**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](#_Toc159344053)

[Introducción 1](#_Toc159344054)

[1.1. Motivación 2](#_Toc159344055)

[1.2. Objetivos 2](#_Toc159344056)

[1.2.1. Objetivo General: 2](#_Toc159344057)

[1.2.2. Objetivos Específicos: 2](#_Toc159344058)

[1.3. Estructura del documento 2](#_Toc159344059)

[Capítulo 2 3](#_Toc159344060)

[Marco Teórico 3](#_Toc159344061)

[2.1. Tráfico web 3](#_Toc159344062)

[2.2. Series de tiempo 3](#_Toc159344063)

[2.1. Pronóstico de series de tiempo de tráfico web 3](#_Toc159344064)

[2.2. Evaluación de desempeño de los modelos predictivos 5](#_Toc159344065)

[Bibliografía 6](#_Toc159344066)

# 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

El trafico web es generado por los usuarios de una página web, el trafico web son los datos que se envían y se reciben correspondientes a los usuarios que visitan la página web[1].

## Series de tiempo

Las series de tiempo son un conjunto de valores medidos en orden secuencial en el tiempo[2]; cuando se miden los datos para generar una serie de tiempo, generalmente se toman los valores con la misma separación en tiempo entre cada valor[3].

## Pronóstico de series de tiempo de tráfico web

El pronóstico de series de tiempo se basa en las observaciones pasadas de la serie de tiempo a pronosticar y otras entradas, siendo el proceso de predecir valores futuros de una serie de tiempo [3].

Entre las formas del pronóstico de series de tiempo de tráfico web los autores en[4] proponen de redes *Generative Adversarial model* (GAN)[4] con *Long Short Term Memory* (LSTM)[5] y un perceptrón multicapa (MLP)[6], donde LSTM y GAN actuarían como generador y el MLP como discriminador, para generar series de tiempo dado el conjunto de datos real, finalmente se realizaría el pronóstico con la librería *Prophet*[7] comparando la combinación de tecnologías anteriores con métodos estadísticos, finalmente llegando a la conclusión de que los autores no obtuvieron una diferencia notable[4].

Entre las formas del pronóstico de series de tiempo de tráfico web existe proponen[8] la técnica de redes neuronales *Long Short Term Memory* (LSTM)[5] con entrenamiento asíncrono distribuido, cuya métrica de desempeño que se utilizó según el autor en[8] es la MAE (*mean absolute error*)[9] y la función de pérdida de Huber[10] para probar la precisión del modelo, han logrado un buen grado de asertividad con 200 épocas, en el documento científico[8] menciona obtuvieron una MAE en promedio menor que 30, lo que consideran un buen resultado[8].

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*)[11] ya que realiza la tarea con mayor velocidad en comparación con los tensores regulares, el *decoder* es TensorFlow GRUBlockCell[11], en este documento científico[1] menciona que realizaron algunos cambios al modelo ganador de *Kaggle*[11] como agregar la mediana de 7,30,90 y 180 días, usaron para medir el desempeño del modelo la métrica *Symmetric Mean Absolute Percentage Error* (SMAPE)[12] donde comentan los autores que obtuvieron un SMAPE de 0.349[1].

En otro artículo[13], proponen utilizar un enfoque híbrido de *particle swarm optimization* y *quantum-inspired* resultando en QPSO[13] y redes *Nonlinear autoregressive exogenous* NARX[14] resultando en NARX -QPSO[13] en dicho artículo se menciona que el modelo NARX –QPSO tiene un rendimiento sobresaliente en los resultados comparados con otros modelos[13].

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

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

*Prophet* y *light gradient boosting machine* (LGBM)[16] donde se menciona que tiene mejores resultados comparado con modelos individuales[16].

Otra tecnología son las redes bidireccionales LSTM (BI-LSTM)[2] las cuales en este artículo[2] fueron probadas en el conjunto de datos *M3-Competition*[17]el cual se usa para probar modelos de pronóstico de series de tiempo dado que posee distintas categorías, en el artículo[2] mencionan que probaron las redes BI-LSTM sobre el conjunto de datos 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): el error cuadrático medio sirve para evaluar las diferencias entre los valores predichos por un modelo y los valores observados.

Donde es el valor predicho y es el valor real[9]

El MAE (*mean absolute error*)[9] es una métrica que sirve para medir el error medio absoluto.

Donde es el valor predicho y es el valor real[9],

La métrica *Symmetric Mean Absolute Percentage Error* (SMAPE)[12] básicamente es el error medio absoluto porcentual simétrico.

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

# 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.

[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, p. 106467, 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] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] “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).

[13] 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.

[14] 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.

[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.