**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

[Capítulo 1 1](#_Toc121065543)

[Introducción 1](#_Toc121065544)

[1.1. Motivación 2](#_Toc121065545)

[1.2. Objetivos 2](#_Toc121065546)

[1.2.1. Objetivo General: 2](#_Toc121065547)

[1.2.2. Objetivos Específicos: 2](#_Toc121065548)

[1.3. Estructura del documento 2](#_Toc121065549)

[Capítulo 2 3](#_Toc121065550)

[Marco Teórico 3](#_Toc121065551)

[2.1. Tráfico web 3](#_Toc121065552)

[2.2. Series de tiempo 3](#_Toc121065553)

[2.3. Pronóstico de series de tiempo de tráfico web 3](#_Toc121065554)

[Capítulo 3 6](#_Toc121065555)

[Descripción del problema 6](#_Toc121065556)

[3.1 La Universidad Nacional de Misiones 7](#_Toc121065557)

[3.2 El problema 7](#_Toc121065558)

[Capítulo 4 8](#_Toc121065559)

[Solución Propuesta 8](#_Toc121065560)

[4.1 Materiales y Métodos 9](#_Toc121065561)

[4.2 Modelos 1](#_Toc121065562)

[3.5 Selección de Modelos 1](#_Toc121065563)

[Bibliografía 2](#_Toc121065564)

# Capítulo 1

# Introducción

## Motivación

## Objetivos

### Objetivo General:

Pronosticar series de tiempo de tráfico web y medir el desempeño de modelos aplicados sobre datos extraídos de las páginas institucionales 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.

## Estructura del documento

# Capítulo 2

# Marco Teórico

En este capítulo se presenta el marco teórico del trabajo realizado, comenzando con la definición de tráfico web y series de tiempo, luego con variados ejemplos de la literatura de distintas tecnologías que realizan el pronóstico de series de tiempo de tráfico web finalizando con las formas de medir el desempeño de los modelos.

## Tráfico web

Trafico web es la cantidad de datos enviados y recibidos por los visitantes de un sitio web lo cual es determinado por el número de visitas y el número de páginas que ellos visitan[1].

## Series de tiempo

Las series de tiempo son una secuencia de datos medidos secuencialmente en el tiempo[2]; los datos de las series de tiempo son registrados en intervalos de tiempos constantes de manera sucesiva, 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, luego está la componente estacional que son movimientos repetitivos durante un ciclo de tiempo la estacionalidad se puede dividir en horas, días, semanas, meses, trimestral y anual; finalmente tenemos al ruido el cual es un componente impredecible[3].

## Pronóstico de series de tiempo de tráfico web

El pronóstico de series de tiempo es el proceso de predecir el valor futuro de los datos de una serie de tiempo basado en las observaciones pasadas y otras entradas[3]

Entre las formas del pronóstico de series de tiempo de tráfico web existe la técnica de redes *Generative Adversarial model* (GAN)[4] con *Long Short Term Memory* (LSTM)[5] como generador y un perceptrón multicapa (MLP)[4] como discriminador para generar series de tiempo dado el dataset real, finalmente se realizaría el pronóstico con la librería prophet[6] comparando la combinación de tecnologías anteriores con métodos estadísticos finalmente llegando a la conclusión de que no obtuvieron una diferencia notable.

Entre las formas del pronóstico de series de tiempo de tráfico web existe la técnica de redes neuronales *Long Short Term Memory* (LSTM)[5] con entrenamiento asíncrono distribuido, cuya métrica de desempeño es la MAE (*mean absolute error*)[7] y la pérdida de Huber[8] función para probar la precisión del modelo, han logrado un buen grado de asertividad con 200 épocas pero en el documento científico[9] 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[9].

Hay otro artículo[1] en el cual se menciona el uso de (*Recurrent Neural Networks*) RNN seq2seq[1] con la ayuda de la arquitectura encoder/decoder, el encoder es cuDNN GRU (*Gate Recurrent Unit*)[10] ya que realiza la tarea con mejor velocidad en comparación con los tensores regulares, el decoder es TensorFlow GRUBlockCell[10] realizando algunos cambios al modelo ganador de *Kaggle*[10] y usando para medir el desempeño del modelo la metrica *Symmetric Mean Absolute Percentage Error* (SMAPE)[11] obtuvieron mejores resultados[1].

En otro artículo[12], utilizan un enfoque híbrido de *particle swarm optimization* y *quantum-inspired* resultando en QPSO[12] y redes *Nonlinear autoregressive exogenous* NARX[13] resultando en NARX -QPSO[12] en dicho artículo se menciona que el modelo NARX –QPSO tiene un rendimiento sobresaliente en los resultados comparados con otros modelos.

En otro artículo[14] se tiene por propósito incorporar un totalmente automatizado y preciso modelo de pronóstico semanal, en donde tiene conjuntos de modelos base que utilizarán junto con arquitecturas de meta-aprendizaje, para generar un modelo que se adapte a más de un caso.

En otro artículo[15] tiene por propósito proponer un *Automated Machine Learning* (Auto ML) framework, una arquitectura neuronal nueva denominada Auto-PyTorch-TS[15] donde comparan con otros modelos y demuestran que su modelo tiene un mejor rendimiento.

También se propone en otro artículo[16] el uso de un modelo combinado

Prophet-LGBM[16] el cual concluye que se tiene mejores resultados comparado con modelos individuales.

Otra tecnología son las redes bidireccionales LSTM (BI-LSTM)[2] las cuales en este artículo[2] fueron probadas en el *M3-Competition*[17]dataset el cual se usa para probar modelos de pronóstico de series de tiempo dado que posee distintas categorías como micro, industria, finanzas, macro, demografía y otros; probaron Bi-LSTM sobre el dataset M3 en el periodo trimestral comparándolo con el modelo *Auto-Regressive Integrated Moving Average* (ARIMA)[3] y otros, donde BI-LSTM tuvo un mejor rendimiento.

## Evaluación de desempeño de los modelos predictivos

El desempeño es la forma de evaluar el modelo, se utiliza para comparar los valores predichos del modelo entrenado, con los valores observados[18]

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]

El MAE (*mean absolute error*)[7] es una métrica que sirve para medir el error medio absoluto.

Donde es el valor predicho y es el valor real[7],

La métrica *Symmetric Mean Absolute Percentage Error* (SMAPE)[11] básicamente es el error absoluto medio porcentual.

Siendo el valor pronosticado, es el valor observado y n es el tamaño de la muestra.

# Capítulo 3

# Descripción del problema

En este capítulo se describe la problemática a abarcar como así también la estructura organizacional.

## **El problema**

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* con permisos cedidos por los administradores.

## Estructura Organizacional

Con el fin de e1 entender el funcionamiento del Departamento de Gestión de Recursos de Redes y Comunicaciones se procede a describir su estructura organizacional, como así también los roles de los trabajadores.

## Organigrama (checar)



Figura 1: Organigrama del Departamento de Gestión de Recursos de Redes y Comunicaciones

El Departamento de Gestión de Recursos de Redes y Comunicaciones, gestiona la infraestructura, además pone a disposición las herramientas para la puesta en servicio de las páginas web y servicios añadidos. Pero no controla cómo los contenidos, diseño o función de los sitios.

* **Jefe de departamento**: Realiza las tareas propias del jefe de departamento, también se encarga de mantener las plataformas de los sitios web, por ejemplo Moodle para Aulas Virtuales.
* **Técnico 1:** Se encarga principalmente de la gestión y mantenimiento de los servicios de VoIP, redes WiFi, gestión de las VLAN y redes de la unidad central de Rectorado, gestión de usuarios en la plataforma de Google Workspace for education, entre otros.
* **Técnico 2 y Técnico 3:** Estos dos técnicos se encargan de la gestión  y ruteo de las redes de la universidad y con los proveedores de servicio, gestión y mantenimiento de los Servidores físicos y virtuales del CPD, plataformas de Backup, sistemas de monitoreo de las plataformas y servicios esenciales de la infraestructura de la universidad como DNS, BGP. OSPF, VPN, etc.

# Capítulo 4

# Solución Propuesta

En este capítulo se describe la metodología seleccionada para la solución, como así también las herramientas utilizadas, y además los modelos.…

## Materiales y Métodos

Para el desarrollo del presente trabajo se utilizará la metodología CRISP-DM[19] como guía, ya que hoy en día es un estándar de facto para los proyectos de ciencia de datos y minado de datos[20].

|  |  |  |
| --- | --- | --- |
| **Actividad arreglar** | **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 | |

## Herramientas y Materiales

### Python

Python según dice en su página oficial es un lenguaje de programación interpretado, soporta múltiples paradigmas como pudiera ser programación funcional o programación orientada a objetos, Python es un lenguaje multipropósito, es decir, sirve para solucionar distintos tipos de problemas[21], para este trabajo se utilizaron las versiones de Python 3.7.x y 3.10.x.

### Google Drive

Google ofrece una plataforma que permite almacenar archivos para uso personal a sus usuarios en la nube, este servicio además permite compartir archivos entre usuarios de Google Drive, como así también la edición de archivos en tiempo real con las aplicaciones que nos permite conectar[22].

### Universal Analytics (UA)

Esta es una propiedad de Analytics, básicamente según el soporte de Google cuando se registran vistas u otro tipo de datos del trafico web se recopilan en la propiedad; Universal Analytics es la generación anterior de Analytics desde el 1 Julio del 2023 dejó de procesar datos[23].

### Google Analytics 4(GA4)

GA4 es una propiedad de Analytics, según el soporte de Google, GA4 recoge datos y los procesa de las páginas web y aplicaciones, mientras que UA no va a funcionar más GA4 es la nueva generación, GA4 por su parte recompila datos desde antes de la fecha en la que UA dejo de funcionar[23].

### Looker Studio

Según la documentación de Google es una herramienta para realizar informes y visualizar los datos, esta herramienta permite la conexión de fuente de datos propios de los usuarios como pudiera ser Google Analytics[23], también se pueden exportar datos de tablas de informes de Looker Studio en archivos CSV(Autoría propia).

### Word

### Excel

### Librerías Python numpy, tensorflow, keras, pandas

### Google Colaboratory

Según su página oficial Colaboratory es un producto de Google Research que permite a cualquier usuario escribir y ejecutar código de Python en la nube. Este producto se pude conectar con google drive para leer los archivos de google drive[24].

### Git

### Github

## **4.2 Modelos**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tecnologías** | **Red Neuronal** | **Resultados  según Artículos científicos** | **Finalidad** | **Aplica** |
| *Multi-Layer Perceptron*  (MLP) | SI | - | Clasificación, predicción | No |
| *Long-Short Term Memory*  (LSTM) | SI | Combinándolo con otras tecnologías se tiene mejor resultado que usarlo individualmente. | Predictivo, hallar patrones | No |
| ARIMA | No | Buenos | No es redes neuronales, es estadística  predictiva (predicción de series temporales) | No |
| *Gate recurrent unit* (GRU) | SI | El modelo ganador de Kaggle utiliza un cuDNN GRU que es más rápido que un GRU normal. | Sistemas predictivos basados en información secuencial. | No |
| *Bidirectional-LSTM* | SI | Mejor rendimiento comparándolo en el *M3-Competition* | Reconocimiento y clasificación | No |
| GAN | SI | Regular, no hay ganancia comparado con ARIMA | Creación artificial | No |
| RNN seq2seq | SI | Buenos | Modelo ganador del concurso de *kaggle* de *web traffic* 2017 | Si |
| NARX -QPSO | SI | Buenos | Predecir el siguiente valor de la señal de entrada | Si |
| *Prophet-LGBM* | SI | Buenos Mejores resultados que modelos individuales | Combinación ligthGBM y Prophet para predecir trafico web | Si |

Tabla 1 Comparación de tecnologías

## **3.5 Selección de Modelos**

En la Tabla 1 se presenta una comparación de las tecnologías, de ahí se procede a seleccionar las tecnologías que si aplican, son las que se consideran que se podrían usar para la solución del problema; la razón de descartar algunas tecnologías son porque no usan redes neuronales, otras son tecnologías cuya finalidad no es la del pronóstico de series de tiempo y por último, de acuerdo a los artículos científicos presentados en el marco teórico, algunas tecnologías no tienen buen rendimiento o el rendimiento suficiente para ser tomado en cuenta.

Entre los modelos que quedan seleccionados están RNN seq2seq, NARX –QPSO y Prophet-LGBM de estos modelos Prophet-LGBM si bien se mencionan las fechas de los datos que utilizaron no está disponible el dataset para compararlo con los datos de la problemática de esta tesis por lo tanto Prophet-LGBM se descarta como posible tecnología para la solución de la problemática de esta tesis, mientras que por otro lado tanto RNN seq2seq y NARX –QPSO ambos utilizan el mismo conjunto de datos el cual está disponible online para que cualquiera lo pueda descargar, de esta manera y dado que RNN seq2seq es el ganador del concurso de Kaggle(citar) y por lo tanto se encuentra más información de la misma en la página de Kaggle se opta por utilizar esta tecnología.

En cuanto a la métrica seleccionada para la medición del desempeño de los modelos, elijo la métrica SMAPE, debido a que es la misma con la que se evaluo el modelo RNN seq2seq es el ganador del concurso de Kaggle.

# Bibliografía

[1] N. Petluri and E. Al-Masri, “Web Traffic Prediction of Wikipedia Pages,” *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 5427–5429, 2019, doi: 10.1109/BigData.2018.8622207.

[2] D. Quoc Nguyen, M. Nguyet Phan, and I. Zelinka, “Periodic Time Series Forecasting with Bidirectional Long Short-Term Memory: Periodic Time Series Forecasting with Bidirectional LSTM,” *ACM Int. Conf. Proceeding Ser.*, pp. 60–64, 2021, doi: 10.1145/3453800.3453812.

[3] V. Kotu and B. Deshpande, “Chapter 12 - Time Series Forecasting,” V. Kotu and B. B. T.-D. S. (Second E. Deshpande, Eds. Morgan Kaufmann, 2019, pp. 395–445. doi: https://doi.org/10.1016/B978-0-12-814761-0.00012-5.

[4] 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, Feb. 2021, doi: 10.1016/J.KNOSYS.2020.106467.

[5] 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.

[6] “Prophet: forecasting at scale.” https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/ (accessed Oct. 24, 2020).

[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] 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.

[9] 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.

[10] “Web Traffic Time Series Forecasting,” 2017. https://www.kaggle.com/c/web-traffic-time-series-forecasting/discussion/39367 (accessed Oct. 12, 2022).

[11] “Choosing the correct error metric: MAPE vs. sMAPE,” 2020. https://towardsdatascience.com/choosing-the-correct-error-metric-mape-vs-smape-5328dec53fac (accessed Oct. 12, 2022).

[12] C. Kuranga and N. Pillay, “A comparative study of nonlinear regression and autoregressive techniques in hybrid with particle swarm optimization for time-series forecasting,” *Expert Syst. Appl.*, vol. 190, Mar. 2022, doi: 10.1016/j.eswa.2021.116163.

[13] 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.

[14] R. Godahewa, C. Bergmeir, G. I. Webb, and P. Montero-Manso, “An accurate and fully-automated ensemble model for weekly time series forecasting,” *Int. J. Forecast.*, 2022, doi: 10.1016/J.IJFORECAST.2022.01.008.

[15] D. Deng, F. Karl, F. Hutter, B. Bischl, and M. Lindauer, “Efficient Automated Deep Learning for Time Series Forecasting,” 2022, Accessed: Oct. 04, 2022. [Online]. Available: https://github.com/automl/Auto-PyTorch.

[16] S. Xu, C. Han, and C. Ran, “A Time Series Combined Forecasting Model Based on Prophet-LGBM,” *ACM Int. Conf. Proceeding Ser.*, May 2021, doi: 10.1145/3469213.3470280.

[17] “The M3-Competition Database.” https://forecasters.org/resources/time-series-data/m3-competition/ (accessed Oct. 18, 2022).

[18] 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

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

[20] 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.

[21] Python Software Foundation, “Python 3.12.1 documentation.” https://docs.python.org/3/ (accessed Feb. 02, 2024).

[22] Google, “Plataforma de archivos compartidos y almacenamiento personal en la nube - Google.” https://www.google.com/intl/es-419\_ar/drive/ (accessed Feb. 02, 2024).

[23] Google, “Ayuda de Google.” https://support.google.com/ (accessed Feb. 02, 2024).

[24] Google, “Google Colab.” https://research.google.com/colaboratory (accessed Feb. 02, 2024).