**Pronóstico de Series de Tiempo de Tráfico web. Estudio de Caso Universidad Nacional de Misiones**

POR: Enrique Gauto Sand

Propuesta de Tesis presentada a la Facultad de Ciencias Exactas, Químicas y Naturales de la Universidad Nacional de Misiones para optar al grado académico / título profesional de Licenciado en Sistemas de Información de la carrera de Licenciatura en Sistemas de Información

Directora: Alice Rambo

Prof. Tit. de la Cátedra “Trabajo Final”: Dr. Kuna, Horacio Daniel

Índice

[1. Presentación de Propuesta 1](#_Toc114647313)

[2. Justificación 2](#_Toc114647314)

[3. Delimitación del Tema 3](#_Toc114647315)

[3.1. Descripción de las variables de forma conceptual y operacional: 3](#_Toc114647316)

[4. Objetivos 4](#_Toc114647317)

[4.1. Objetivo General: 4](#_Toc114647318)

[4.2. Objetivos Específicos: 4](#_Toc114647319)

[5. Marco Teórico 5](#_Toc114647320)

[5.1. Tráfico web 5](#_Toc114647321)

[5.2. Series de tiempo 5](#_Toc114647322)

[5.3. Pronóstico de tráfico web usando series de tiempo 5](#_Toc114647323)

[5.4. LSTM (*Long Short Term Memory*) 6](#_Toc114647324)

[6. Planificación Metodológica 7](#_Toc114647325)

[8. Resultados Esperados 10](#_Toc114647326)

[9. Bibliografía: 11](#_Toc114647327)

|  |
| --- |
| Catedra: TESIS DE GRADO  Carrera: Licenciatura en Sistemas de Información  Facultad: FCEQyN  UNaM  PROPUESTA DE TESIS |

# Presentación de Propuesta

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre del  Trabajo | PRONÓSTICO DE SERIES DE TIEMPO DE TRÁFICO WEB.  ESTUDIO DE CASO UNIVERSIDAD NACIONAL DE MISIONES | | |
| Directora | Alice Rambo | | |
| Codirector | No se ha elegido un codirector  todavía. | | |
| Alumnos | Nro. Alumno | Apellido y Nombres | Dirección, Teléfono y e-mail |
|  | 1 | Gauto Sand, Enrique | Casa 30 barrio Los Álamos  Posadas, Misiones +54-3764-247399  egsand98@gmail.com |
| Clasificación | Investigación  Aplicada | | |
| Plazo de Ejecución | 01/08/2022 – 07/07/2023 | | |

# Justificación

El pronóstico del tráfico web es un problema importante hoy en día ya que saber cuánta demanda se tendrá ayuda a que no se exceda la carga de trabajo de los servidores y brindar un buen servicio a los clientes, un mal tiempo de respuesta por parte de los servidores pudiera llevar a la pérdida de clientes[1].

El uso del pronóstico del tráfico web no tiene como fin únicamente el de evitar que los servidores no se congestionen, además sirve para cuestiones de venta de espacios publicitarios, porque los sitios publicitarios necesitan estimar cuántas visitas tendrán las páginas, donde residirán sus anuncios, para poder estimar ganancias y saber si les será rentable publicitar en dicha página[2].

Las páginas que residen en la nube usan el pronóstico de tráfico web para determinar una estrategia eficaz para el equilibrio de carga[3], aplicando tecnologías de auto-escalado de servidores de manera que el rendimiento de las páginas no decaen, debido a que esta tecnología permite crear y remover recursos en la nube adaptándose a la demanda[4].

El presente trabajo pretende hacer análisis de tráfico web para poder estimar y predecir futuras demandas en las páginas institucionales de la Universidad Nacional de Misiones – U.Na.M. Para lo cual se consideraran indicadores como volumen de tráfico, cantidad de conexiones, cantidad de visitantes, velocidad de carga de las paginas, páginas vistas, promedio de páginas por vista, porcentaje de rebote, entre otros indicadores según la disponibilidad de los mismos. Siendo provisto por parte del personal técnico referente institucional acceso a las herramientas de captura de datos estadístico de los siguientes enlaces:

* Sitio Institucional de la U.Na.M. disponible en <https://unam.edu.ar/>
* Sitio de la editorial Universitaria disponible en <https://editorial.unam.edu.ar/>
* Portal de acceso al contenido generado por la radio y la televisión de la Universidad disponible en <https://transmedia.unam.edu.ar/>

De los sitios mencionados anteriormente se extraerán los datos para utilizar en este proyecto con la finalidad de realizar el pronóstico de series de tiempo de tráfico web. Inicialmente como herramienta de acceso a los datos se dispone *google analytics*[5]con permisos cedidos por los administradores.

# Delimitación del Tema

El alcance de este estudio será descriptivo, ya que se pretende describir el desempeño de un conjunto de modelos para el pronóstico de series de tiempo de tráfico web sobre los datos de U.Na.M. (<https://transmedia.unam.edu.ar/> ).

## Descripción de las variables de forma conceptual y operacional:

|  |  |
| --- | --- |
| Variables | Conceptual |
| Tráfico web | El tráfico web es cada vez que un usuario visita la página web a través de internet[3]. |
| Desempeño | El desempeño es la forma de evaluar el modelo, se utiliza para comparar los valores predichos del modelo entrenado, con los valores observados[6]. |
| Variables | Operacional |
| Trafico web | El tráfico web se mide registrando cada vez que alguien visita la página web, una plataforma que realiza esta tarea es *google analytics,* y realiza la medición del tráfico web en forma de series de tiempo[5]. |
| Desempeño | Para evaluar el desempeño de los modelos se opta por usar el *Root Mean Square Errors* (RMSE): cuadrática media o el error cuadrático medio es una medida de las diferencias entre los valores predichos por un modelo y los valores observados.  Donde es el valor predicho y es el valor real[7]. |

# Objetivos

## Objetivo General:

Pronosticar series de tiempo de tráfico web y medir el desempeño de modelos aplicados sobre datos extraídos de la página institucional de la U.Na.M.

## Objetivos Específicos:

-Relevar y analizar distintos modelos de pronóstico de series de tiempo de tráfico web.

-Determinar la adecuación de cada uno de los modelos propuestos de la bibliografía a la problemática.

-Definir indicadores para medir el desempeño de los modelos.

-Generar los modelos a utilizar.

-Realizar la evaluación del desempeño de los modelos.

-Validar los modelos.

# Marco Teórico

## Tráfico web

El tráfico web es cada vez que un usuario visita la página web a través de internet[3].

## Series de tiempo

Son una sucesión de datos medidos secuencialmente en un tiempo.

Entre los componentes de las series de tiempo está la tendencia la cual es el movimiento sostenido a largo plazo de la serie, otra componente es la cíclica que serían las fluctuaciones que se repiten en forma periódica mayores a un año alrededor de la tendencia, luego está la componente estacional que son movimientos oscilatorios durante el año, cuya duración es menor al año como por ejemplo las estaciones del año y por último está la componente irregular que serían movimientos esporádicos, aleatorios que no se pueden predecir.

## Pronóstico de tráfico web usando series de tiempo

El pronóstico de series de tiempo se utiliza para predecir valores futuros basados en el valor observado en el pasado[3]. Entre las formas del pronóstico de series de tiempo de tráfico web existe la técnica de redes neuronales *Long Short Term Memory* (LSTM)[8] con entrenamiento asíncrono distribuido, cuya métrica de desempeño es la MAE (*mean absolute error*)[7] y la pérdida de Huber[9] función para probar la precisión del modelo han logrado un buen grado de asertividad con 200 épocas pero en el documento científico[1] menciona que mejora con 500 y 100 siendo una conclusión no del todo certera la de si se aumentan las épocas y reduciendo considerablemente el MAE el pronóstico de series de tiempo es cada vez más preciso[1], también hay otro modelo para el pronóstico de series temporales de tráfico web usando la técnica de *Discrete Wavelet Transform* (DWT)[10], *Auto-Regressive Integrated Moving Average* (ARIMA)[11] y LSTM *Recurrent Neural Network* (RNN)[12], dividieron los datos en dos conjuntos uno de entrenamiento y otro de pruebas aplican DWT para separar los datos en lineales y no lineales, a los lineales le aplica ARIMA y a los no lineales LSTM sus resultados fueron que los pronósticos dados por LSTM RNN mostraron un patrón que detecta picos con precisión, unos de los problemas que puede tener es que la función de pérdida del gradiente puede decaer exponencialmente con el tiempo en redes RNN lo cual debería solucionarse con LSTM[12]; también está el pronóstico de tráfico web usando el modelo de series de tiempo Prophet[2], Prophet agrega la contribución importante de utilizar un modelo de series de tiempo descomponible con tres componentes principales del modelo: tendencia, estacionalidad y días festivos(*holidays*), siendo los *holidays* la importante diferencia con modelos como ARIMA ya que por ejemplo en navidad, otros días festivos y eventos, generan perturbaciones predecibles en una serie de tiempo[2],

Otra tecnología para predecir el tráfico web son las redes *Generative Adversarial model* (GAN)[13] con LSTM como generador y un perceptrón multicapa (MLP)[13] como discriminador para generar series de tiempo dado el dataset real, finalmente se realizaría el pronóstico con la librería prophet[13], también está como tecnóloga de pronóstico el modelo state augmented Poisson-gamma state space ((sa-)PGSS)[14] cuya estructura se adapta de manera flexible a los conteos recién observados y admite un algoritmo de Monte Carlo[14] secuencial para actualizaciones en línea, esto para distribuciones predictivas posteriores, el cual es un algoritmo de rápida adaptación para el aprendizaje de parámetros secuenciales, el monitoreo del sistema y la previsión de la demanda[14].

Referente a antecedentes que se basen en datos de *google analytics*, hay algunos donde se pretende pronosticar el tráfico web para predecir los recursos web a utilizar [15], donde se menciona el uso del modelo ARIMA, RNN, *Convolutional neural networks* (CNN)[16] y redes *Nonlinear autoregressive exogenous* (NARX)[15], además hay otro antecedente donde se usan datos de google analytics[17] donde se mencionan tecnologías como *Gaussian Process*[18]*, MLP, Linear Regression*[19] *y Sequential Minimal Optimization*[20]*.*

## LSTM (*Long Short Term Memory*)

Las redes de memoria a corto y largo plazo son una de las variantes de redes neuronales recurrentes que son capaces de aprender los patrones a largo plazo y corto plazo[8].

# 

# Planificación Metodológica

Se opta por utilizar la metodología CRISP-DM[21] ya que hoy en día es un estándar de facto para los proyectos de ciencia de datos y minado de datos [22].

Las fases generales de la metodología CRISP-DM son las siguientes:

* Comprensión del negocio:

Esta fase inicial se enfoca en comprender los objetivos y requisitos del proyecto desde una perspectiva de negocios, y luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar del proyecto diseñado para lograr los objetivos[23].

* Comprensión de los datos:

Esta fase inicia con recolectar los datos y continúa con actividades para familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir los primeros conocimientos de los datos o detectar subconjuntos interesantes para formular hipótesis[23].

* Preparación de los datos:

La fase de preparación cubre todas las actividades para construir el dataset final a partir de los datos iniciales. Las tareas de preparación de datos se realizan varias veces y no en ningún orden prescrito[23].

* Modelado de los datos:

En esta fase, se seleccionan y aplican varias técnicas de modelado, y sus parámetros se calibran a valores óptimos[23].

* Evaluación:

En la fase de evaluación, los resultados son evaluados con los objetivos comerciales que se definieron. Por lo tanto, los resultados deben interpretarse y deben definirse acciones adicionales[22].

* Desarrollo:

La creación del modelo generalmente no es el final del proyecto. Según los requisitos, la fase de desarrollo puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minado de datos repetible[23].

1. **Planificación Temporal**

|  |  |  |
| --- | --- | --- |
| **Actividad** | **Tiempo Estimado** | **Referencia** |
| Relevar y analizar distintos modelos de pronóstico de series de tiempo de tráfico web.   * Selección de los buscadores académicos * Generación de filtros de búsqueda * Aplicación de filtros de búsqueda y evaluación de resultados. | 1 mes | A |
| Tratamiento de los datos | 3 meses | B |
| Determinar la adecuación de cada uno de los modelos propuestos de la bibliografía a la problemática.   * Listar los modelos propuestos resultantes de la búsqueda bibliográfica. * Comparar los modelos. * Seleccionar los modelos que se adecuen a la problemática. | 1 mes | C |
| Definir indicadores para medir el desempeño de los modelos.   * Realizar una búsqueda bibliográfica de los indicadores para medir el desempeño. * Analizar y seleccionar los indicadores para medir el desempeño. | 1 mes | D |
| Generar los Modelos   * Generar la prueba de diseño. * Construir los modelos. | 1 meses | E |
| Realizar la evaluación del desempeño de los modelos.   * Evaluar y ajustar los modelos. | 1 mes | F |
| Validar los modelos. | 1 mes | G |
| Realizar la escritura del documento de tesis. | 11 meses | H |
| TOTAL | 11 meses | |

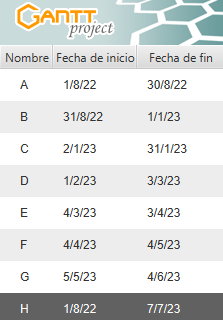


Figura 1. Tiempo estipulado para cada actividad.

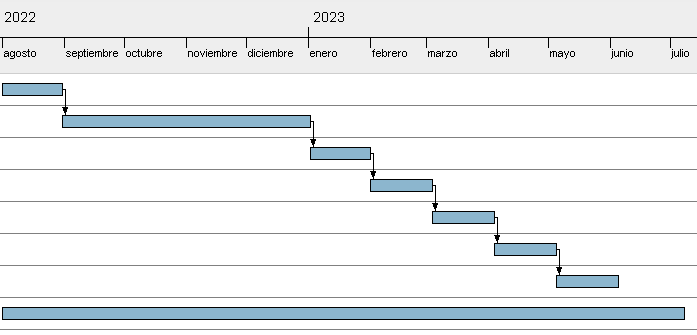


Figura 2. Diagrama de Gantt.

Según lo estipulado en la planificación temporal el desarrollo de este proyecto tendrá una duración de 11 meses.

# Resultados Esperados

Este proyecto de investigación brindaría a la U.Na.M. un modelo capaz de pronosticar el tráfico web, se espera que el desempeño del modelo justifique el uso del mismo.

Entre algunos puntos importantes se pueden citar:

* Detectar patrones de estacionalidad y tendencia, referente al tráfico web de las páginas institucionales de la U.Na.M.
* Detectar e informar indicadores que permitan predecir el comportamiento de los servidores para permitir la toma de decisiones que posibilite el mayor rendimiento de los mismos.
* Ayudar a entender el comportamiento de los usuarios de las páginas institucionales.

Las hipótesis que guiarán a este estudio son las siguientes:

* Existe algún modelo que pueda representar el pronóstico de series de tiempo de tráfico web en la página institucional de la U.Na.M.

# Bibliografía:

[1] R. Casado-Vara, A. M. del Rey, D. Pérez-Palau, L. De-La-fuente-valentín, and J. M. Corchado, “Web Traffic Time Series Forecasting Using LSTM Neural Networks with Distributed Asynchronous Training,” *Math. 2021, Vol. 9, Page 421*, vol. 9, no. 4, p. 421, Feb. 2021, doi: 10.3390/MATH9040421.

[2] A. Subashini, S. K, S. Saranya, and U. Harsha, “Forecasting Website Traffic Using Prophet Time Series Model,” *Int. Res. J. Multidiscip. Technovation*, pp. 56–63, Jan. 2019, doi: 10.34256/IRJMT1917.

[3] K. P. Paun and C. H. Makwana, “A Survey on Web Traffic Forecasting on Time Series Data,” *J. Appl. Sci. Comput.*, vol. VI, no. 3588, pp. 3588–3594, 2019.

[4] N.-M. Dang-Quang and M. Yoo, “An Efficient Multivariate Autoscaling Framework Using Bi-LSTM for Cloud Computing,” *Appl. Sci.*, vol. 12, no. 7, p. 3523, Mar. 2022, doi: 10.3390/APP12073523.

[5] B. Plaza, “Monitoring web traffic source effectiveness with Google Analytics An experiment with time series,” *Aslib Proc. New Inf. Perspect.*, vol. 61, no. 5, pp. 474–482, Sep. 2009, doi: 10.1108/00012530910989625.

[6] A. Botchkarev, “Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology,” 2018, [Online]. Available: http://arxiv.org/abs/1809.03006.

[7] W. Wang and Y. Lu, “Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model,” in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 324, no. 1, doi: 10.1088/1757-899X/324/1/012049.

[8] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[9] G. P. Meyer, “An Alternative Probabilistic Interpretation of the Huber Loss,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5257–5265, doi: 10.1109/CVPR46437.2021.00522.

[10] M. Weeks and M. Bayoumi, “Discrete Wavelet Transform: Architectures, Design and Performance Issues,” *J. VLSI signal Process. Syst. signal, image video Technol.*, vol. 35, no. 2, pp. 155–178, 2003, doi: 10.1023/A:1023648531542.

[11] Aki Taanila, “Time series forecasting 1 Time series forecasting,” *Data Sci.*, no. vi, p. 25, Jan. 2010, doi: 10.1016/B978-0-12-814761-0.00012-5.

[12] T. Shelatkar, S. Tondale, S. Yadav, and S. Ahir, “Web Traffic Time Series Forecasting using ARIMA and LSTM RNN,” *ITM Web Conf.*, vol. 32, p. 03017, 2020, doi: 10.1051/ITMCONF/20203203017.

[13] K. Zhou, W. Wang, L. Huang, and B. Liu, “Comparative study on the time series forecasting of web traffic based on statistical model and Generative Adversarial model,” *Knowledge-Based Syst.*, vol. 213, p. 106467, Feb. 2021, doi: 10.1016/J.KNOSYS.2020.106467.

[14] K. Irie, C. Glynn, and T. Aktekin, “Sequential modeling, monitoring, and forecasting of streaming web traffic data,” *Ann. Appl. Stat.*, vol. 16, no. 1, Mar. 2022, doi: 10.1214/21-AOAS1505.

[15] P. Khanarsa, A. Luangsodsai, K. Sinapiromsaran, I. F. Astachova, K. A. Makoviy, and Y. V Khitskova, “Possibilities for predicting the state of usability web resources,” *J. Phys. Conf. Ser.*, vol. 1902, no. 1, p. 012029, May 2021, doi: 10.1088/1742-6596/1902/1/012029.

[16] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv Prepr. arXiv1511.08458*, 2015.

[17] C. Napagoda and H. Singh, “Web Site Visit Forecasting Using Data Mining Techniques,” *Int. J. Sci. Technol. Res.*, vol. 2, 2013, Accessed: Jun. 27, 2022. [Online]. Available: www.ijstr.org.

[18] M. SEEGER, “GAUSSIAN PROCESSES FOR MACHINE LEARNING,” *Int. J. Neural Syst.*, vol. 14, no. 02, pp. 69–106, Apr. 2004, doi: 10.1142/S0129065704001899.

[19] X. Yan, X. Su, and World Scientific (Firm), “Linear regression analysis : theory and computing,” p. 328, 2009.

[20] J. Platt, “Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,” 1998. Accessed: Sep. 19, 2022. [Online]. Available: https://www.researchgate.net/publication/2624239\_Sequential\_Minimal\_Optimization\_A\_Fast\_Algorithm\_for\_Training\_Support\_Vector\_Machines.

[21] P. Chapman *et al.*, “CRISP-DM 1.0: Step-by-step data mining guide,” *SPSS inc*, vol. 9, no. 13, pp. 1–73, 2000.

[22] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: 10.1016/J.PROCS.2021.01.199.

[23] W. Rüdiger and J. Hipp, “CRISP-DM : Towards a Standard Process Model for Data Mining,” *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.