**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](#_Toc164369620)

[Introducción 1](#_Toc164369621)

[1.1. Motivación 2](#_Toc164369622)

[1.2. Objetivos 2](#_Toc164369623)

[1.2.1. Objetivo General: 2](#_Toc164369624)

[1.2.2. Objetivos Específicos: 2](#_Toc164369625)

[1.3. Estructura del documento 2](#_Toc164369626)

[Capítulo 2 3](#_Toc164369627)

[Marco Teórico 3](#_Toc164369628)

[2.1. Tráfico web 4](#_Toc164369629)

[2.2. Series de tiempo 4](#_Toc164369630)

[2.1. Pronóstico de series de tiempo de tráfico web 4](#_Toc164369631)

[2.2. Evaluación de desempeño de los modelos predictivos 6](#_Toc164369632)

[Capítulo 3 7](#_Toc164369633)

[Descripción del problema 7](#_Toc164369634)

[3.1 El problema 8](#_Toc164369635)

[3.2 Estructura Organizacional 8](#_Toc164369636)

[Organigrama 8](#_Toc164369637)

[Capítulo 4 10](#_Toc164369638)

[Solución Propuesta 10](#_Toc164369639)

[4.1 Materiales y Métodos 11](#_Toc164369640)

[4.2 Herramientas 13](#_Toc164369641)

[4.2.1 Python 13](#_Toc164369642)

[4.2.2 Google Drive 13](#_Toc164369643)

[4.2.3 Universal Analytics (UA) 13](#_Toc164369644)

[4.2.4 Google Analytics 4 (GA4) 13](#_Toc164369645)

[4.2.5 Looker Studio 13](#_Toc164369646)

[4.2.6 Microsoft Word 13](#_Toc164369647)

[4.2.7 Microsoft Excel 14](#_Toc164369648)

[4.2.8 Numpy 14](#_Toc164369649)

[4.2.9 Tensorflow 14](#_Toc164369650)

[4.2.10 Keras 14](#_Toc164369651)

[4.2.11 Pandas 14](#_Toc164369652)

[4.2.12 Matplotlib 14](#_Toc164369653)

[4.2.13 Google Colaboratory 14](#_Toc164369654)

[4.2.14 Git 14](#_Toc164369655)

[4.2.15 Hiperparametros (*Hyperparameters*) 15](#_Toc164369656)

[4.2.16 Época (*Epoch)* 15](#_Toc164369657)

[4.2.17 Trial 15](#_Toc164369658)

[4.2.18 Unidad Recurrente con Puerta – GRU (*Gate Recurrent Unit*) 15](#_Toc164369659)

[4.2.19 Arquitectura Codificador-Decodificador (*Encoder-Decoder*) 17](#_Toc164369660)

[4.2.20 Algoritmo Optimización Bayesiana (*Bayesian Otimization*) 17](#_Toc164369661)

[4.2.21 Algoritmo Hyperband 17](#_Toc164369662)

[4.2.22 Algoritmo Búsqueda Aleatoria (*Random Search*) 17](#_Toc164369663)

[4.2.23 Detención Temprana (*Early Stopping)* 18](#_Toc164369664)

[4.2.24 Github 18](#_Toc164369665)

[4.3 Comprensión de los datos 19](#_Toc164369666)

[4.3.1 Recolección de datos iniciales 19](#_Toc164369667)

[4.3.2 Preparación de los datos 20](#_Toc164369668)

[4.3.3 Predictor Flag 22](#_Toc164369669)

[4.3.4 Detección de Anomalías 23](#_Toc164369670)

[4 Modelos 24](#_Toc164369671)

[3.5 Selección de Modelos 24](#_Toc164369672)

[Desarrollo 26](#_Toc164369673)

[Pruebas Editorial Universitaria 33](#_Toc164369674)

[Conversión de Modelos de UA a GA4 de Tum Transmedia 34](#_Toc164369675)

[Conversión de Modelos de UA a GA4 de Editorial Universitaria 34](#_Toc164369676)

[Capítulo 5 37](#_Toc164369677)

[Conclusión 37](#_Toc164369678)

[5.1 Conclusiones 38](#_Toc164369679)

[5.2 Futuras líneas de investigación 38](#_Toc164369680)

[Anexo 40](#_Toc164369681)

[Bibliografía 42](#_Toc164369682)

# 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 redes Modelo Generativo Adversario - GAN(*Generative Adversarial model*)[4] con reders de Memoria a Corto y Largo Plazo – LSTM (*Long Short Term Memory*)[5] y un perceptrón multicapa - MLP (*Multilayer Perceptron*)[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 que existen, los autores en [8] proponen la técnica de redes neuronales LSTM[5] con entrenamiento asíncrono distribuido, cuya métrica de desempeño que se utilizó según el autor en[8] es el Error Medio Absoluto - 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] mencionan que 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 Redes Neuronales Recurrentes -RNN (*Recurrent Neural Networks*) seq2seq[1] con la ayuda de la arquitectura Codificador –Decodificador (*encoder/decoder)*[11], el *encoder* es cuDNN-GRU

Es decir una Unidad Recurrente con Puerta (*Gate Recurrent Unit*)[12] con backend de CUDA, ya que realiza la tarea con mayor velocidad en comparación con los tensores regulares, el *decoder* es TensorFlow GRUBlockCell[12], en este documento científico[1] menciona que realizaron algunos cambios al modelo ganador de *Kaggle*[12] como agregar la mediana de 7,30,90 y 180 días, usaron para medir el desempeño del modelo la métrica Error Absoluto Medio Porcentual Simétrico -SMAPE (*Symmetric Mean Absolute Percentage Error*)[13] donde comentan los autores que obtuvieron un SMAPE de 0.349[1].

En otro artículo[14], proponen utilizar un enfoque híbrido de optimización de enjambre de partículas (*particle swarm optimization)* e inspiración cuántica (*quantum-inspired)* resultando en QPSO[14] y redes Exogenas Autorregresivas No Lineales – NARX(*Nonlinear autoregressive exogenous*)[15] resultando en NARX -QPSO[14] en dicho artículo se menciona que el modelo NARX –QPSO tiene un rendimiento sobresaliente en los resultados comparados con otros modelos[14].

En otro artículo[16] propone un diseño de un sistema Aprendizaje Automático Automatizado - *Automated Machine Learning* (Auto ML) [16], una arquitectura neuronal nueva denominada Auto-PyTorch-TS[16] donde lo comparan con otros modelos y demuestran que su modelo tiene un mejor rendimiento[16].

También se propone en el artículo[17] el uso de un modelo combinado

*Prophet* y Maquina de Empuje de Gradiente de Luz *-* LGBM(*light gradient boosting machine*)[17] donde se menciona que tiene mejores resultados comparado con modelos individuales[17].

Otra tecnología son las redes bidireccionales LSTM (BI-LSTM)[2] las cuales en este artículo[2] fueron probadas en el conjunto de datos (*dataframe*) *M3-Competition*[18]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 Media móvil integrada autorregresiva ARIMA (*Auto-Regressive Integrated Moving Average*)[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[19]

La Distancia media cuadrática mínima - RMSE(*Root Mean Square Errors*): el error cuadrático medio es una medida de las diferencias entre los valores predichos por un modelo y los valores observados.

(1)

En la formula (1) es el valor predicho y es el valor real[9]

El error medio absoluto - MAE(*mean absolute error*)[9] se calcula de la siguiente forma:

(2)

En la formula (2) es el valor predicho y es el valor real[9],

La métrica Error Absoluto Medio Porcentual Simétrico - SMAPE (*Symmetric Mean Absolute Percentage Error*)[13] se calcula de la siguiente forma:

(3)

En la formula (3) 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 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



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 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 voz sobre protocolo de Internet - VoIP, redes WiFi, gestión de las Redes de Área Local Virtual - VLAN (Virtual Local Area Network) 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 Centro de Procesamiento de Datos - CPD, plataformas de respaldo de datos, sistemas de monitoreo de las plataformas y servicios esenciales de la infraestructura de la universidad como Sistema de Nombres de Dominio – DNS (agregar), Protocolo de Puerta de Enlace Fronteriza – BGP, Camino Más Corto Primero – OSPF (*Open Shortest Path First*), Red Privada Virtual – VPN (Virtual Private Network), etc.
* **Poner en ingles esp-**abreviacio-(ingles)

# 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 utilizó la metodología CRISP-DM[20] como guía, ya que hoy en día es un estándar de facto para los proyectos de ciencia de datos y minado de datos[21].

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

* Comprensión del negocio:

En esta fase inicial se comprenden los objetivos y los requisitos para cumplir el proyecto de Pronóstico de Series de Tiempo de Tráfico web, donde también se realizaron entrevistas a los trabajadores del área de Departamento de Gestión de Recursos de Redes y Comunicaciones.

* Comprensión de los datos:

En esta fase se comenzó a ver como extraer los datos de *Google analytics*, en que formato se pueden descargar los datos de la nube, de los datos que están en la nube sobre tráfico web correspondientes a las páginas de la universidad entender como son recolectados esos datos y que significa cada uno.

* Preparación de los datos:

En esta fase se procedió a descargar los datos de *Google analytics,* esto a través de la herramienta de Looker Studio, también se limpiaron los datos nulos, y se hizo una eliminación de los datos anómalos del *dataset*.

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[22].

* Modelado de los datos:

En esta fase se seleccionó la arquitectura *encoder/decoder* y la red neuronal GRU a utilizar, como así también se probaron distintos algoritmos que afinaron los modelos a valores óptimos.

En esta fase, se seleccionan y aplican varias técnicas de modelado, y sus parámetros se calibran a valores óptimos[22].

* Evaluación:

En esta fase se diseñaron y realizaron distintas pruebas a los modelos, de ser necesario se volvía a la fase de modelado a ejecutar nuevamente los algoritmos de calibración de los modelos, finalmente los modelos obtenidos con los datos de UA eran probados y reentrenados para ser usados con los datos de GA4.

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][21].

* Despliegue:

Los últimos modelos obtenidos son funcionales con los datos de GA4, es decir, se tienen modelos que funcionan con los datos del entorno real de más actual, dado que los datos de UA ya no serán recolectados por *Google analytics*.

En esta última fase se realizan las conclusiones del proyecto de Pronóstico de Series de Tiempo de Tráfico web y se analizan las futuras líneas de investigación.

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

## Herramientas

### 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[23], 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[24].

### 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[25].

### 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[26].

### 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*[26], también se pueden exportar datos de tablas de informes de Looker Studio en archivos CSV.

### Microsoft Word

Microsoft Word un procesador de texto, este software es de Microsoft, sirve para crear, editar, leer, escribir documentos de texto[27].

### Microsoft Excel

Microsoft Excel un software para poder crear, leer y editar planillas de cálculos, este software es de Microsoft[27].

### Numpy

Según su documentación oficial es una librería*,* es un estándar para trabajar con datos numéricos en Python, la librería posee estructuras de datos como Arrays, además posee funciones para realizar cálculos matemáticos con dichos Arrays [28].

### Tensorflow

Es una biblioteca según su documentación oficial es de código abierto para el aprendizaje automático[29].

### Keras

Es una biblioteca que sirve de Interfaz de Programación de Aplicaciones-API  (*Aplication Programing Interface*) de alto nivel de Tensorflow, actualmente Keras se encuentra dentro de la biblioteca Tensorflow y es una API de alto nivel oficial de Tensorflow[29].

### Pandas

Es una librería de Python para el análisis y manipulación de datos, posee el objeto dataframe, también funciones para manipular los datos del dataframe y funciones para leer y escribir archivos CSV (valores separados por coma)[30].

### Matplotlib

Matplotlib según su página oficial es una librería para la creación de gráficas, animaciones, distintas formas de visualización de datos en Python[31].

### 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 de esta manera tener acceso y poder leer y escribir los archivos de Google Drive[32].

### Git

Git es un sistema de control de versiones, sirve para controlar las versiones de un proyecto, un proyecto es un repositorio de Git, un proyecto comprende todo el conjunto de carpetas y archivos al que se le realiza el control de versiones, de manera que se guardan los cambios y también se pueden recuperar las versiones anteriores del proyecto[33].

### Hiperparametros (*Hyperparameters*)

Son las variables que se utilizaron durante el entrenamiento del modelo, como así también pueden definir la topología del modelo, estas variables tienen un impacto en el rendimiento del modelo[34].

### Época (*Epoch)*

Según la documentación oficial de *Keras* una época *(Epoch)*, es cuando el mdoelo pasa por todo el conjunto de datos[35], en *Keras*, se puede agregar *callbacks* al final de una época para guardar o imprimir información, como por ejemplo guardar el modelo, de forma que si se ejecutan x cantidad de épocas, se pueda recuperar el modelo con mejor rendimiento de la mejor época.

### Trial

Según su documentación oficial, un trial es una prueba, donde cada trial tiene x cantidad de épocas, en un trial se evalúa un conjunto de valores de hyperparámetros, es decir, en cada trial se evalúa una configuración diferente del modelo[36].

### Unidad Recurrente con Puerta – GRU (*Gate Recurrent Unit*)



Figura 2: Red Neuronal GRU (2).

Donde r es la compuerta de reinicio (*reset gate)* y z es la compuerta de actualización (*update gate*) y h’ es el candidato a activación y h es la activación[11].

La siguiente formula (4) es para cada j-esimo valor de la compuerta de reset.

(4)

En la formula (4) es la función sigmoidea, W y U son matrices que aprenden, es el estado oculto anterior y x es la entrada

La siguiente formula (5) es para cada j-esimo valor de la compuerta de actualización.

(5)

En la formula (5) es la función sigmoidea W y U son matrices que aprenden, es el estado oculto anterior y x es la entrada

La siguiente formula (6) es para cada j-esimo valor de activación, o también llamado estado oculto.

(6)

En la formula (6) es el candidato a activación y se define en la siguiente fórmula para cada j-esimo valor.

(7)

En la formula (7) es la función de tangente hiperbólica y es el producto de Hadamard, W y U son matrices que aprenden, es el estado oculto anterior y x es la entrada[11].

### Arquitectura Codificador-Decodificador (*Encoder-Decoder*)

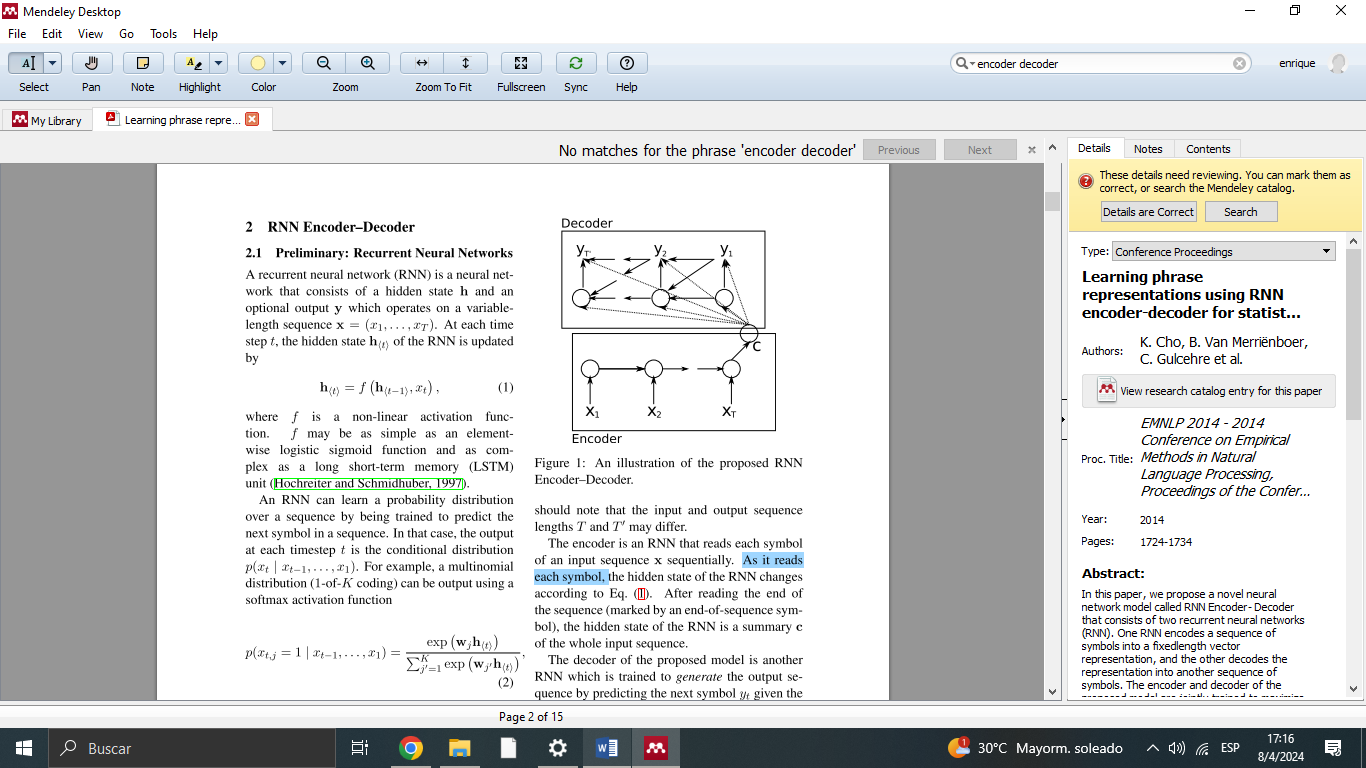


Figura 3: Arquitectura *encoder-decoder* GRU (3).

La arquitectura Codificador-Decodificador consta de un codificador, siendo el codificador una capa de GRU que lee los valores de la secuencia de entrada en orden, y modifica su estado oculto con cada nuevo valor, por otro lado el decodificador también es una capa de GRU que toma como entrada un vector c de los estados ocultos del codificador, de esta manera el decodificador tendría que aprender a predecir el siguiente valor de la secuencia[11].

### Algoritmo Hyperband

Es un algoritmo que se usa para el afinamiento de los hiperparametros, este algoritmo ejecuta el algoritmo de *Successive Halving*[37]variando como entrada n y r, siendo n el número de configuraciones y r el número de recursos, como ser numero de epocas, para de esta manera encontrar la mejor configuración de hiperparametros[37], a diferencia de la búsqueda aleatoria y la Optimizacion bayesiana el algoritmo *Hyperband* no tiene como dato de entrada un número máximo de trials, solo los valores de R (épocas) y n el cual por defecto es 3.

### Detención Temprana (*Early Stopping)*

La detención temprana es una configuración que sirve para cortar el entrenamiento de manera temprana[38], esto es, se configura de manera que si el rendimiento no mejora en un número x de épocas se detiene el trial. Por ejemplo si se configura un modelo para que entrene 150 épocas pero la detención temprana está configurada en 20, y al entrenar el modelo no mejora el rendimiento desde la época 40 a la 60, este se detiene y no continúa el entrenamiento.

En este documento científico[39] se menciona el uso de una detención temprana de 10, además en este otro documento científico[38] se menciona el uso la detención temprana para evitar el sobre entrenamiento.

### Github

Github según su página oficial es una plataforma que hospeda repositorios de Git, ofreciendo además el servicio de que los equipos de desarrollo puedan trabajar en grupo[33].

## 4.3 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[22].

### 4.3.1 Recolección de datos iniciales

Se tienen acceso a datos históricos de las siguientes 3 páginas webs:

* 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 las cuales tenemos mayor cantidad de datos de registrados por UA y menor cantidad de datos expresados en días registrados en GA4, las siguientes tablas explican los datos correspondientes por fecha que se tienen de dichas páginas web.

|  |  |  |
| --- | --- | --- |
| Página Web | Desde | Hasta |
| <https://unam.edu.ar/> | 28-05-2018 | 04-07-2023 |
| <https://editorial.unam.edu.ar/> | 08-10-2018 | 04-10-2023 |
| <https://transmedia.unam.edu.ar/> | 22-08-2018 | 01-06-2023 |

Tabla 1 de datos de *Universal Analytics*

|  |  |  |
| --- | --- | --- |
| Página Web | Desde | Hasta |
| <https://unam.edu.ar/> | 14-06-2022 | 23-11-2023 |
| <https://editorial.unam.edu.ar/> | 14-06-2022 | 24-01-2024 |
| <https://transmedia.unam.edu.ar/> | 14-06-2022 | 01-01-2024 |

Tabla 2 de datos de *Google Analytics* 4

Para poder descargar los datos se procedió a conectar *Google Analytics* como fuente de datos de *Looker Studio*, de esta manera se pueden pasar los datos a tablas en *Looker Studio* y posteriormente descargar los datos en archivos CSV.

### 4.3.2 Preparación de los datos

Identificando algunos problemas de calidad de los datos la página web <https://unam.edu.ar/> desde noviembre del 2023 no está capturando el tráfico correctamente en GA4, por lo que se procede a descartar este conjunto de datos.

Lo siguiente que se procedió a realizar, comenzando por la página web <https://transmedia.unam.edu.ar/> se descargaron los datos en archivos csv del *Looker Studio* de UA, son 5 archivos CSV que se procede a unificar en un solo archivo a través de código en Python.

Se descargaron los siguientes datos en 5 archivos CSV:

Vistas por Sistemas Operativos, Vistas por Agrupación de canales predeterminados, número de vistas de página, Usuarios Nuevos, Usuarios, Numero de Sesiones Por usuario, Sesiones, Vistas por Categoría de dispositivo, Vistas por País

Se procedieron a transformar los datos obtenidos de los archivos CSV de Looker Studio para obtener las columnas del dataframe.

De vistas por sistema operativo salieron las siguientes columnas:

Windows, Android, Linux, Macintosh, Sistema operativo (not set), iOS, Windows Phone, Chrome OS, Tizen, Sistema operativo promedio, Sistema operativo std,

mediana Sistema operativo.

Siendo Sistema operativo (not set) donde se registran las vistas los sistemas operativos que UA no pudo detectar, Sistema operativo promedio es el promedio de las otras columnas de sistemas operativos, Sistema operativo std el desvió estándar de las columnas de Sistema operativo y el resto son vistas de sistemas operativos conocidos, por ejemplo la columna Windows tendrá en cada fila datos de las vistas de los usuarios que accedieron a la página web a través de un dispositivo cuyo sistema operativo es Windows.

De vistas por Agrupación de canales predeterminados salieron las siguientes columnas:

Organic Search, Social, Direct, Referral, Agrupación de canales predeterminada promedio, Agrupación de canales predeterminada std, mediana Agrupación de canales predeterminada

Donde Organic Search, Social, Direct y Referral son los canales predeterminados, es decir, las fuentes de tráficos más comunes[40], Agrupación de canales predeterminada promedio es el promedio de los canales predeterminados, Agrupación de canales predeterminada std es el desvío estándar y mediana Agrupación de canales predeterminada es la mediana de las columnas de agrupación de canales predeterminados.

De vistas por Categoría de dispositivo salieron las siguientes columnas:

Mobile, Desktop, Tablet, Categoría de dispositivo promedio, Categoría de dispositivo std., mediana Dispositivos.

Donde Mobile, Desktop y Tablet son la Categoría de dispositivo, es decir, el tipo de dispositivo que tenía el usuario cuando se registró la visita en la página web[41], además Categoría de dispositivo promedio es el promedio de vistas, Categoría de dispositivo std es el desvió estándar y mediana Dispositivos es la mediana de las columnas de dispositivos.

Las siguientes son tal cual del CSV no se necesitó mayor trabajo de transformación:

Número de vistas de página, Usuarios Nuevos, Usuarios, Número de Sesiones Por usuario, Sesiones.

Es decir una columna de Usuarios Nuevos del CSV se pasó al dataframe como una columna de Usuarios Nuevos sin mayores complicaciones. De vistas por País se procedió a dividir en 2 columnas, las vistas que son de Argentina y las que son de otros países.

Luego, se agregaron las columnas mediana de medianas de acuerdo al documento [1] donde toman las medianas rodantes de 6, 12, 18, 30, 48, 78, 126, 203, 329 días de la columna Número de vistas de página y de ahí se procede a calcular la mediana de medianas, según lo explicado en [1].

También dado que en el código original [42] agregaba valores rezagados, es decir valores anteriores los agregaba como columnas, entonces se agregaron valores rezagados de 3, 6, 9,12 meses como las columnas:

lag\_90, lag\_180, lag\_270, lag\_360

Los nombres de las columnas vienen del inglés “lag” que se utiliza para referirse al rezago, o también se puede traducir como retraso, también un \_ para separar del número y el numero en días correspondiente al més de retraso tomando 1 mes como 30 días aplicado esa regla para hacer las columnas de valores retrasados. A continuación Se procedió a confirmar la existencia de NaN (Not an number), y reemplazarlos por su promedio.

Quedando como resultado del preprocesado las siguientes columnas de la Television Universitaria Misionera - Tum transmedia, que fueron las mismas columnas que se tomaron en cuenta para las otras páginas web de la institución.

Listado de columnas del *Dataframe*:

Número de vistas de página, Usuarios nuevos, Usuarios ,

Número de sesiones por usuario, Sesiones, Argentina,

Otros Países , mobile , desktop , tablet ,

Categoría de dispositivo promedio, Categoría de dispositivo std ,

Windows , Android , Linux , Macintosh ,

Sistema operativo (not set) , iOS , Windows Phone , Chrome OS ,

Tizen , Sistema operativo promedio , Sistema operativo std ,

Organic Search , Social , Direct , Referral ,

Agrupación de canales predeterminada promedio ,

Agrupación de canales predeterminada std , mediana\_de\_medianas ,

lag\_90,lag\_180, lag\_270, lag\_360, mediana Dispositivos ,

mediana Sistema operativo,

mediana Agrupación de canales predeterminada.

Siendo un total de 37 columnas, por supuesto que desde *Google Analytics* se registran más datos, sin embargo debido al cambio de UA a GA4[43] solo se eligieron los datos que a nivel tanto conceptual como de medición, fueran más parecidos entre sí.

### 

### 4.3.4 Detección de Anomalías

Finalmente se realizó una limpieza de los datos anómalos (*outliers)* detectados usando los k vecinos más cercanos - KNN (*K Nearest Neighbors*), donde se tomaron los 7 vecinos más cercanos, se consideran más probable que sean *outliers* aquellos datos del *dataset* que estén más alejados de sus vecinos más cercanos[44].

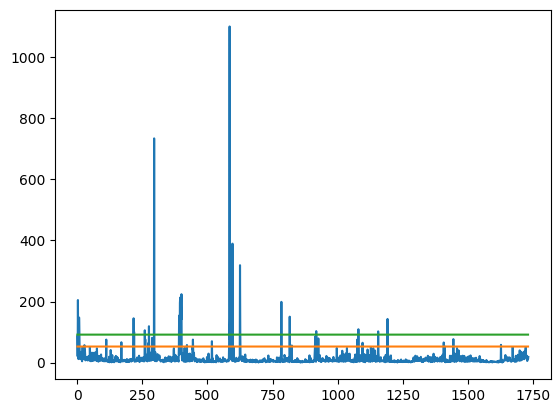


Figura 4: Distancia Promedio de los 7 Vecinos más cercanos para cada fila del dataframe de Tum Transmedia.

En la figura anterior muestra la distancia Promedio de los 7 vecinos más cercanos para cada fila del conjunto de datos, de esta manera podemos considerar los que se encuentren más lejanos de sus 7 vecinos más cercanos como un *outlier*, en este caso se quitaron las filas del conjunto de datos que estén por encima de 500, finalizando así el proceso de limpieza de los datos de Tum Transmedia. En el paper [45] que trata sobre outliers en series temporales se menciona que a los outliers normalmente se los elimina o se los reemplaza con algún valor, en este caso se decidió eliminarlos del conjunto de datos.

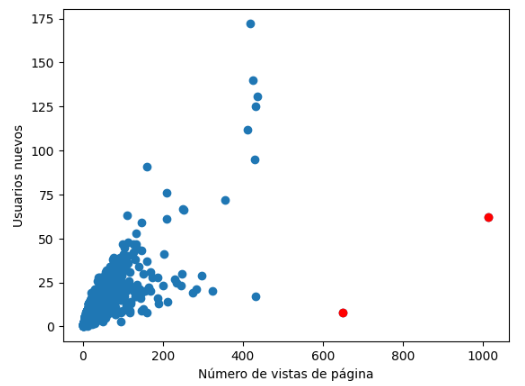


Figura 5: Grafico de dispersión de Número de vistas de página y Usuarios nuevos del dataframe de Tum Transmedia.

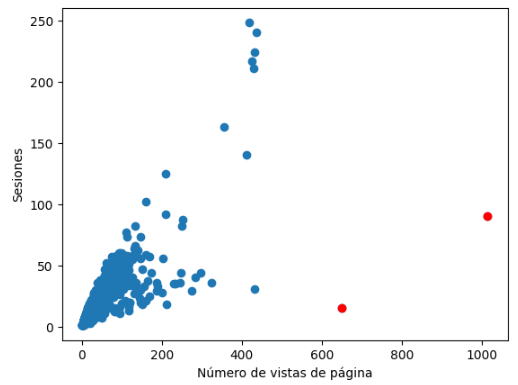


Figura 6: Grafico de dispersión de Número de vistas de página y Sesiones del dataframe de Tum Transmedia.

En los gráficos de dispersión mostrados se puede ver que en rojo aparecen los outliers de manera que considero que el valor probado con los 7 vecinos más cercanos para el knn es el correcto.

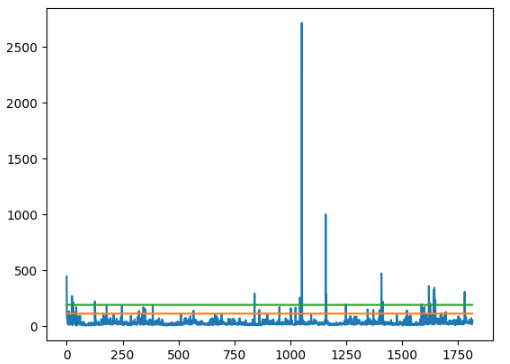


Figura 8: Distancia Promedio de los 7 Vecinos más cercanos para cada fila del dataframe de Editorial Universitaria.

Se tomaron como Outliers de Editorial Universitaria a partir de una distancia 500 mayor que 500 de media de sus 7 vecinos más cercanos.

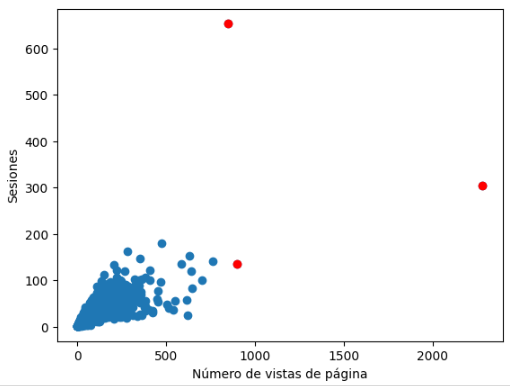


Figura 6: Grafico de dispersión de Número de vistas de página y Sesiones del dataframe de Editorial Universitaria.

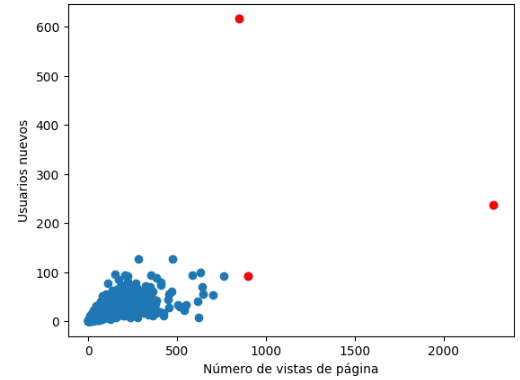


Figura 6: Grafico de dispersión de Número de vistas de página y Usuarios nuevos del dataframe de Editorial Universitaria.

### 4.3.5 Partición de Los datos

Los datos se parten en conjuntos de entrenamiento y prueba, al igual que en [8] partimos los datos en 80-20, es decir, donde el 80% de los datos se utilizaran para entrenar el modelo, y el 20% para probar el rendimiento del mismo.

En el caso de los datos de GA4 dado hay una menor cantidad de datos y necesitamos al menos 120 datos en el conjunto de prueba se parte en 70-30, 70% para entrenamiento y 30% para prueba. A su vez los datos se parten en muestras de a 120, 60 días de datos de entrada del modelo y 60 días de datos que se comparan con la salida del modelo. Las muestras de 60 días de entrada usados en entrenamiento los llamaremos X\_train mientras que para prueba las llamaremos X\_test. Las muestras de 60 días de salida usados en entrenamiento las llamaremos Y\_train mientras que para prueba las llamaremos Y\_test.

### 4.3.6 Normalización

Para poder ingresar los datos a la red neuronal se normalizaron los datos pasándolos primero por log1p es decir hacer log(1+x) donde x es el dato y luego normalización z-score (citar) esto tanto para entrenamiento y prueba de acuerdo al código original (citar), otro tipo de normalización que se aplico fue solo z-score esto para las pruebas en los que el modelo tuviera Huber Loss, en el plan de pruebas se explica con más detalle.

## **4.4 Modelos**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tecnologías** | **Red N.** | **Resultados  según Artículos científicos** | **Finalidad** | **Aplica** |
| MLP | SI | - | Clasificación, predicción | No |
| 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 |
| GRU | SI | El modelo ganador de Kaggle utiliza un GRU que es más  Rápido que un GRU normal. | Sistemas predictivos basados en información secuencial. | No |
| *Bi-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 3 Comparación de tecnologías

## **4.4.1 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.

Porque el modelo no es lo suficientemente complejo, y tampoco hay tantos datos.

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, eligió la métrica MAE que se utilizó en [8], también la métrica SMAPE debido a que es la misma con la que se evaluó el modelo RNN seq2seq, que es el ganador del concurso de Kaggle[12].

## 4.5 Desarrollo

### 4.5.1 Plan de pruebas

El plan de prueba consiste generar varios modelos GRU con la arquitectura seq2seq, y se procederá a medir usando la métricas MAE,RMSE y SMAPE, en primera instancia se utilizarán los datos de Tum transmedia de UA, posteriormente se utilizará el mejor modelo resultante en los datos de Tum transmedia de GA4, también en los datos de editorial universitaria UA y el mejor modelo resultante de editorial universitaria UA se utilizará para generar un modelo y probarlo con los datos de editorial universitaria GA4.

La función de perdida de los modelos que se utilizaron fueron huber loss (citar) utilizado en este documento científico [8] y el MAE, esto a nivel entrenamiento y prueba, ya una vez los modelos entrenados se midió el rendimiento usando la métrica MAE,RMSE y SMAPE sobre los datos desmoralizados.

Un SMAPE aceptable de acuerdo a [1] es alrededor o por debajo de 0.350, mientras que un MAE aceptable de acuerdo a [8] es alrededor o por debajo de 30.

Se entrenaron modelos con los datos de TUM UA normalizados usando solo z-score con función de perdida huber loss y se entrenó otro modelo con el conjunto de datos normalizado usando log1p y z-score con función de perdida mae.

Agregar algo de RMSE

### 4.5.2 Construcción de los Modelos

En esta sección se detallan las características de los modelos generados, básicamente en primera instancia se generaron modelos utilizando el algoritmo de ajuste de hiperparametros, siendo el algoritmo de ajuste de hiperparametros Hyperband, En las pruebas realizadas se busca obtener los mejores valores de los parámetros e hiperparametros del modelo.

### 4.5.3 Cantidad de Trials por Modelo

El algoritmo hyperband se configuró con un factor de 3 (el cual es el valor por defecto) y con objetivo de 200 épocas ya que es el número utilizado en este paper (citar), se utilizó un early stopping de 10 que se aumentó dicho valor en las pruebas finales o directamente se lo retiró.

### 4.5.4 Hiperparametros que se Afinaron

|  |  |  |
| --- | --- | --- |
| Nombre | Valores posibles | Definición |
| Unidades | 128-512 con un paso de 32 | Es la cantidad de unidades de GRU por capa. |
| Ratio de Aprendizaje | Entre 1e-4 y 1e-2 | Es el ratio de aprendizaje, *Learning Rate* en ingles  Se utiliza un Optimizador Adam. |
| Batch size | 32 a 128 con un paso de 32 | Tamaño de lote a tomar, normalmente, esto es mara dividir la cantidad de muestras a tomar, ya que no se puede pasar todo el conjunto de datos entero de una sola vez. |
| Numero de Capas | 1-2 | Es el número de capas de la red neuronal |

Tabla 4 Hiperparametros que se afinaron.

Un concepto muy importante que no se podía afinar con la librería Keras es el N-Días el cual es básicamente la cantidad de días hacia atrás que toma como datos de entrada la red neuronal, ya que este valor no se puede afinar por la librería Keras, se toma como entrada por 60 días ya que según lo mencionado por el ganador del concurso de Kaggle (citar) es un numero con el cual se puede obtener resultados aceptables.

Otra variable muy importante que se podía afinar con la librería Keras, pero se hizo manualmente fue el número de capas, dado que con los algoritmos de afinamiento de hiperparametros generaría cientos de modelos por cada algoritmo y que al aumentar el número de capas también aumenta el tamaño de almacenamiento requerido para guardar los modelos, por lo que rápidamente me quedaría sin recursos suficientes de Google Colab y Google Drive, debido a esa limitante se afinó manualmente el número de capas.

A través del siguiente código se configuró la semilla en keras para poder hacer pruebas replicables:

seed\_value = 42

import random

np.random.seed(seed\_value)

random.seed(seed\_value)

tf.random.set\_seed(seed\_value)

La primera prueba que se realizó fue la de correr el algoritmo hyperband con los datos normalizados mediante log1p + z-score, con función de perdida MAE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Unidades | capas | lr | batch\_size |
| 1 | 224 | 1 | 0.0012 | 64 |

Tabla 5 Prueba de Algoritmo de Afinamiento de Hiperparametros.

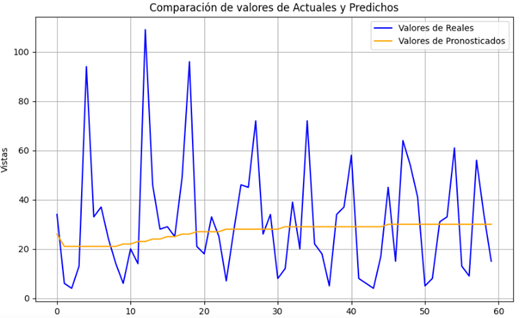
****

Figura 7: Grafico de Los Valores Reales y los Valores Pronosticados de una parte de conjunto de prueba.

La segunda prueba que se realizó fue la de correr el algoritmo hyperband con los datos normalizados mediante z-score, con función de perdida huber loss.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Unidades | Capas | lr | batch\_size |
| 2 | 192 | 1 | 0.0018 | 128 |

Tabla 6 Prueba de Algoritmo de Afinamiento de Hiperparametros.

Comparaciones de rendimientos las pruebas 1 y 2

|  |  |  |  |
| --- | --- | --- | --- |
| # | MAE | SMAPE | RMSE |
| 1 | 11.126461988304094 | 0.5977127357014901 | 16.491247900324023 |
| 2 | 13.274267093758834 | 0.6972292246352332 | 17.523084885859184 |

Tabla 7 Comparaciones de Rendimientos entre Pruebas.

En la tabla 7 se puede ver que la prueba 1 tiene mejores rendimientos en todas las métricas, por lo tanto en pruebas siguientes se seguirán utilizando los parámetros de la prueba 1.

**Prueba de capas**

Esta prueba consiste en agregar más capas al mejor modelo obtenido, se entrnó por 500 épocas con 2 capas de *encoder* y 2 capas de *decoder*.

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo | MAE | SMAPE | RMSE |
| 1 Capa | 11.126461988304094 | 0.5977127357014901 | 16.491247900324023 |
| 2 Capas | 12.152266081871344 | 0.6348007961755141 | 17.790148975902234 |

Tabla 8 Comparaciones de Rendimientos entre Pruebas.

Como se puede ver en la tabla 8 el primer modelo sigue siendo el mejor modelo, En el código de github del ganador del concurso de kaggle (citar), se hacen pruebas de 1 y 2 capas como máximo, por lo tanto no continuamos con las pruebas de capas, siendo la prueba de 2 capas la prueba tercera prueba.

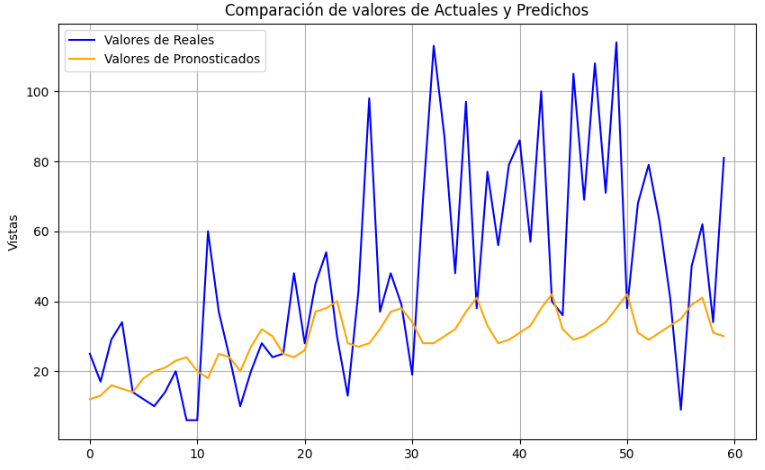


Figura 8: Grafico de Los Valores Reales y los Valores Pronosticados de una parte de conjunto de prueba Y\_test, de la prueba 3.

### 4.5.5 Pruebas Editorial Universitaria

Concluidas las pruebas en Tum Transmedia se seleccionó el modelo con mejor rendimiento y se pasaron a probar con los datos la página web de Editorial U.Na.M.

Para utilizar el mejor modelo de TUM en Editorial UA se construyó el dataset con las mismas columnas que Tum Transmedia, y se utilizó ese dataset para pruebas posteriores de Editorial Universitaria UA.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Modelo | MAE | SMAPE | RMSE |
| 4 | El mejor modelo de TUM UA | 258.66926229 | 1.1840436456 | 364.641053760 |

Tabla 9 Prueba 4

En la tabla 9 se ve el rendimiento del mejor modelo de Tum Transmedia UA aplicado y reentrenado por 500 epocas con los datos de Editorial Universitaria UA.

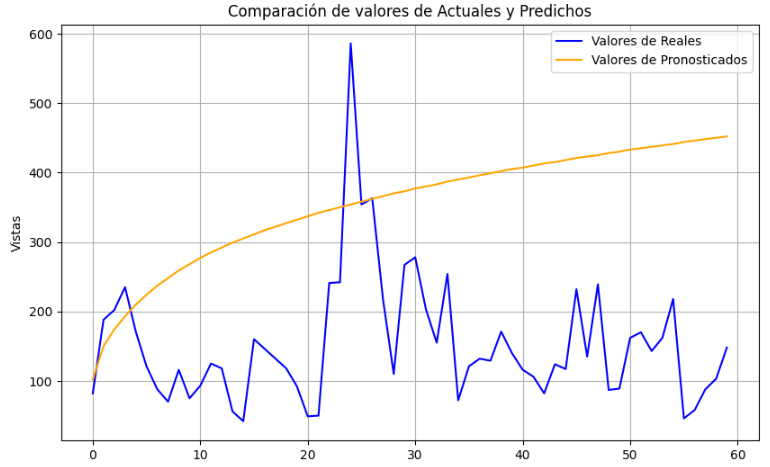


Figura 9: Grafico de Los Valores Reales y los Valores Pronosticados de una parte de conjunto de prueba Y\_test, de la prueba 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Unidades | Capas | lr | batch\_size |
| 5 | 160 | 1 | 0.0001338028856513705 | 192 |

Tabla 10 Prueba 5 de Editorial Universitaria de UA

Dado que la prueba 4 no tuvo un MAE por debajo de 30, decidí hacer una búsqueda de hiperparametros usando el algoritmo hyperband, en la tabla 10 se puede ver los mejores parámetros obtenidos.

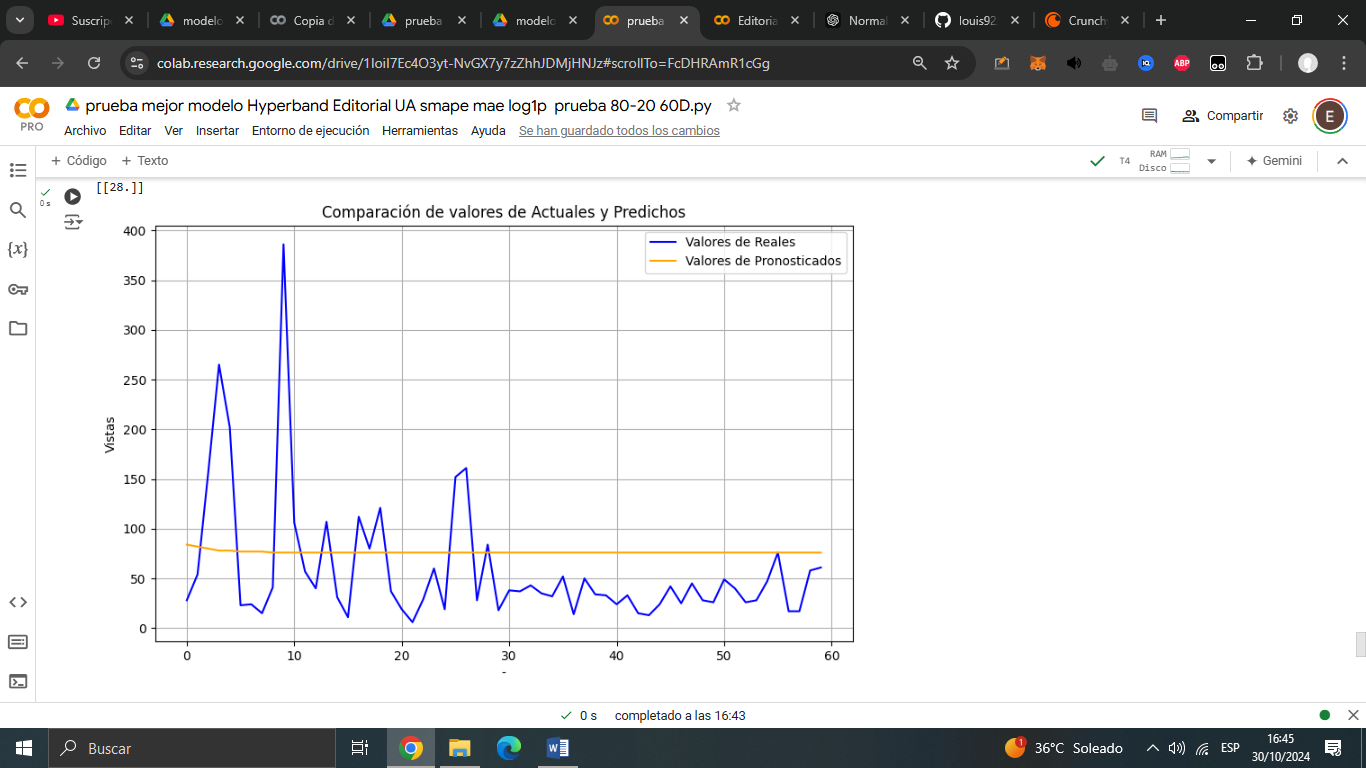


Figura 10: Grafico de Los Valores Reales y los Valores Pronosticados de una parte de conjunto de prueba Y\_test, de la prueba 5.

|  |  |  |  |
| --- | --- | --- | --- |
| # | MAE | SMAPE | RMSE |
| 4 | 258.66926229 | 1.1840436456 | 364.641053760 |
| 5 | 81.53907103825136 | 0.609703658254883 | 130.59741930437121 |

Tabla 11 Comparaciones de Rendimientos entre Pruebas 4 y 5

Como se puede observar en la tabla 11 se pudo obtener un MAE de 81.53 comparando mi mejor resultado con lo que obtuvieron en [8] tenemos un resultado intermedio entre el mejor MAE menor a 30 y el peor MAE de 132.26 que en [8] lo tuvieron en el dataset de inglés.

### 4.5.6 Conclusiones Pruebas con datos de UA

Una vez finalizadas las pruebas tanto para Editorial Universitaria UA como para Tum Transmedia UA, procedemos a resumir los mejores resultados en la siguiente tabla.

|  |  |  |
| --- | --- | --- |
| Prueba # | Conjunto de Datos | MAE |
| 1 | Tum Transmedia UA | 11.126461988304094 |
| 5 | Editorial Univeristaria UA | 81.53907103825136 |

Tabla 12 Resumen Pruebas con Mejores Resultados con datos de UA.

En la tabla 12 se puede ver que el resultado en Tum Transmedia UA comparado con [8] fue bastante mejor, mientras que el resultado de Editorial Univeristaria UA es un resultado intermedio ya que el mejor rendimiento está entre su mejor rendimiento [8] que fue un mae de 27 y su peor mejor rendimiento que fue un mae de 132.26, esto ya que en [8] tienen varios resultados, uno por cada idioma.

### 4.5.7 Conversión de Modelos de UA a Tum Transmedia GA4

Se procedió a reentrenar el mejor modelo de Tum Transmedia UA con los datos de Tum Transmedia GA4. Para utilizar el mejor modelo de Tum Transmedia UA en Tum Transmedia GA4 se construyó el dataset con las mismas columnas que Tum Transmedia UA, y se utilizó ese dataset para pruebas posteriores en Tum Transmedia GA4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Modelo | MAE | SMAPE | RMSE |
| 6 | El mejor modelo de TUM UA | 26.5404255319 | 0.57887830853 | 34.6621485053 |

Tabla 13 Prueba Número 6

En la tabla 6 se puede ver el rendimiento del mejor modelo de TUM UA aplicado a TUM GA4, como se puede observar tiene un MAE por debajo de 30 en lo que concluimos que es un buen resultado y se decide no continuar realizando más pruebas en TUMGA4

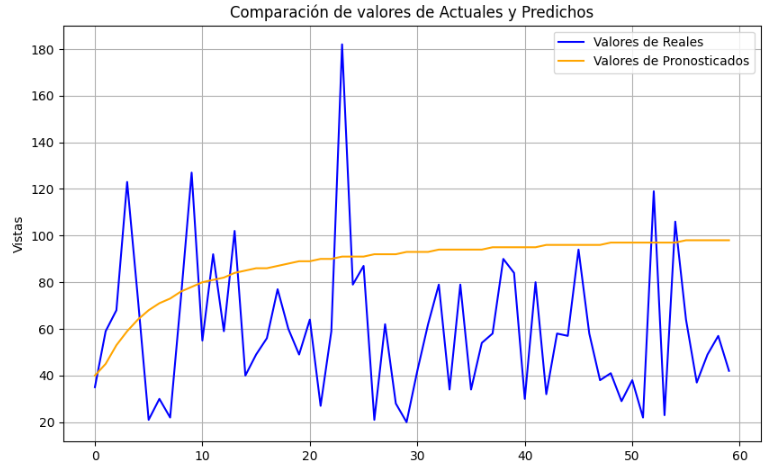


Figura 11: Grafico de Los Valores Reales y los Valores Pronosticados de una parte del conjunto de prueba Y\_test, de la prueba 6.

### 4.5.8 Pruebas Modelos de Editorial Universitaria GA4

Se procedió a reentrenar el mejor modelo de Tum Transmedia UA con los datos de Editorial Universitaria GA4 y también se reentrenó el mejor modelo de Editorial Universitaria UA con los datos de Editorial Universitaria GA4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Modelo | MAE | SMAPE | RMSE |
| 7 | El mejor modelo de TUM UA | 89.944642857 | 0.9568072360 | 129.76339273 |
| 8 | El mejor modelo de Editorial UA | 71.077976190 | 0.6042800485 | 104.79610588 |

Tabla 14 Pruebas Número 7 y 8

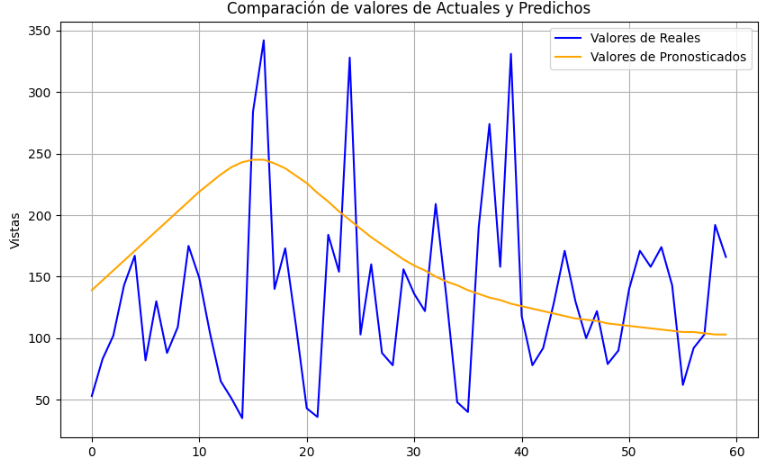


Figura 13: Gráfico de Los Valores Reales y los Valores Pronosticados de una parte del conjunto de prueba Y\_test, de la prueba 7.

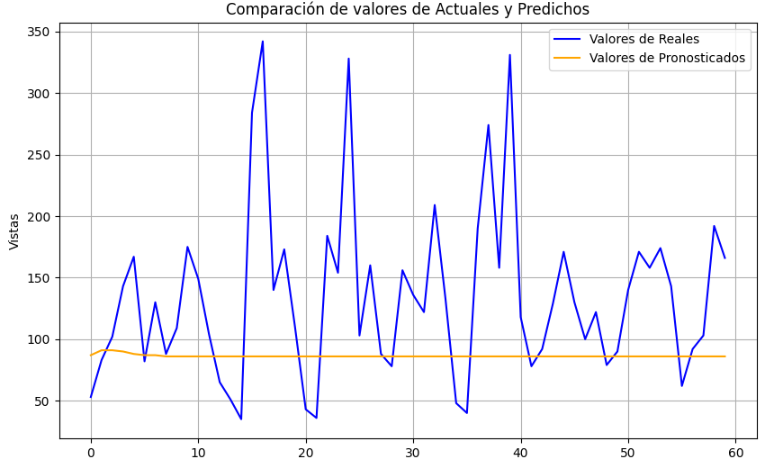


Figura 13: Grafico de Los Valores Reales y los Valores Pronosticados de una parte del conjunto de prueba Y\_test, de la prueba 8.

La tabla 14 resume los 2 pruebas realizadas con los datos de Editorial Universitaria GA4, donde el mejor modelo es el que corresponde a la prueba 8, el rendimiento obtenido de acuerdo a [8] está dentro de los mejores que ellos obtuvieron, siendo su mejor rendimiento para el conjunto de dato de las paginas en ingles de 132.26, visualmente se puede ver en la figura 12 que el modelo de la prueba 7 se adapta mejor al hacer el pico más potente, sin embargo, el modelo 8 como se puede ver en la figura 13 tiene una salida que está más cerca de una mayor cantidad de puntos de datos, esto de acuerdo al error medio absoluto y las otras métricas, esto se explica en [8] donde menciona que los modelos no con capaces de predecir picos, esto debido a que es un comportamiento impredecible proveniente de que los visitantes de las páginas web son humanos y tienen sus formas propias de actuar [8].

Agregar cantidad de epocas entrenadas

# Capítulo 5

# Conclusión

## 5.1 **Conclusiones**

El desarrollo de este trabajo aporta a la U.Na.M modelos de Pronóstico de Tráfico Web para las páginas institucionales Tum Transmedia y Editorial Universitaria. Para la página web de Tum Transmedia, el mejor modelo con los datos de UA tiene un MAE de 11.126461988304094 , mientras que el mejor modelo en el documento científico [8] donde se pronostican páginas web de Wikipedia, tiene un MAE de 17.4330 sobre el dataset de idioma chino de Wikipedia, dado que menores números de MAE se traducen en un mejores resultados, concluyo que es un buen resultado el rendimiento del modelo obtenido para Tum Transmedia.

Para la página web de la Editorial Universitaria tenemos el mejor modelo generado, este modelo tiene un MAE de 81.53907103825136, en el documento científico [8] se menciona como un buen rendimiento un MAE por debajo de 30, lo cual para este conjunto de datos no se ha podido obtener.

Respecto a medir el rendimiento de los modelos usando SMAPE según el código del ganador de Kaggle [1] tienen un SMAPE de 0.349, en mi caso los mejores modelos tienen un SMAPE que ronda los 0.6, por lo que podemos decir que medido en SMAPE el rendimiento no fue el esperado.

Durante el desarrollo de la tesis hubo un suceso inesperado, y este fue el cambio de la forma de registrar el trafico web desde la propiedad Universal Analytics a Google Analytics 4 [43] siendo este último el único en continuar funcionando, Se realizaron pruebas de pasar los modelos con los datos de UA a GA4.

El mejor modelo de Tum Transmedia UA y Editorial Universitaria UA se probaron con los datos de Editorial Universitaria GA4 el cual mejoró levemente el rendimiento pasando de un MAE de 81 a 71 aun así es un rendimiento que está por debajo de lo esperado, por otro lado, el mejor modelo de Tum Transmedia UA se probó en Tum Transmedia GA4 manteniendo un buen rendimiento con un MAE por debajo de 30.

Dado que los modelos de GA4 tuvieron un conjunto reducido de datos para poder reentrenarse, está la posibilidad de que si contáramos con una mayor cantidad de datos, pudiéramos tener un mejor rendimiento en el pronóstico.

## 5.2 Futuras líneas de investigación

A partir del desarrollo de esta Tesis surgen varias líneas de trabajo que son enumeradas a continuación:

* Desarrollar un sistema que realice el pronóstico de series de tiempo de tráfico web en tiempo real.
* Realizar un estudio de caso comparativo de diferentes IAs con los datos de las páginas web de la U.Na.M.
* Generar nuevos modelos agregando nuevos datos que solo se miden en GA4 que no se medían en UA.
* Utilizar los resultados del Pronóstico de Series de Tiempo de Tráfico Web para pronosticar el uso de recursos de hardware.
* Probar los Modelos generados con otras páginas web de la U.Na.M.

* Entrenar modelos de pronóstico de series de tiempo de tráfico web cuando se tengan una mayor cantidad de datos de GA4.

# Anexo

Detalle de pruebas donde se variaron el orden de los datos de entrenamiento prueba en fecha.

* Prueba 1 variando las fechas de entrenamiento y prueba

train\_df,test\_df = dataset[1:950], dataset[950:1250]

train\_df2 = dataset[1250:1670]

test\_df2 = dataset[1670:]

Hyperparameters:

Unidades: 288

Ratio de Aprendizaje: 0.0016406586383193105

*Dropout*: 0.1

*Recurrent Dropout*: 0.4

*Batch size*: 64

*Score*: 0.18816892802715302

* Prueba 2 variando las fechas de entrenamiento y prueba

train\_df,test\_df = dataset[1:950], dataset[950:1250]

train\_df2 = dataset[1250:1670]

test\_df2 = dataset[1670:]

Misma configuración de datos train-test que el anterior pero con más épocas.

Hiperparametros:

Unidades: 160

Ratio de Aprendizaje: 0.0007479781936524693

*Dropout*: 0.30000000000000004

*Recurrent* *Dropout*: 0.2

*Batch size*: 32

Epocas:200

*Score*: 0.18816962838172913

* Prueba 3 variando las fechas de entrenamiento y prueba

train\_df,test\_df = dataset[0:100], dataset[950:1250]

train\_df2 = dataset[250:500]

train\_df3 = dataset[1360:1490]#dataset[1250:1670]

test\_df2 = dataset[1670:]

Hiperparametros:

Unidades: 192

Ratio de Aprendizaje: 0.0028156208463983876

*Dropout*: 0.0

*Recurrent Dropout*: 0.0

*Batch size*: 32

*Score*: 0.1882011443376541

To do

Hacer conclusiones

Hacer Futuras lineas de investigación

Poner en alguna parte lo de replicar el modelo

Citar outliers

Citar predictor flag

Donde pongo el código, y cito qué parte de código, saqué de qué paginas webs

Pruebas con datos de GA4

Agregar anexo de preprocesamiento de datos de UA a GA4

Agregar knn de Ga4 tum y editorial

Agregar knn de 5 en vez de 7 para mostrar que se probó con diferentes valores

Agregar comparación de UA con GA4

Agregar lof y agregar promedio y desvió estándar ?

Agregar las columnas que se quitan de GA4?

Preguntas

1. Corregir todos al formato español – acrónimo (nombre en inglés) si-no
2. Deberia explicar mas en detalle los algoritmos de hypertuner? si-no
3. Debería de citar antes las variables que se afinaron? Ya está

Dado que para las pruebas utilizo la palabra prueba; mientras que para entrenamiento y prueba también utilizo la palabra prueba debería de cambiar la palabra?

# 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.

[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.

[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.

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

[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.

[7] “Prophet: forecasting at scale.” [Online]. Available: https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/. [Accessed: 24-Oct-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.

[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.

[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.

[11] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, pp. 1724–1734.

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

[13] “Choosing the correct error metric: MAPE vs. sMAPE,” 2020. [Online]. Available: https://towardsdatascience.com/choosing-the-correct-error-metric-mape-vs-smape-5328dec53fac. [Accessed: 12-Oct-2022].

[14] 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.

[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.

[16] D. Deng, F. Karl, F. Hutter, B. Bischl, and M. Lindauer, “Efficient Automated Deep Learning for Time Series Forecasting,” 2022.

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

[18] “The M3-Competition Database.” [Online]. Available: https://forecasters.org/resources/time-series-data/m3-competition/. [Accessed: 18-Oct-2022].

[19] A. Botchkarev, “Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology,” 2018.

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

[21] 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.

[22] 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.

[23] Python Software Foundation, “Python 3.12.1 documentation.” [Online]. Available: https://docs.python.org/3/. [Accessed: 02-Feb-2024].

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

[25] Google, “Ayuda de Google.” [Online]. Available: https://support.google.com/. [Accessed: 02-Feb-2024].

[26] Google Lcc, “Ayuda de Google,” *Ayuda de Google*, 2023. [Online]. Available: https://support.google.com/.

[27] “Microsoft: Nube, aplicaciones y juegos.” [Online]. Available: https://www.microsoft.com/es-ar/. [Accessed: 16-Apr-2024].

[28] Numpy.org, “NumPy Documentation.” [Online]. Available: https://numpy.org/doc/. [Accessed: 05-Feb-2024].

[29] Google Lcc, “TensorFlow es una plataforma de código abierto de extremo a extremo para el aprendizaje automático.” [Online]. Available: https://www.tensorflow.org/. [Accessed: 05-Feb-2024].

[30] “About pandas.” [Online]. Available: https://pandas.pydata.org/about/. [Accessed: 06-Feb-2024].

[31] “https://matplotlib.org/.” [Online]. Available: https://matplotlib.org/. [Accessed: 06-Feb-2024].

[32] Google, “Google Colab.” [Online]. Available: https://research.google.com/colaboratory. [Accessed: 02-Feb-2024].

[33] GitHub, “Acerca de Git.” [Online]. Available: https://docs.github.com/es/get-started/using-git/about-git. [Accessed: 06-Feb-2024].

[34] “Introduction to the Keras Tuner.” [Online]. Available: https://www.tensorflow.org/tutorials/keras/keras\_tuner. [Accessed: 15-Feb-2024].

[35] “Keras FAQ.” [Online]. Available: https://keras.io/getting\_started/faq/#. [Accessed: 15-Feb-2024].

[36] “The base Tuner class.” [Online]. Available: https://keras.io/api/keras\_tuner/tuners/base\_tuner/. [Accessed: 15-Feb-2024].

[37] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *J. Mach. Learn. Res.*, vol. 18, pp. 1–52, 2018.

[38] A. Baghbanpourasl, E. Lughofer, P. Meyer-Heye, H. Zorrer, and C. Eitzinger, “Virtual quality control using bidirectional lstm networks and gradient boosting,” in *IEEE International Conference on Industrial Informatics (INDIN)*, 2019, vol. 2019-July, pp. 1638–1643.

[39] J. Tourille, O. Ferret, X. Tannier, and A. L. Nvol, “Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers,” in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 2, pp. 224–230.

[40] “[UA] Información sobre las agrupaciones de canales.” [Online]. Available: https://support.google.com/analytics/answer/6010097. [Accessed: 07-Feb-2024].

[41] “[GA4] Informe ‘Detalles de la tecnología.’” [Online]. Available: https://support.google.com/analytics/answer/12980150. [Accessed: 07-Feb-2023].

[42] A. Suilin, “kaggle-web-traffic.” [Online]. Available: https://github.com/Arturus/kaggle-web-traffic. [Accessed: 07-Nov-2024].

[43] “[UA→GA4] Diferencias entre los datos de Universal Analytics y Google Analytics 4.” [Online]. Available: https://support.google.com/analytics/answer/9964640. [Accessed: 10-Feb-2024].

[44] T. T. Dang, H. Y. T. Ngan, and W. Liu, “Distance-based k-nearest neighbors outlier detection method in large-scale traffic data,” *Int. Conf. Digit. Signal Process. DSP*, vol. 2015-Septe, pp. 507–510, 2015.

[45] A. Blázquez-Garc\’\ia, A. Conde, U. Mori, and J. A. Lozano, “A Review on Outlier/Anomaly Detection in Time Series Data,” *ACM Comput. Surv.*, vol. 54, no. 3, 2021.