplo) es de gran apoyo al momento de realizar modelos predictivos, en este caso en particular con los datos recopilados con el paso de los meses ayudan para poder ofrecer un modelo lo suficientemente efectivo. Se recomienda para trabajos a futuro sobre modelos predictivos el uso de Python, puesto que Weka es una herramienta que poco a poco está quedando en el pasado.

VIII Referencias

[1] Organización Mundial de la Salud, «PAndemia de Enfermedad por coronavirus(Covid-19),» OMS, Noviembre 2020.

[2]

G. Wei-Jie, N. Zheng-Yi, L. Wen-Hua, O. Chun-quan, H. Shan y L. Chun-liang, «Clinical Charasteristics of Coronavirus Disease 2019 in China,» The New England Journal of Medicine, vol. 382, pp. 1708-1720, Abril 2020.

[3]

S. Y. K.-H. K. K. K.-W. T. H. C. J. Y. Jasper Fuk-Woo Chan*, «A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmition: a study of a family cluster,» The Lancet, vol. 395, nº 10223, pp. 514-523, January 2020.

[4]

de Inteligencia Artificial Avanzada, Barcelona, Universitat Oberta de Catalunya, 2014, p. 298.

[5]

«Time Series Analysis and Forecasting with Weka - Pentaho Data Mining - Pentaho Wiki,» [En línea]. Available:

wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka. [

Último acceso: 20 Noviembre 2020].

[6]

F. Tak-chung, «A review on time series data mining,» ELSEIVER, nº 24, pp. 164-181, Septiembre 2011.

[7]

S. Md y

K. Alam, «Cointegration and causal relationships between energy consumption and output: Assessing the evidence from Australia,» Energy Economics,

vol. 34, pp. 2182-2188, 2012.

[8]

A. Ethem, Introduction to Machine Learning, Segunda ed., Londres, 2010.

[9]

T. L. DANIEL y D. L. CHANTAL, de DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining, Segunda ed., New Jersey.

[10]

J. Han, K. Micheline y J. Pei, Data Mining Concepts and Techniques, Tercera ed.

[11]

V. R. Nancy, 2020. [En línea]. Available: https://rpubs.com/nanrosvil/283121. [Último acceso: Febrero 2021].

[12]

E. M. M. Antonio, «WEKA, ÁREAS DE APLICACIÓN Y SUS ALGORITMOS: UNA REVISIÓN SISTEMÁTICA,» ECOCIENCIA, Noviembre 2018.

[13]

HoltWinters, 2020. [En línea]. Available: https://weka.sourceforge.io/doc.packages/timeseriesForecasting/weka/classifiers/timeseries/HoltWinters.html. [Accessed: 22-Dec-2020].. [Último acceso: 25 Febrero 2021].

[Metadata removed]

Curiginal

Hit and source - focused comparison, Side by Side

Submitted text	As student entered the text in the submitted document.
Matching text	As the text appears in the source.

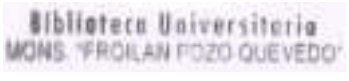
El Bibliotecario de la Sede Azogues

CERTIFICA:

Que, **Josué Ismael Siguencia Verdugo** portador de la cédula de ciudadanía N° 0302165626 de la Carrera de **Ingeniería de Sistemas,** Sede Azogues, Modalidad de estudios presencial no adeuda libros, a esta fecha.

Azogues, **27 de julio de 2021**

Eco. Fabián Rodríguez Herrera

1.- DATOS GENERALES	
Tema del trabajo de titulación:	
Unidad Académica:	
Carrera:	
Modalidad de estudio:	
Matriz, Sede o Extensión:	
2.- DESCRIPCIÓN	
2.1.- DOCUMENTOS HABILITANTES PARA LA APROBACIÓN DEL TRABAJO DE TITULACIÓN EN BIBLIOTECA	SI
Revisión del resumen y palabras clave previa traducción en el Centro de Idiomas	
Carátulas actualizadas (pasta azul y blanca)	X
Declaración de Autoría y Responsabilidad con firma (FORMATO F-DB-34)	
Certificación del tutor con firma	X
Autorización de publicación en el repositorio digital (FORMATO F-DB-30)	X
Certificado del Sistema de prevención de plagio (Turnitin)	X
Certificado de no adeudar libros a biblioteca	X
Revisión del contenido del CD (Documento Word: Carátula con fondo blanco, resumen y abstract;	
Documento pdf del trabajo de titulación, (Apellidos y nombres en cada archivo)	X
Solicitud de embargo de obra (cuando aplique)	X
Declaración de embargo de obra (cuando aplique)	
  Eco. Fabián Rodríguez Herrera BIBLIOTECARIO CARRERA DE INGENIERÍA EN SISTEMAS	
FECHA: 02/08/2021	