SATHIRÍ Sembrador

MINERÍA DE DATOS PARA SERIES TEMPORALES Y SU APLICACIÓN EN LAS PRECIPITACIONES PLUVIALES EN EL CANTÓN HUACA ECUADOR

DATA MINING FOR TIME SERIES AND ITS APPLICAT.ION IN RAINFALL PRECIPITATIONS IN THE CANTON HUACA ECUADOR AREA

Recibido: 28/10/2021 - Aceptado: 11/01/2023

Marco Antonio Yandún Velasteguí

Magíster en Auditoría de Tecnologías de la Información - Universidad de Especialidades Espíritu Santo Guayaquil

Docente de la Universidad Politécnica Estatal del Carchi

marco.yandun@unmsm.edu.pe
marco.yandun@upec.edu.ec
https://orcid.org/0000-0001-5627-9838

Carlos Fernando Montenegro Argoti

Ingeniero en Informática por la Universidad Politécnica Estatal del Carchi

carlos.montenegro@upec.edu.ec
https://orcid.org/0000-0001-8360-3162

Cómo citar este artículo:

Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

Resumen

En el cantón San Pedro de Huaca existe la Finca Experimental San Francisco, cuenta con una estación meteorológica que recolecta datos relacionados con el clima y, entre ellos, también se obtiene registros de la precipitación pluvial. Con estos datos se hace un análisis mínimo incluido a los datos histórico-pluviales, no se ha evidenciado el uso de herramientas de minería de datos, para determinar las precipitaciones o generar un pronóstico. Lo anterior se obtiene por medio de conocimientos ancestrales de agricultores. Con el uso de los datos almacenados por la estación meteorológica con sus variables y unidades de medida, se realizó la investigación que se basa en la aplicación de técnicas de minería de datos predictiva aplicando la metodología Cross Industry Standard Process for Data Mining CRISP-DM (Proceso estándar de la industria para la minería de datos), cumpliendo todas sus fases. Para la etapa de modelado se aplica los pronósticos móviles, para el pronóstico de series de tiempo, donde hace predicciones con los datos históricos, en este caso 2019 y 2020. Para llevar a cabo todo este proceso se hizo uso del lenguaje de programación R, ya que presenta grandes ventajas en el análisis de datos. Los resultados obtenidos corresponden a los pronósticos de precipitación pluvial obtenidos para cada mes del año 2021, la unidad de medida de estos datos se encuentra en milímetros por metro cuadrado. Esta información es desplegada en un aplicativo web shinnyapp.io que muestra y descarga el análisis de los datos, disponible al público.

Palabras claves: Minería de datos, análisis de datos, CRISP-DM, series de tiempo, promedios móviles

Abstract

In the San Pedro de Huaca canton there is the San Francisco Experimental Farm, it has a meteorological station that collects data related to the climate and among them also records of rainfall are obtained, with these data a minimum analysis is made including the data historical rainfall, I do not know has evidenced the use of data mining tools to determine rainfall or generate a forecast, the above is obtained through ancestral knowledge of farmers. With the use of the data stored by the meteorological station with its variables and units of measurement, the research was carried out that is based on the application of predictive data mining techniques applied the Cross Industry Standard Process for Data Mining CRISP-DM methodology, (Industry standard process for data mining), fulfilling all its phases. For the modeling stage, mobile forecasts are applied, for the forecast of time series, where it makes predictions with historical data in this case 2019 and 2020, to carry out all this process the programming language R was used since presents great advantages in data analysis, the results obtained correspond to the rainfall forecasts obtained for each month of the year 2021, the unit of measurement of these data is in millimeters per square meter, this information is displayed in a web application shinnyapp.io displaying and downloading the data analysis, publicly available.

Key words: Data mining, data analysis, CRISP-DM, time series, moving averages.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

Introducción

Los datos históricos que son almacenados por estaciones meteorológicas constituyen información importante para el análisis de datos. Uno de estos análisis que se puede realizar son las predicciones de precipitaciones pluviales en series de tiempo, las mismas que sirven para saber cuándo es buen tiempo para realizar actividades agrícolas, obteniendo un producto de gran calidad, sin perjudicar los cultivos.

León Guzmán (2017) indica que la minería de datos consiste en una gran herramienta de estrategia, cuya funcionalidad es analizar los datos históricos desde los diferentes puntos estratégicos, para así poder transformar esta información, ordenarla, clasificarla y filtrarla.

Rivera et al. (2018) mencionan que los datos que mejor se manejan deben ser generados por una estación meteorológica. Para el presente trabajo se utilizaron los datos de la finca experimental San Francisco de Huaca de la Universidad Politécnica Estatal del Carchi (UPEC), previo a la autorización de uso de los datos que ahí se generan, además que esta autorización fue realizada para Carlos Montenegro como parte de su trabajo de titulación como ingeniero en Informática bajo la dirección de Marco Yandún, tomando en cuenta que se han realizado análisis mínimos sobre el histórico de datos generados por la estación metereologíca, evidenciando que no se han aplicado herramientas de minería de datos relacionados con las precipitaciones pluviales, que sirvan de ayuda para los diferentes agricultores de la zona cuya actividad económica principal es la agronomía, permitiendo de esta forma mitigar los riesgos por la variación de clima y otros fenómenos relacionados con la hidrometereológica.

Ortiz Farro (2017) afirma que dentro del área agrícola se puede aplicar la minería de datos con la finalidad de proponer herramientas que permitan al usuario tener acceso a la información precisa donde se realicen predicciones sobre la evolución futura de actividades que se desarrollan, obteniendo resultados a corto plazo, que permitirán asegurar la confiabilidad de estos, sirviendo de apoyo para las decisiones futuras que se puedan tomar sobre las mismas. Los beneficiaros directos serán los agricultores del cantón San Pedro de Huaca, al igual que otros agricultores de diferentes zonas que cuenten con datos históricos sobre las precipitaciones, ya que la propuesta del aplicativo informático permitirá visualizar los resultados de predicciones obtenidos con sus propias variables almacenadas mediante el uso de un modelo predictivo realizado con la técnica de promedios móviles. Ilbay Yupa (2019) indica que el modelo predictivo permite la toma de mejores decisiones para obtener productos de mejor calidad, resultando con mayor demanda para el mercado logrando un mayor desarrollo económico.

Para Zamora Villalobos (2018), con los datos recolectados y el análisis de los mismos se puede desarrollar un instrumento para recolectar y analizar datos, ya que el proceso de extracción del conocimiento se la realiza a través de varias fases (preparación de datos, exploración, auditoria, minería de datos, evaluación, difusión y utilización de modelos) y la incorporación de diversas técnicas (árboles de decisión, regresión lineal, redes neuronales artificiales, máquinas de soporte vertical); además, permite una presentación variada del problema mediante su clasificación, categorización, estimación, regresión, agrupamiento etc. IONOS (2018), además de poder realizar diversas técnicas, explica que esto permite obtener los patrones, pudiendo así hacer una comparativa de cual técnica es la más viable al momento de obtener datos. El número de técnica de minería de datos es muy grande, además que, cabe indicar, no existe una técnica específica para resolver un problema determinado, como también la aplicación de diferentes metodologías de minería de datos las cuales todas llevan a un mismo propósito.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

Materiales y métodos

En la presente investigación se aplicó un enfoque cuantitativo, utilizando la recolección de datos para analizar la información, con base en la medición numérica y el análisis estadístico, para establecer patrones de comportamiento y probar teorías. Como teorías se pueden determinar los testimonios y conocimientos ancestrales de los agricultores de la zona; con ellos se realizaron diferentes encuestas, de acuerdo a Hernández Sampieri et al. (2018), "utiliza la recolección de datos sin medición numérica para descubrir o afinar preguntas de investigación en el proceso de interpretación" así como datos demográficos del cantón San Pedro de Huaca ubicado en la provincia del Carchi, cuyos datos fueron obtenidos del Instituto Nacional de Estadísticas y Censos (INEC), realizado en el 2010, en donde se pudo observar que cuenta con 6868 habitantes, de los cuales 2331 corresponden a población activa económicamente desde hace cinco años o más, eligiendo a 324 habitantes que se dedican a la agricultura como actividad ocupacional; ellos contribuyen con la información que será contrastada con la información que se genera por la estación metereologíca.

Para el proceso de minería de datos para series temporales y su aplicación en las precipitaciones pluviales se utilizaron herramientas digitales que permiten desarrollar el aplicativo informático. Mancero (2017) indica que entre una metodología de programación utilizada CRISP-DM, cumpliendo con todas las etapas, haciendo énfasis en las etapas de modelado, evaluación y despliegue ya que es ahí donde se concentra la problemática de este proyecto. Unir (2019) muestra una herramienta utilizada para la elaboración del modelado y aplicativo informático de analítica de datos es el lenguaje de programación R, ya que en el ámbito de datos permite manipular, procesar y visualizar gráficos del análisis de datos de alta calidad, además de ser un lenguaje que puede ser utilizado en todas las fases de análisis de datos. Permitiendo aplicar los promedios móviles simples, correspondiente a la expresión matemática utilizada para la obtención de los pronósticos pluviales. También se hace uso de R Studio que no es más que un entorno de desarrollo.

Resultados y discusión

Se analizó los milímetros de precipitación pluvial a través de una base de datos que contiene información diaria desde 2018-06-03 hasta 2021-08-02, procediendo a realizar una limpieza de la información al determinar la cantidad de datos faltantes o fuera de los rangos estándares. Para ello se creó una secuencia de fechas artificial que contiene los datos completos desde 2018-06-03 hasta 2021-08-02, a la cual se añadieron los datos completos de precipitación, seguido por una agregación de datos por día y no por hora como los datos sin tratar que son los originales.

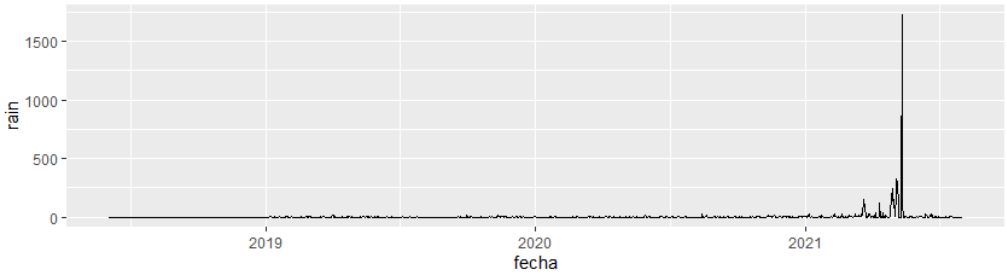

Figura 1. Serie de tiempo

Como se puede observar en la Figura 1, corresponde a la serie de tiempo donde los datos a partir del año 2021 contienen valores muy elevados, llegando hasta un valor máximo de 1725mm, lo cual es irreal en comparación con los datos históricos e incompatible con los rangos de precipitación documentados. Se procede entonces a realizar un corte de la información tomando solo los valores anteriores a 2021-01-01, quedando una serie de tiempo de agregados diarios.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

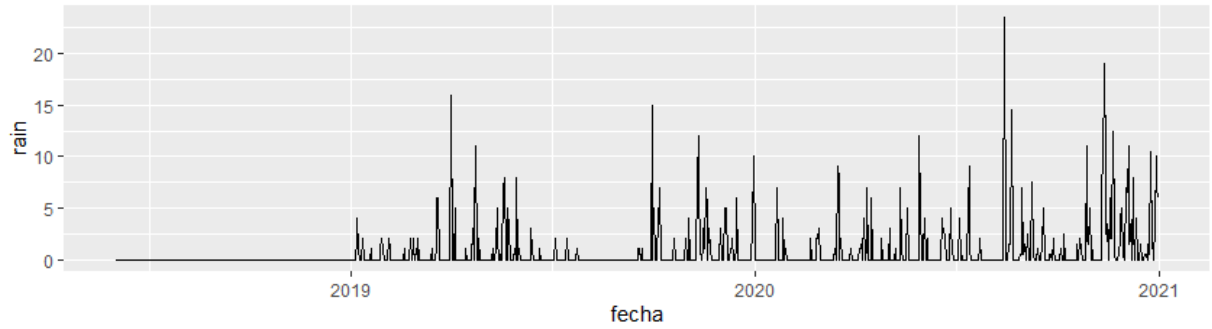

Figura 2. Serie de tiempo con corte de información

Como se puede distinguir en la Figura 2 sobre la nueva serie de tiempo, los datos del 2018 no contienen valores diferentes de 0, lo cual podría tratarse de un error de adquisición de datos o su registro, por lo cual se procedió a excluir los datos anteriores al 2019.

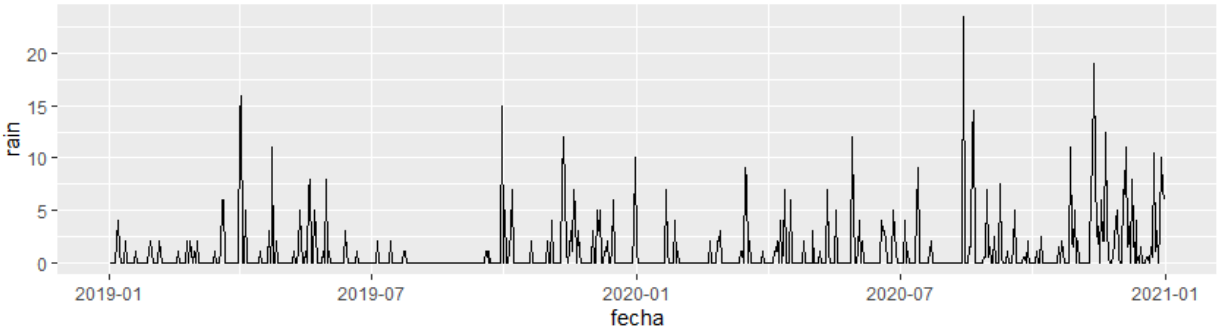

Figura 3. Serie de tiempo 2019 – 2021

Con esta delimitación de información descrita en la Figura 3, se puede observar una mejor continuidad de la información, con lo cual haremos agregados mensuales para verificar si es posible encontrar alguna estacionalidad en la serie de tiempo.

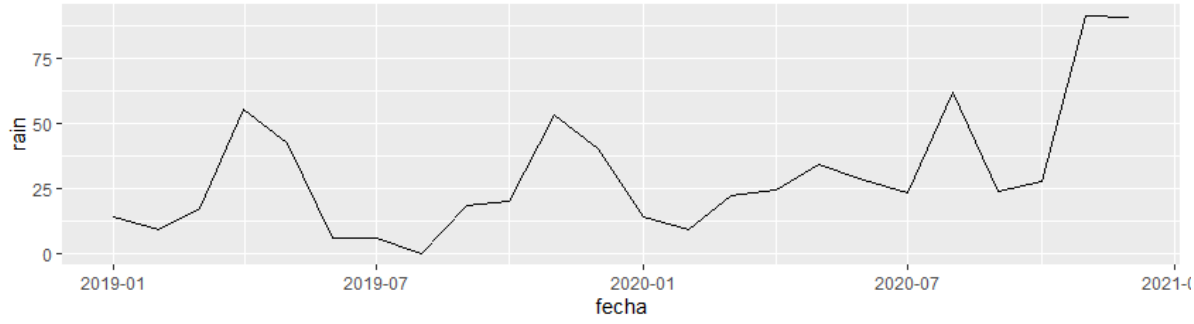

Figura 4. Estacionalidad en serie de tiempo

Dado que se tienen dos años completos de datos, se procedió a realizar una descomposición aditiva de la serie de tiempo de agregados mensuales como se puede apreciar en la Figura 4, para determinar el comportamiento general de los datos (Escutia, 2019), quedando como sigue:

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

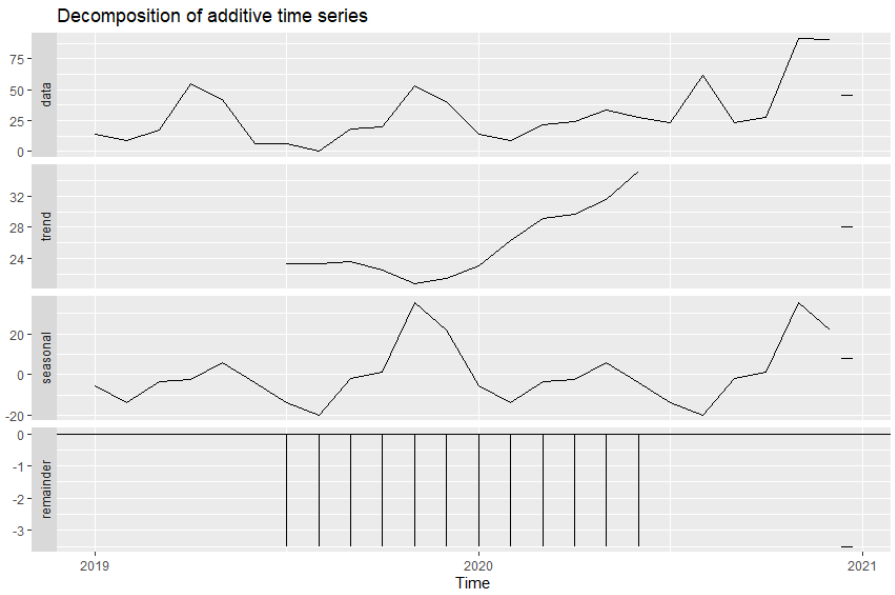

Figura 5. Descomposición aditiva de series de tiempo

Con la Gráfica 5 se puede observar como la parte tendencial (trend) tiene un crecimiento notable, lo que significaría que la precipitación pluvial tiende a aumentar con el paso de los años. Por otro lado, vemos como la parte estacional (seasonal) refleja que aproximadamente durante los segundos y los últimos trimestres de cada año se tienen las mayores cantidades de precipitación. Similar al criterio de Vazquez (2018).

Con los datos agregados mensuales procedimos a realizar una proyección de datos a 6 meses, usando un pronóstico a partir de promedios móviles, como se observa la Tabla 1. Adicional se podría generar tendencias aplicando regresión lineal con el método de los mínimos cuadrados que son base de las reglas Gaussianas y las Redes Neuronales como el estudio propuesto por Hidalgo Guijarro et al. (2019).

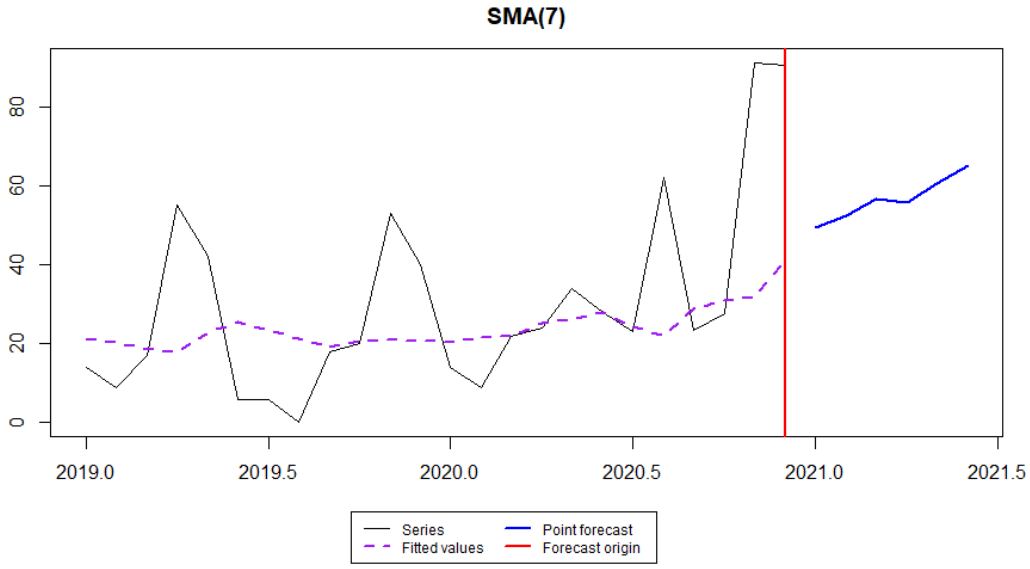

Figura 6. Histograma SMA Pronóstico

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

Tabla 1.
Forecast año 2021

Forecast	Jan	Feb	Mar	Apr	May	Jun
2021	49,35714	52,40816	56,60933	55,83923	60,45912	65,16757
	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2021	61,47722	57,33111	58,47025	59,33626	59,72583	60,28105

Para poder desplegar los datos analizados se registra de forma gratuita en la página de shinyapps.io versión server con el acceso al servidor en donde puede ser alojada, una vez creada la cuenta se dispone de un Dashboard con toda la información referente al aplicativo web para tener un manejo más cauteloso, como se puede denotar en la Figura 7.

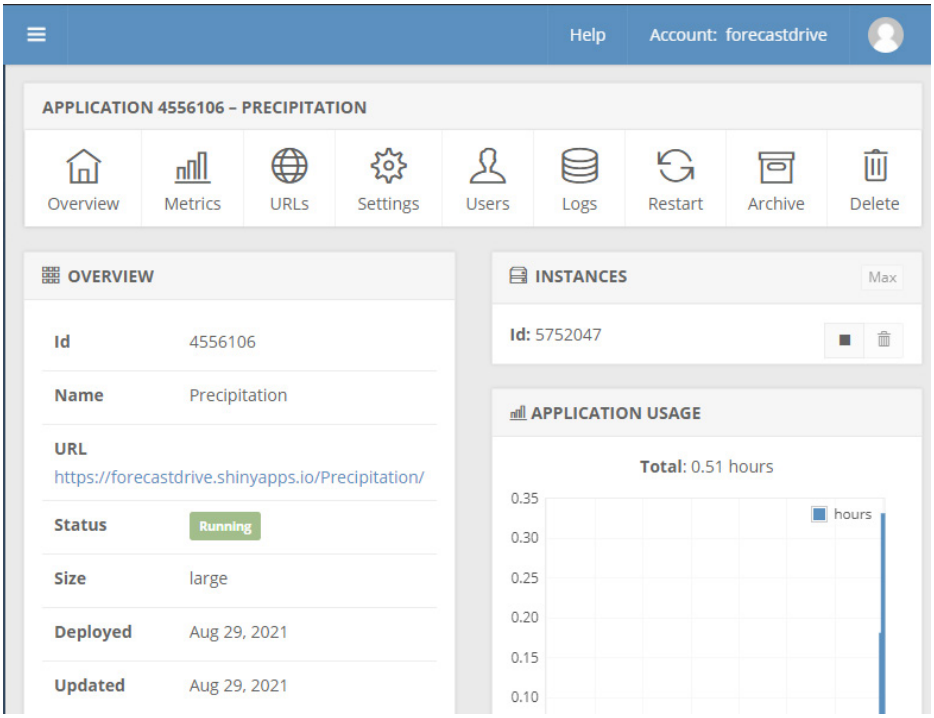

Figura 7. Dashboard Shinyapps.io server

La página principal de la aplicación https://forecastdrive.shinyapps.io/Precipitacion/ mostrada en la Figura 8, cuenta con un informativo de las funciones que presenta, en la parte inferior se encuentra un botón (Browser), el cual cumple la función de buscar y subir el archivo preparado con la información y separado con comas para realizar el análisis predictivo, mientras que en la parte izquierda encontramos dos menús adicionales exploración y pronóstico, un input del rango de fechas indicando las que van a ser tomadas en cuenta, y un slidebar que indicar los meses de pronóstico que son requeridos.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

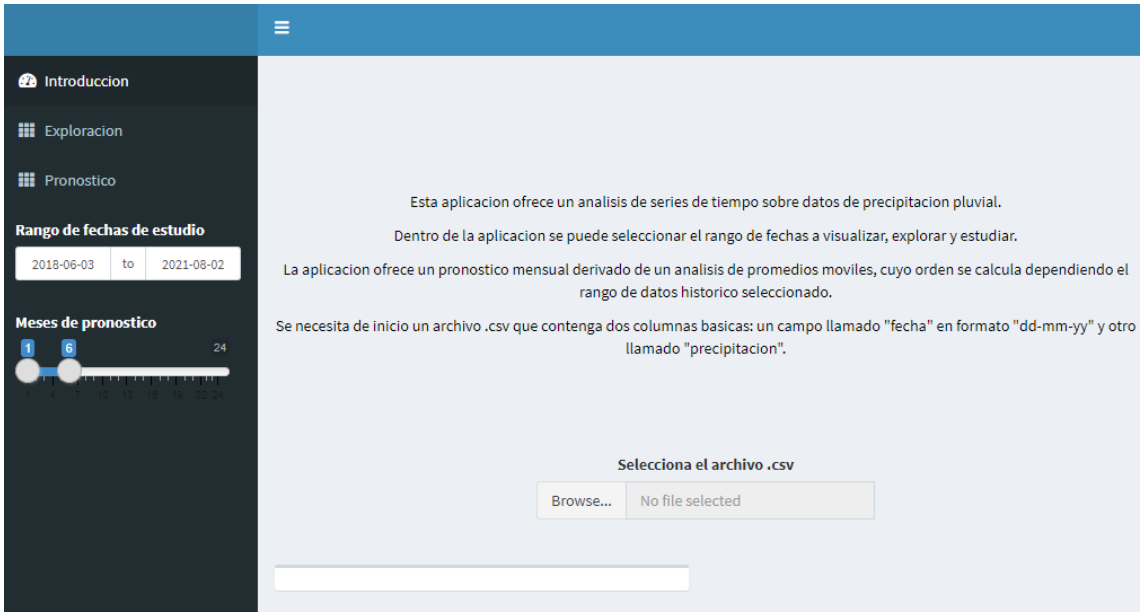

Figura 8. Interfaz aplicación web de despliegue de los datos

En el menú de exploración indicado en la Figura 9 lo primero que se observa es la serie de tiempo de la relación fecha y precipitación de nuestra tabla de datos base, a medida que modificamos los campos del rango de tiempo estos cambiarán mostrando al usuario la serie de tiempo deseada en el momento deseado.

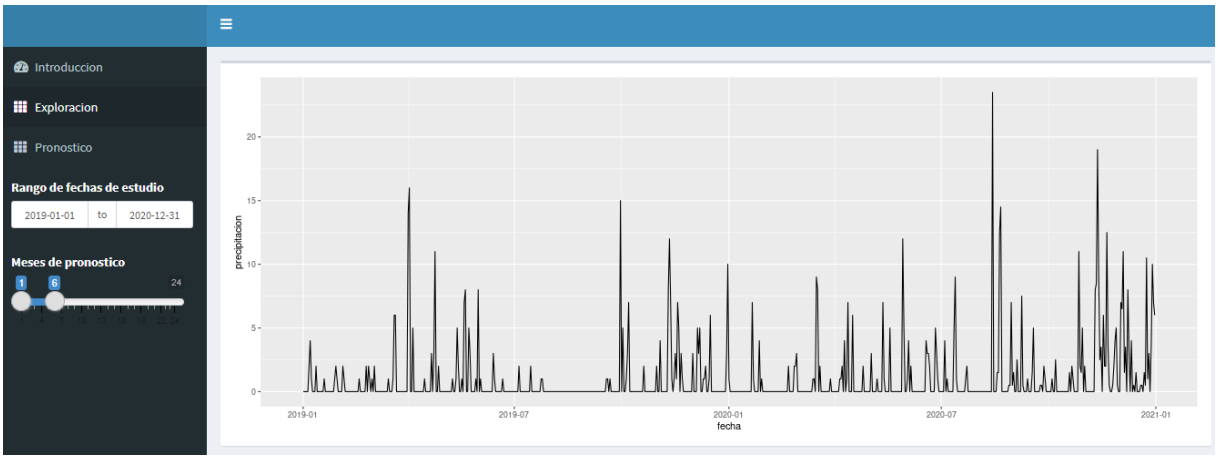

Figura 9. Serie de tiempo Exploración

En la parte superior del menú de exploración dos tablas que muestran 10 entradas este campo pueden ser modificados, en esta instancia se muestra la fecha y el valor de precipitación por día, y la precipitación acumulada, en esta parte de la exploración ya comienza a trabajar el modelo anteriormente realizado.

Además, este apartado cuenta con tres botones de copia, CSV, Excel, permitiendo descargar en estos dos tipos de formatos los datos que sean requeridos como se puede visualizar en las figuras 10 y 11.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

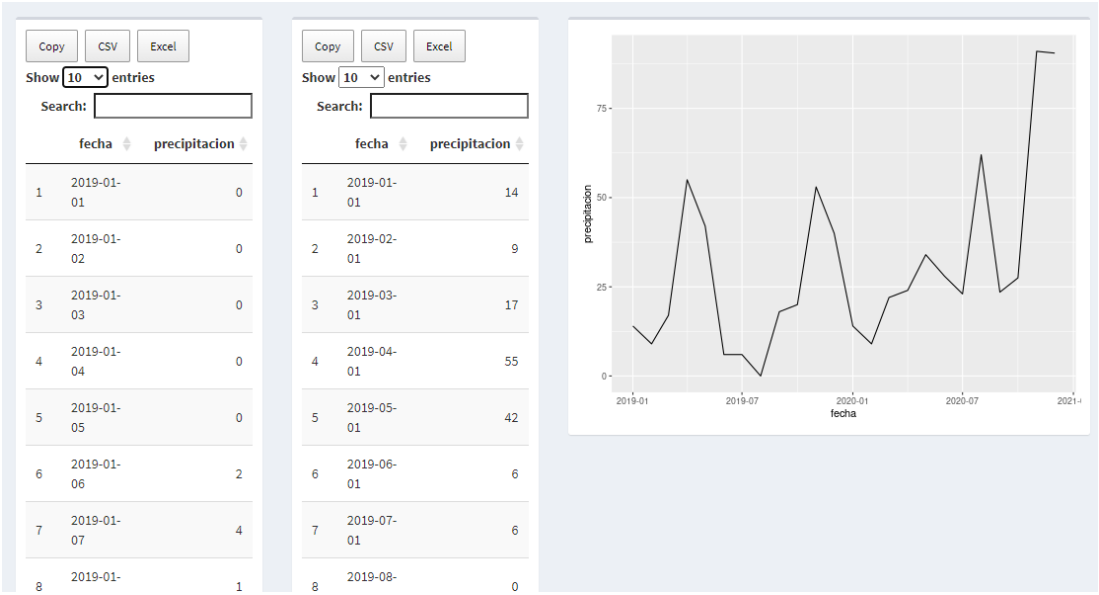

Figura 10. Tablas de Exploración

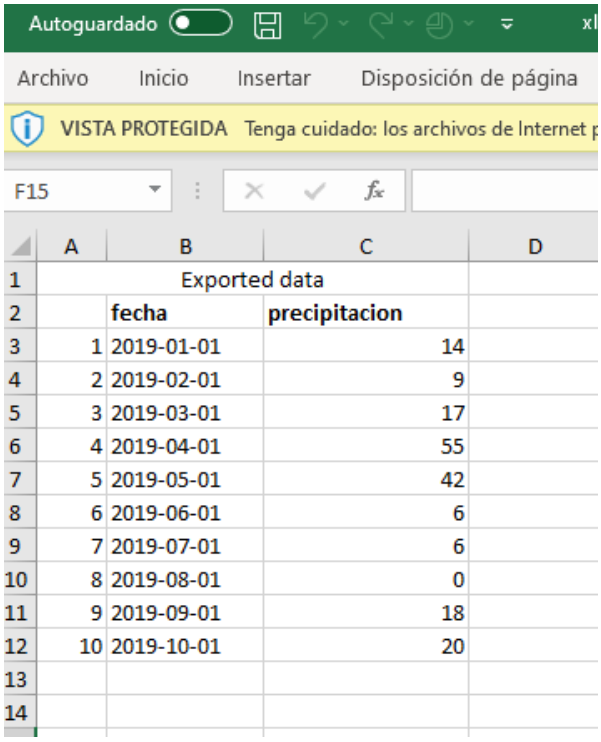

Figura 11. Archivo descargado en formato xlsx

El ultimo menú muestra el histograma con el pronóstico de precipitaciones pluviales aplicado con el modelo de SMA representado por las figuras 12 y 13, además de igual forma muestra las tablas con los mismos botones de descarga, pero con la diferencia de que aquí obtenemos los datos de los valores ajustados y de los pronósticos para cada mes respectivamente.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

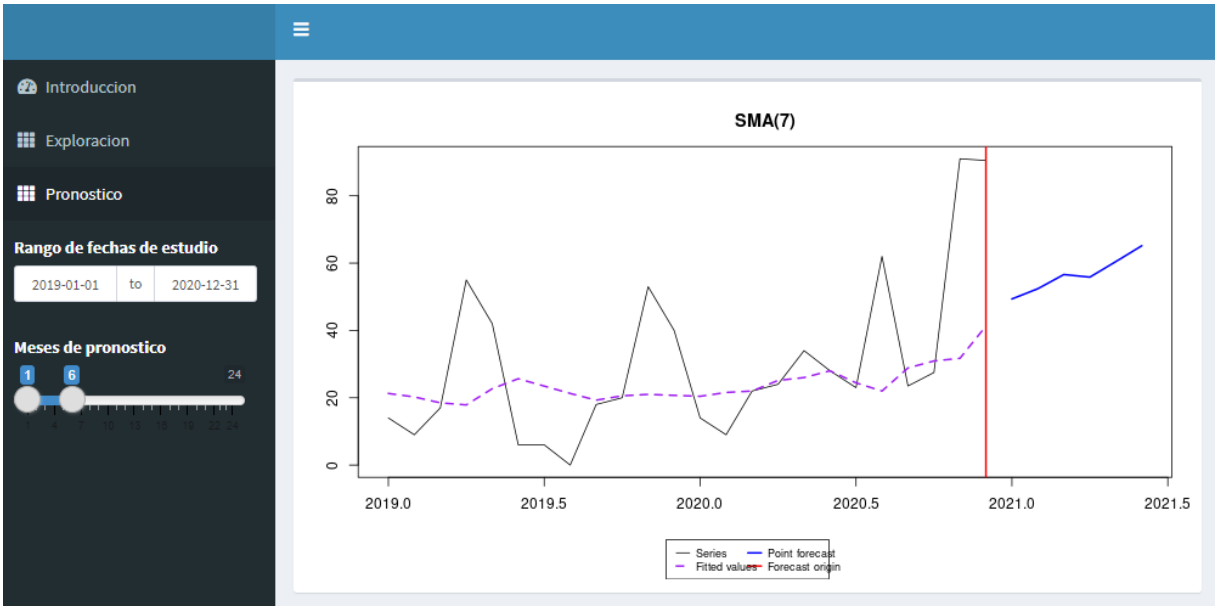

Figura 12. Histograma SMA Pronóstico

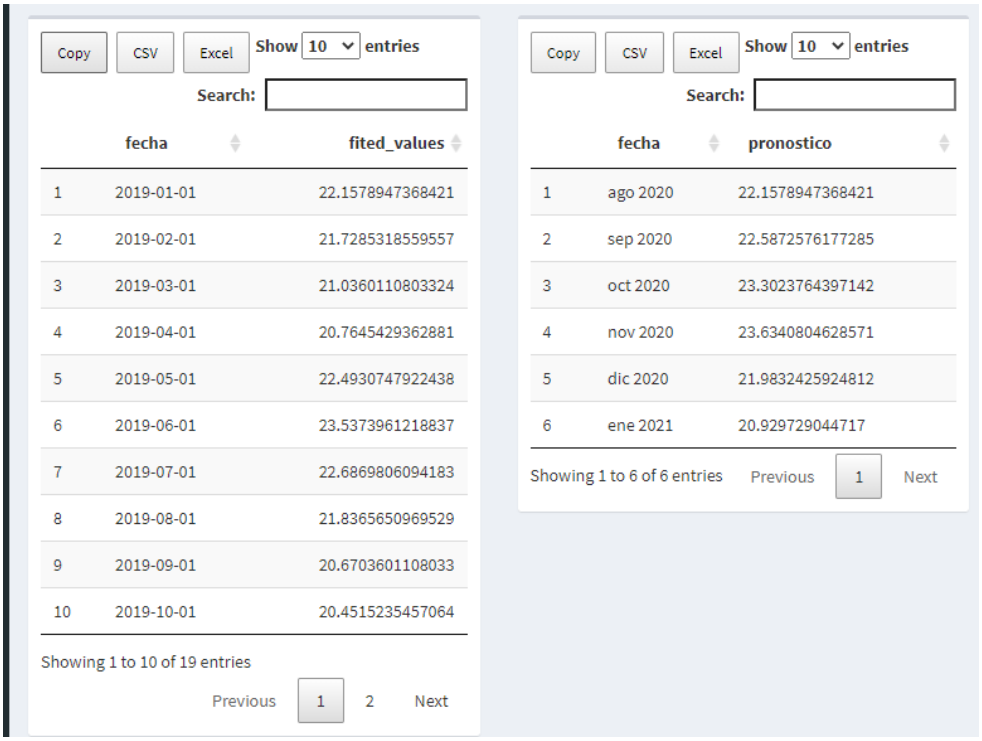

Figura 13. Tablas de datos valores ponderados y pronostico

Cabe destacar que el aplicativo informático es responsive, esto quiere decir que la aplicación puede ser abierta desde el navegador web de cualquier dispositivo móvil con acceso a internet sin perder ninguna de las funcionalidades que tiene, adaptándose al tamaño de pantalla disponible.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

Conclusiones

La utilización de diferentes métodos y metodologías fueron de gran ayuda para la presente investigación, al igual que las herramientas de recolección de datos que permitieron generar conocimiento a partir de la información obtenida de la estación metereologíca, permitiendo conocer como las precipitaciones afectan la calidad, oferta y demanda de los cultivos al ser una actividad que depende mucho de los fenómenos climáticos. El presente modelo de pronóstico basado en promedios móviles es una buena aproximación hacia el pronóstico acumulado mensual de precipitación mensual. Su implementación en aplicaciones interactivas permite una exploración de datos muy práctica y flexible, permitiendo interpretar el análisis de datos históricos de la estación meteorológica de la fina experimental San Francisco de Huaca de una manera más entendible para predecir futuras precipitaciones. Finalmente, el análisis de series de tiempo resultan ser una poderosa herramienta para la evaluación proyectos de investigación.

Recomendaciones

Conseguir una base de datos con tanto histórico como sea posible, ya que mientras más datos se tenga mayor será la precisión de la predicción, permitiendo de esta manera tener resultados óptimos y tomar mejores decisiones. Es posible generar mejores modelos a partir de más puntos de colección de datos, por lo que agregar información de otras estaciones meteorológicas sería recomendable. Hacer pruebas de calidad sobre la obtención de datos, ya que en el presente trabajo se observa como los valores a partir del año 2021 se vuelven fuera del comportamiento normal, y esto puede ser debido a fallas técnicas de la estación. Generar modelos de pronóstico diario es posible si se agrega información con filtros de calidad adecuados, así como también sistematizar las estaciones meteorológicas para tener un mejor manejo de los datos. Generar modelos de pronóstico a partir de otras técnicas y metodologías de minería de datos para obtener diferentes resultados y hacer un estudio de que técnica es más aceptable en la aplicación de predicciones en precipitaciones pluviales.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201

Referencias

Escutia, I. (2019). Descomposición de series de tiempo. Recuperado de https://rstudio-pubs-static.s3.amazonaws.com/546278_da6272365edd4444a4e7ccf6b17978fe.html

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. del P. (2018). Metodología de la Investigación (Quinta Edi). México: Mc Graw Hill.

Hidalgo-Guijarro, J., Yandún-Velasteguí, M., Bolaños-Tobar, D., Borja-Galeas, C., Guevara, C., Varela-Aldás, J., ... & Rivera, R. (2019, September). Preprocessing Information from a Data Network for the Detection of User Behavior Patterns. In International Conference on Human Systems Engineering and Design: Future Trends and Applications (pp. 661-667). Springer, Cham.

Ilbay Yupa, M. L. (2019). "tendencia espacio-temporal de la precipitación, su agresividad y concentración en la región interandina del ecuador". Universidad Nacional Agraria La Molina.

IONOS. (2018). Software de data mining: realiza análisis de datos más efectivos. Recuperado de https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

León Guzmán, E. (2017). Minería de Datos Módulo Diplomado. Módulo Universidad Nacional de Colombia, 1–19. Recuperado de http://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf

Mancero, H. (2017). Estimación del coeficiente de hurst de las series temporales de tráfico vehicular en zonas urbanas. Universidad de las Fuerzas Armadas.

Ortiz Farro, P. E. (2017). "Minería de datos con series de tiempo en el desarrollo e implementación del sistema inteligente que predice la producción de arroz en el ámbito de la Gerencia Regional de Agricultura – Lambayeque." Universidad Señor De Sipán.

Rivera, S. H., Lema, L. Z., Freire, A. M., Rojas, L. V., & Villa, A. E. (2018). Métodos de clasificación en minería de datos meteorológicos. Revistas ESPOCH, 107–113.

Unir. (2019). Lenguaje R, ¿qué es y por qué es tan usado en Big Data? Recuperado de https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/

Vazquez, M. (2018). Minería de datos para generación de reglas de tendencia de precipitación pluvial en el estado de Morelos. Ingeniería-Revista Académica de la facultad de Ingeniería, 22, 9–24.

Zamora Villalobos, T. F. (2018). Aplicación De Tecnicas De Mineria De Datos Para Pronósticos Del Sector Agrícola. Pontificia Universidad Católica De Valparaíso.

Cómo citar este artículo:
Yandún, M. & Montenegro, C. (Enero - Junio de 2022). Minería de datos para series temporales y su aplicación en las precipitaciones pluviales en la zona del cantón Huaca Ecuador. Sathiri (18)1, 230-241. https://doi.org/10.32645/13906925.1201