ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

Modelo de predicción de la producción de energía de la Central Hidroeléctrica Coca Codo Sinclair, basado en técnicas de aprendizaje computacional

Alvarez Chiriboga, Daniel Alejandro

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión de Sistemas de

Información e Inteligencia de Negocios

Msc. Díaz Zúñiga, Magi Paúl

9 de septiembre de 2020

URKUND

Document Information

Analyzed document	TESIS Daniel Alvarez.pdf (D85028129)
Submitted	11/12/2020 11:36:00 PM
Submitted by	DIAZ ZUÑIGA PAUL
Submitter email	mpdiaz@espe.edu.ec
Similarity	2%
Analysis address	mpdiaz.espe@analysis.urkund.com

Sources included in the report

W	URL: https://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2016/02/Regula ... Fetched: 11/13/2020 3:14:00 AM	5
W	URL: https://www.regulacionelectrica.gob.ec/wp-content/plugins/download-monitor/downloa ... Fetched: 11/13/2020 3:14:00 AM	3
W	URL: https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining Fetched: 11/13/2020 3:14:00 AM	2
W	URL: https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html Fetched: 11/13/2020 3:14:00 AM	1
W	URL: https://decidesoluciones.es/calidad-o-cantidad-de-datos-para-ia/ Fetched: 11/13/2020 3:14:00 AM	1
W	URL: https://www.xataka.com/robotica-e-ia/machine-learning-y-deep-learning-como-entende ... Fetched: 11/13/2020 3:14:00 AM	2

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Modelo de predicción de la producción de energía de la Central Hidroeléctrica Coca Codo Sinclair, basado en técnicas de aprendizaje computacional**" fue realizado por el señor **Alvarez Chiriboga, Daniel Alejandro** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 13 de noviembre de 2020

Firma:

MAGI PAUL DIAZ

Msc. Díaz Zúñiga, Magi Paúl

Director

C.C.: 1707249072

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo **Alvarez Chiriboga, Daniel Alejandro**, con cédula de ciudadanía n° 1717980088, declaro que el contenido, ideas y criterios del trabajo de titulación: **Modelo de predicción de la producción de energía de la Central Hidroeléctrica Coca Codo Sinclair, basado en técnicas de aprendizaje computacional** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 13 de noviembre de 2020

Firma

DANIEL ALEJANDRO ALVAREZ CHIRIBOGA

Ing. Alvarez Chiriboga, Daniel Alejandro

C.C.: 1717980088

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo **Alvarez Chiriboga, Daniel Alejandro** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Modelo de predicción de la producción de energía de la Central Hidroeléctrica Coca Codo Sinclair, basado en técnicas de aprendizaje computacional** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 13 de noviembre de 2020

Firma

Ing. Alvarez Chiriboga, Daniel Alejandro

C.C.: 1717980088

Dedicatoria

A mis padres, a mis amadas esposa e hija, pero en especial este trabajo va dedicado a mi hermana Carlita Nataly Alvarez Chiriboga, que Jehová te tenga en su memoria y a través su misericordia nos permita volver a verte. Mientras tanto nos queda el triste consuelo de ver que en nuestros sueños recorras los caminos que esta vida te quedó debiendo.

Agradecimiento

En primer lugar a Jehová Dios, que es quién permite que haya podido culminar este logro a través de sus infinitas bendiciones.

A mis padres, quienes desde pequeño me guiaron para ser cada día una mejor persona y me dieron la formación en valores para poder llegar a ser quién soy ahora.

A mi esposa e hija, que comparten mi día a día y son mi motivación para seguir adelante.

A mis docentes incluido mi tutor, que fueron la fuente de conocimientos necesarios para poder culminar con éxito el presente trabajo de investigación.

A mis amigos de vida y de aula, con quienes he compartido valiosos momentos.

Índice de Contenido

Índice de Tablas

Índice de Figuras

Resumen

El presente trabajo de investigación aborda el uso del aprendizaje computacional para la implementación de modelos de predicción. El caso específico abordado corresponde a la determinación de la cantidad de energía que puede producir la Central Hidroeléctrica Coca Codo Sinclair a través de la predicción del caudal del Río Coca que provee los recursos hídricos a la central. Para lograr este objetivo se utilizó los registros históricos del caudal promedio diario del Río Coca entre el año 1972 y 2020. Los datos históricos se estructuraron mediante una serie temporal y sirvieron como base para la generación y entrenamiento de los modelos construidos usando Redes Neuronales. Mediante un proceso automático se generaron 240 modelos; los cuales fueron evaluados con respecto al error generado entre las predicciones entregadas y los valores reales de caudal. Con la evaluación se identificó el Modelo cuya arquitectura y parámetros de ajuste presentó la mejor precisión en el cálculo de pronósticos del caudal. Las predicciones del Modelo Final fueron evaluadas con aquellas calculadas actualmente en la CHCCS con un Modelo Lineal (Arima), arrojando un RSME de 95,99 y MAPE de 25,70% para el Modelo Propuesto frente a un RSME DE 129,21 y un MAPE de 32,46 para el Modelo Lineal (ARIMA). Finalmente se concluye que la aplicación de un Modelo de Predicción, basado en técnicas de aprendizaje automático, puede mejorar la precisión de la programación de la energía a generar por la CHCCS.

PALABRAS CLAVE:

- **PREDICCIÓN / PRONÓSTICO**
- **REDES NEURONALES**
- **CAUDAL DEL RÍO**
- **SERIES TEMPORALES**

Abstract

This research work addresses the use of computational learning for the implementation of prediction models. The specific case addressed corresponds to the determination of the amount of energy that the Coca Codo Sinclair Hydroelectric Plant can produce through the prediction of the flow of the Coca River that provides the water resources to the plant. To achieve this objective, the historical records of the daily average value of the Coca River between 1972 and 2020 were used. The historical data were structured through a univariate time series and served as a basis for the generation and training of the models built using Artificial neural networks. Through an automatic process, 240 models were generated; which were evaluated with respect to the error generated between the predictions delivered and the real flow values. As a result of the evaluation, the Model whose architecture and adjustment parameters presented the best precision in the calculation of forecasts of the flow of the Coca River was identified. The predictions of the Final Model were evaluated with those currently calculated in the CHCCS with a Linear Model (Arima), yielding a RSME of 95.99 and MAPE of 25.70% for the Proposed Model compared to an RSME OF 129.21 and a MAPE of 32.46 for the Linear Model (ARIMA). Finally, it is concluded that the application of a Prediction Model, based on machine learning techniques, can improve the precision of the programming of the energy to be generated by the CHCCS.

KEYWORDS:

- **PREDICTION / FORECAST**
- **NEURONAL NETWORK**
- **RIVER FLOW**
- **TEMPORAL SERIES**

Capítulo I: El Problema de Investigación

Antecedentes

El Ecuador ha hecho uso complementario entre la energía hidroeléctrica (renovable) y la energía térmica (no renovable). De acuerdo a las condiciones geográficas propias del país, desde inicios del siglo pasado la región sierra contaba con pequeñas centrales hidroeléctricas, mientras que debido a la falta de caudales de agua junto al factor de altura, las poblaciones de la costa optaron por el desarrollo de centrales térmicas (Mena, 2017). Con el pasar del tiempo el Estado Ecuatoriano asumió como deber satisfacer las necesidades de energía eléctrica en el país, a través del aprovechamiento óptimo de sus recursos naturales, considerando al suministro de energía como un servicio de utilidad pública e interés nacional.

Con este antecedente se estableció el Sector Eléctrico Ecuatoriano, el cual se comprendía entre otros componentes al Centro Nacional de Control de Energía (actualmente CONELEC), Empresas Eléctricas de: generación, transmisión (actualmente CELEC) y las Empresas eléctricas de distribución y comercialización. El CENACE coordina el despacho de todos sus agentes generadores mediante la elaboración de la Planificación Operativa Energética, cuyo objetivo es establecer una política óptima para el uso de los recursos disponibles para la generación eléctrica. Para la realización de la Planificación Operativa Energética diaria, cada uno de los agentes del MEM tiene la obligación de suministrar al CENACE, a través de medios electrónicos, la curva de generación prevista de cada agente generador con un umbral de predicción de 24 horas.

La Central Hidroeléctrica Coca Codo Sinclair (CHCCS), se encuentra bajo la administración y operación de la Unidad de Negocio COCA CODO SINCLAIR, la cual

pertenece a la Empresa Pública Estratégica Corporación Eléctrica del Ecuador CELEC E.P. La Unidad de Negocio Coca Codo Sinclair, a través de la Jefatura de Operación de la CHCCS tiene la responsabilidad de elaborar y remitir, con periodicidad diaria, al CENACE la “programación de la producción de energía” con un umbral de 24 horas de antelación. Esta programación corresponde a la “curva de generación prevista” y es usada como guía para la ejecución del despacho diario por parte del personal técnico del CENACE, y define de manera preliminar la asignación del aporte de cada agente generador, para suplir toda la demanda energética del MEM.

El Problema de Investigación

Contexto del Problema

La programación de la energía a producir por la CHCSS, es un proceso de cálculo numérico cuya precisión en la actualidad está relacionada exclusivamente a los datos históricos de las variables que influyen la generación eléctrica y a la experiencia de los operadores/supervisores de la CHCCS. Este método de elaboración, está sujeto a una estimación poco precisa de los valores de energía a generar por la CHCCS para umbrales futuros, esto es consecuencia principalmente de la naturaleza estocástica y no lineal entre las variables inmersas en el proceso de generación hidroeléctrica (Sauhats, Petrichenko, Baltputnis, Broka, & Varfolomejeva, A multi-objective stochastic approach to hydroelectric power generation scheduling, 2016)

Planteamiento del Problema

La estimación poco precisa en la predicción (umbrales de 24 horas de anticipación) de la energía que la CHCCS puede producir, constituye la problemática a ser abordada en el presente trabajo de investigación. Podemos señalar que el problema planteado afecta de manera directa en el objetivo establecido por el CENACE para las

tareas de Operación y Mantenimiento de la CHCCS, el cual se enfoca en que las tareas citadas anteriormente busquen alcanzar el “aprovechamiento óptimo de los recursos hídricos disponibles de cada agente generador del MEM, en nuestro caso particular la CHCCS.

En contraposición a este objetivo, como consecuencia del problema suscitado por la imprecisión en la estimación de la energía a producir, se producen efectos como:

- Pérdida de Recursos Económicos para CELEC EP, debido a una disminución en la facturación de la energía generada por la CHCCS.
- Aumento del periodo de retorno de la inversión realizada por el Estado Ecuatoriano en la construcción de la central.
- Incumplimiento de los aportes programados de energía de la CHCCS, de acuerdo a la operación del CENACE. Este incumplimiento se refleja en la necesidad del uso de otras fuentes no renovables y con costos de generación más altos (generación termoeléctrica, importación de energía).
- Reducción de indicadores de eficiencia de Operación de la CHCCS.

Una de las causas directamente relacionada con el problema detallado, corresponde al mecanismo de cálculo de la estimación de la programación de la energía (neta y bruta) que la CHCCS estará en capacidad de generar con un umbral de anticipación de 24 horas. El método de cálculo actual para establecer la curva de predicción de la energía para períodos futuros se puede considerar manual debido a toda la carga operativa que los responsables de dicha actividad deben ejecutar. A esto se suma el hecho que en la mayoría de ocasiones se añaden correcciones a los cálculos manuales, los cuales son producto del discernimiento empírico y la experiencia acumulada de los funcionarios responsables de la Supervisión de la Operación.

Objetivos

Objetivo General

Implementar un Modelo de Predicción de la producción de energía de la CHCCS, a través de técnicas de aprendizaje computacional, para mejorar el aprovechamiento de los recursos hídricos disponibles.

Objetivos Específicos

OE1: Realizar una Revisión Inicial de Literatura para identificar los métodos de elaboración de modelos predictivos, asociados a la generación de energía hidroeléctrica; y definir el método y los modelos que mejor se ajusten a la realidad de la CHCCS.

OE2: Identificar las fuentes de información relacionadas con las variables que intervienen en la predicción de la generación hidroeléctrica de la CHCCS.

OE3: Implementar un modelo predictivo que permita elaborar la programación de la energía generada por la CHCCS.

OE4: Validar la efectividad del modelo implementado, para la predicción de la energía generada por la CHCCS, en comparación con los registros históricos.

Hipótesis

El uso de técnicas de aprendizaje computacional permite obtener una mejor precisión en la predicción de la producción de energía de la Central Hidroeléctrica Coca Codo Sinclair.

Señalamiento de Variables

Variable dependiente

Programación de producción de energía hidroeléctrica.

Variable independiente

Modelo Predictivo basado en técnicas de aprendizaje computacional.

Justificación, Importancia y Alcance

De acuerdo a la Ley del Régimen del Sector Eléctrico el CENACE tiene la administración de todas las transacciones del MEM, la CHCCS al ser un ente generador suscrito al MEM se encuentra en obligación de cumplir con los reglamentos y regulaciones emitidos por el CENACE. Entre las disposiciones emitidas se indica que la CHCCS debe proporcionar, de manera oportuna y veraz, toda la información que le sea solicitada, con el objetivo de poder efectuar la planificación operativa y el despacho de la central (ARCONEL, 2019).

La información solicitada por el CENACE debe ser enviada de manera obligatoria y periódica por la CHCCS a través de la programación de energía, los datos proporcionados corresponden a la “curva de generación horaria prevista para el día siguiente”. La programación debe ser enviada hasta las 10H00 a.m. de cada día y responder a políticas óptimas de operación del de la central embalse (mínimo costo de producción, previsión de vertimientos, etc.), incluyendo los datos pronósticos del caudal de la fuente hídrica que alimenta al embalse. El incumplimiento de esta obligación está a sujeta a sanciones por parte del ente regulador (ARCONEL, 2000).

Utilizar un modelo de predicción, basado en aprendizaje computacional, que mejore la precisión de la energía a producir por la CHCCS, es sin duda un aporte en pos de un abastecimiento eléctrico bajo condiciones de soberanía y priorizando el uso de energías renovables, Y aporta en la sustitución progresiva de fuentes no renovables, mismas que son más costosas y contaminantes; como el caso de la generación térmica (MEER, 2016).

Con estas consideraciones el alcance del presente trabajo de investigación está orientado en mejorar la precisión de la programación de la energía a generar por la CHCCS, con un umbral de 24 horas de antelación, mediante la implementación de un modelo predictivo basado en técnicas de aprendizaje computacional. Para la consecución de este objetivo general se iniciará con la fase de “entendimiento del negocio”, lo que implica empoderarse de la forma actual de elaboración de la programación de la energía a producir por la CHCCS.

De manera paralela podremos definir el estado del arte a través de una revisión sistemática de literatura; lo que nos permitirá determinar el modelo predictivo y la técnica de aprendizaje computacional más adecuados para abordar de manera exitosa la problemática planteada de acuerdo a su naturaleza. Con estos objetivos alcanzados se realizará la fase más extensa, que consiste en el proceso de extracción, transformación y limpieza de los datos que servirán de entrada para el modelo a implementarse. Finalmente, de acuerdo a nuestra metodología de desarrollo de la investigación (Werick, 2017), se someterá el modelo implementado a datos nuevos y se validará la efectividad del mismo con respecto a la hipótesis planteada.

Preguntas de Investigación

OE1 – RQ1.1: ¿Existe algún método formal para la elaboración de la programación de energía a generar por la CHCCS?

OE1– RQ1.2: ¿Cuáles son métodos, técnicas y modelos que reportan mayor precisión en la predicción de generación de energía hidroeléctrica?

OE2 – RQ2.1: ¿Cuáles son los registros de operación de la CHCCS que

intervienen de modo directo en la elaboración de la programación de la energía a generar?

OE2 – RQ2.2: ¿Qué técnicas de extracción, transformación y limpieza son necesarias ejecutar para preparar los datos relacionados con la programación de la energía a generar?

OE3 – RQ3.1: ¿Cuáles fueron los criterios de selección para definir el modelo implementado?

OE3 – RQ3.2: ¿Cuáles son los parámetros de ejecución utilizados y definidos para la implementación del Modelo?

OE4 – RQ4.1: ¿Qué margen de error arroja el modelo con respecto a los datos utilizados para su implementación?

OE4 – RQ4.2: ¿La precisión de la programación de energía a generar por la CHCCS, calculada a través del modelo implementado, mejoró en comparación con los valores generados de forma habitual?

Metodologías

Metodología de Investigación

La metodología de investigación utilizada será cualitativa, y se desarrollará a través de la técnica de estudio de caso. Los objetivos que persigue esta técnica son:

- Elaboración de una o varias hipótesis

- Confirmación de hipótesis
- Descripción y registros de los hechos
- Comprobación o comparación de situaciones similares

Fases de la Metodología de Estudio de Caso. De modo tradicional, la ejecución de un Estudio de Caso está compuesto por cinco fases, citadas a continuación (Rovira, s.f.):

Selección de Caso. Esta fase inicia con la definición del problema y los objetivos que queremos alcanzar. Con estas definiciones resueltas procedemos a establecer el ámbito para nuestro estudio y seleccionamos un caso apropiado y relevante.

Elaboración de Preguntas. Una vez seleccionado nuestro caso y tema de estudio establecemos un conjunto de preguntas que se quiere despejar una vez finalizado el trabajo de investigación.

Localización de fuentes y recopilación de datos. Para poder identificar las fuentes de la información necesaria para nuestro trabajo se puede hacer uso de: técnicas de observación, entrevistas con los involucrados, etc. Con las fuentes establecidas planteamos las técnicas necesarias de acuerdo a cada caso particular para poder realizar los procesos de recopilación centralizada de los datos necesarios.

Análisis e interpretación de la información y los resultados. Con los datos recopilados y procesados pasamos a la comparación de nuestros resultados con la hipótesis planteada en un inicio. Y como resultado de este análisis comparativo él o los investigadores generar una seria de conclusiones sobre el resultado del trabajo de investigación.

Elaboración de Informe. Finalmente se prosigue con el registro detallado y cronológico de los datos de nuestro estudio de caso. Se necesita especificar los pasos seguidos para llegar a las conclusiones finales, así como el proceso de obtención de la información. Una característica primordial de este informe es que su redacción debe ser en lenguaje claro y que permita la comprensión al lector, de todos sus puntos.

Metodología de Desarrollo

El presente trabajo de investigación se llevará cabo mediante una técnica híbrida entre las fases de CRISP – DM (CRoss-Industry Standard Process for Data Mining) y la Metodología ágil.

El objetivo de utilizar esta técnica híbrida es poder aprovechar las ventajas que entrega el poder adaptar las fases tradicionales de CRISP-DM a la forma de trabajo iterativa de la Metodología Ágil (SCRUM):

- Descomponer el trabajo necesario, para alcanzar nuestro objetivo, en piezas más pequeñas (sprints) y más manejables.
- Poder entregar un producto potencialmente usable después de cada sprint.
- La ejecución de los sprints generará un proceso incremental de ajuste entre los resultados obtenidos y el objetivo; retroalimentado con los requerimientos de los interesados.

CRISP – DM. Se puede describir a CRISP-DM como una metodología que sigue una estructura jerárquica, es un modelo bastante flexible que pretende adaptarse a las necesidades específicas de minería de datos de una organización.

Esta estructura tiene como fin la obtención de resultados de forma más rápida y eficiente y comprende las siguientes fases:

Comprensión del Negocio. En esta fase se busca la comprensión de los objetivos del Negocio. Esto se traduce en tener la capacidad de definir los datos necesarios para cumplir los objetivos definidos.

Estudio y Comprensión de los Datos. Las actividades de esta fase comprenden la recolección inicial de datos, familiarización de los datos, reconocimiento de problemas de calidad de los datos e identificación preliminar del comportamiento y patrones de los datos.

Preparación de los Datos. Con base en la recolección de datos en la fase previa, esta fase se encargar de realizar los procesos de limpieza y preparación de los datos. Estas tareas se enfocan en que los datos se estructuren de tal manera que puedan ser utilizados en fase de modelado.

Modelado. En esta fase seleccionamos y aplicamos técnicas de modelado que permitan alcanzar el objetivo planteado. Cada técnica de manera general tiene sus propios parámetros e incluso puede ser necesario preparar realizar estructuras adicionales de datos para ciertos modelos.

Evaluación. Para este punto ya se deben haber desarrollado varios modelos o varias iteraciones mejoradas de un modelo específico que nos permite realizar comparaciones y llegar a conclusiones sobre la capacidad del modelo de haber alcanzado nuestro objetivo.

Despliegue. La fase final es presentar modelo realizado al cliente final, esto puede comprender ejecuciones periódicas del modelo o incluso procesos de implementación de procesos automatizados, sistemas, aplicativos, etc.

Figura 1

Diagrama de las fases de CRISP-DM

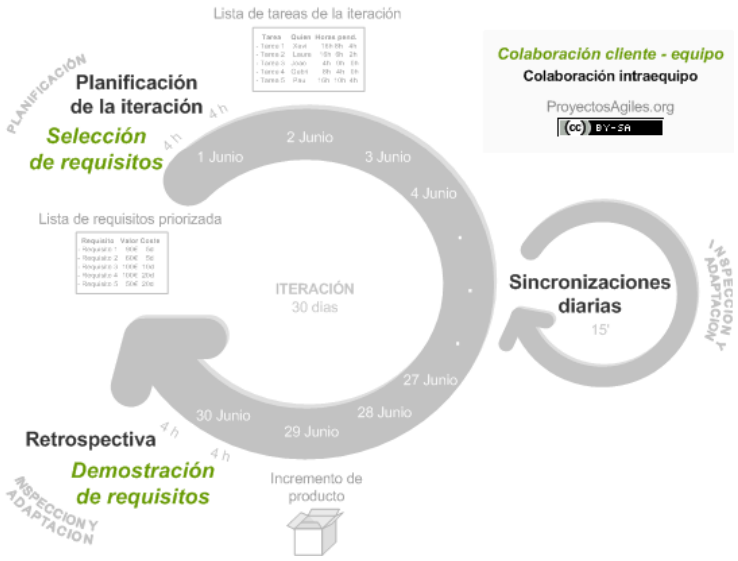
Nota: Tomado de (Cross Industry Standard Process for Data Mining, 2020)

Metodología Ágil Scrum. SCRUM es una de las más conocidas metodologías ágiles, en la cual un grupo de trabajo ejecuta tareas definidas de trabajo en un período de tiempo determinado, períodos conocidos como “sprints” o iteraciones. Estos sprints generalmente no duran más de un mes. Para arrancar con un sprint se elabora una lista de funcionalidades que deberá contener nuestro producto final, las funcionalidades definidas en esta lista son priorizadas de acuerdo a la importancia que guardan con los objetivos del negocio (Martinez, s.f.). Las funcionalidades que fueron escogidas como prioritarias se convierten en los objetivos a entregar al final del primer sprint. Cuando se ha finalizado el primer sprint, a través de retroalimentación, se definen los refinamientos a realizarse en el producto final, y se toman las funcionalidades que siguen dentro del

orden de prioridad y se realiza un nuevo sprint. Este proceso tal como se muestra en la Figura 2 se repite de manera iterativa hasta alcanzar el objetivo final y cumplir con todas las funcionalidades requeridas del producto final.

Figura 2

Metodología Ágil SCRUM



Nota: Tomado de (Proyectos Ágiles, s.f.)

Metodología Híbrida Propuesta.

La metodología que se plantea combina las fases definidas de CRISP-DM con los sprints iterativos de la metodología SCRUM. Para esto se propone dividir la ejecución del presente trabajo de investigación en tres fases:

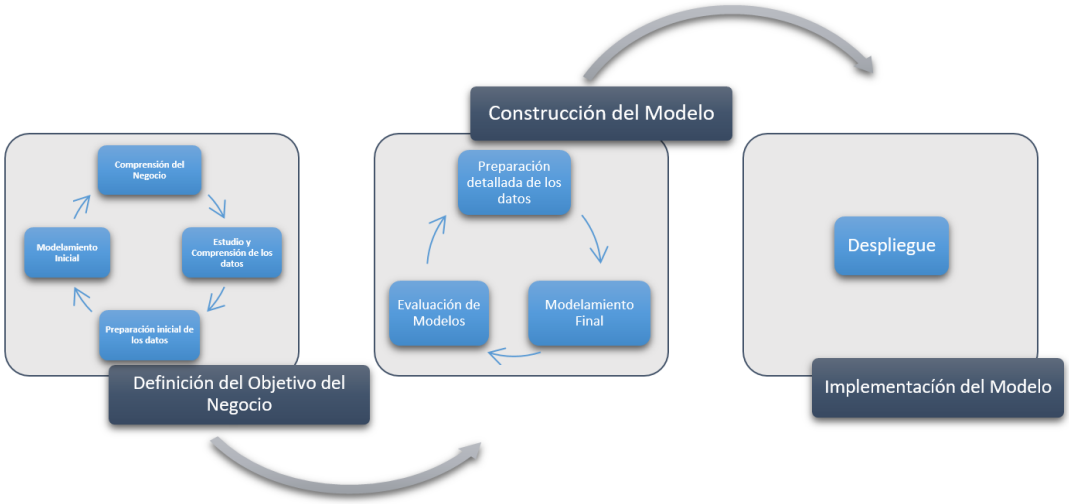
- Definición del Objetivo del Negocio
- Construcción del Modelo
- Implementación del Modelo

Cada una de estas fases se las ejecutará mediantes sprints, hasta que el

resultado final de cada una pueda ser el insumo de arranque para la siguiente fase. El diagrama de la metodología híbrida propuesta se muestra en la Figura 3, donde se plantean tres fases principales y dentro de cada una se utiliza una o varias de las tareas definidas en la metodología CRISP-DM.

Figura 3

Diagrama de Metodología Híbrida de Investigación



Nota: Elaboración Propia, 2020

Definición del Objetivo de Negocio. En esta fase se plantea realizar sprints que nos permite entender el negocio a través de varias iteraciones que vayan desde el entendimiento del negocio hasta la construcción de modelos iniciales. Estos modelos iniciales buscan entender el comportamiento y naturaleza de los datos iniciales. A su vez estos modelos iniciales se pueden analizar con los interesados para identificar cuáles son los datos que pueden aportar a la consecución del objetivo del negocio.

Construcción del Modelo. Una vez comprendido el objetivo de Negocio, a través de los sprints realizados en la fase anterior, estamos en capacidad de poder

escoger el conjunto de datos con los cuales se construirán los distintos modelos propuestos. En los sprints de esta fase se vuelven a repetir las actividades de la fase anterior: Preparación de los Datos, y Modelamiento; sin embargo en este punto ambas actividades son de carácter más complejo ya que están orientados a la construcción de los modelos que serán evaluados como solución de la problemática planteada.

Despliegue. Finalmente con el modelo o modelos escogidos como solución, se realiza la implementación para el llamado usuario final. Y también contiene una evaluación final de la solución desplegada. Sin embargo se considera que esta actividad está implícita en el despliegue y en teoría se esperaría que en esta fase final se ejecute en dos sprints.

Relación Entre los Objetivos Específicos y la Metodología de Desarrollo Híbrida propuesta.

La consecución de cada uno de los objetivos serán alcanzados a través de la ejecución de las fases que componen la metodología de investigación híbrida propuesta previamente. En la Tabla 1 se detalla la relación establecida entre los Objetivos Estratégicos planteados con la o las fases que componen la Metodología Híbrida propuesta; además se incluyen las tareas necesarias para ejecutar cada fase.

Tabla 1

Relación: Objetivos Específicos - Fases de Metodología de Investigación

OBJETIVOS ESPECÍFICOS	**FASES DE LA METODOLOGIA HÍBRIDA PROPUESTA**
OE1: Realizar una Revisión Inicial de Literatura para identificar los métodos de elaboración de modelos predictivos, asociados a la generación de energía hidroeléctrica; y definir el método y los modelos que mejor se ajusten a la realidad de la CHCCS.	Definición del Objetivo de Negocio: (Comprensión del Negocio – Estudio y Comprensión de los Datos)

OBJETIVOS ESPECÍFICOS	FASES DE LA METODOLOGIA HÍBRIDA PROPUESTA
OE2: Identificar las fuentes de información relacionadas con las variables que intervienen en la predicción de la generación hidroeléctrica de la CHCCS.	Definición del Objetivo de Negocio: (Preparación Inicial de los datos – Modelamiento Inicial)
OE3: Implementar un modelo predictivo que permita elaborar la programación de la energía generada por la CHCCS.	Construcción del Modelo: (Preparación Detallada de los Datos – Modelamiento Final)
OE4: Validar la efectividad del modelo implementado, para la predicción de la energía generada por la CHCCS, en comparación con los registros históricos.	Construcción del Modelo: (Evaluación del Modelo) Implementación del Modelo (Despliegue)

Nota: Elaboración Propia

Capítulo II: Revisión de Literatura y Marco Teórico

Revisión de Literatura

Para analizar el estado del arte se utilizó el proceso de Revisión Inicial de Literatura (RIL) Con el problema ya planteado en el capítulo 3 de este documento continuaremos con las siguientes fases, a saber:

- Definición del objetivo de la Revisión Inicial de Literatura
- Criterios de inclusión y exclusión
- Grupo de Control
- Cadena de Búsqueda
- Proceso de Selección
- Resultados y síntesis

Motivación de la Revisión Inicial de Literatura

La ejecución de la RIL tiene como propósito poder determinar el estado del arte de la problemática que representa la estimación de la programación de energía a generar en centrales hidroeléctricas relacionado con causa de dicha problemática, en concordancia con el alcance definido en el presente trabajo de investigación.

Criterios de Inclusión y Exclusión

La definición de los siguientes criterios permitió definir estudios que se encuentren enmarcados en el propósito del trabajo de investigación de una forma más explícita.

Inclusión. Estudios que hayan sido publicados desde el año 2014 hasta la actualidad, con el objetivo de abarcar investigaciones actuales con respecto a la problemática planteada.

- Estudios publicados en bibliotecas digitales reconocidas y redactados en inglés.
- Estudios de investigación relacionados con el aprovechamiento de recursos hídricos y programación de energía de centrales hidroeléctricas, con soluciones orientadas a técnicas de aprendizaje computacional.
- Estudios que reporten comparaciones entre soluciones orientadas al aprendizaje computacional y otros métodos alternativos.

Exclusión. Estudios que utilicen técnicas de aprendizaje computacional que aborden la problemática.

- Estudios realizados en las llamadas “Small Hydropower”, es decir centrales de generación hidroeléctrica de baja potencia instalada. Por no tener relación con la potencia instalada en la CHCCS.
- Estudios que traten de sistemas de programación o planificación a largo plazo.

Estrategia de Búsqueda

La ejecución de la RIL tiene como propósito poder determinar el estado del arte de la problemática que representa la estimación de la programación de energía a generar en centrales hidroeléctricas y de la causa de esta problemática de acuerdo al alcance definido en el presente trabajo de investigación. Para lograr este cometido se ejecutan las siguientes tareas:

Revisión Inicial. Se ejecutó la búsqueda inicial en distintas bibliotecas digitales, y se realizó una selección preliminar de los estudios que conformarían en grupo de control.

Validación Cruzada. Con apoyo por parte del docente evaluador se analizaron los estudios preseleccionados, y se verificó que cumplan con los criterios de inclusión y exclusión planteados.

Integración del Grupo de Control. En base a las dos actividades precedentes y a través de la discusión de los estudios resultantes de dichas actividades, se conformó el Grupo de Control con los estudios mostrados en la Tabla 2.

Tabla 2

Grupo de Control

ESTUDIO DE CONTROL	TÍTULO	PALABRAS CLAVES
EC1	Multistep-ahead Streamflow and Reservoir Level Prediction Using ANNs for Production Planning in Hydroelectric Stations	Prediction, ANNs, Hydroelectric, Reservoir Level, Streamflow
EC2	Comparison of multiple linear regression, artificial neural network, extreme learning machine, and support vector machine in deriving operation rule of hydropower reservoir	Hydropower reservoir, operation rule derivation, multiple linear regression, artificial neural network, extreme learning machine, support vector machine, dynamic programming
EC3	Hybrid metaheuristic of artificial neural network — Bat algorithm in forecasting electricity production and water consumption at Sultan Azlan shah Hydropower plant	Artificial Neural Network (ANN), Bat Algorithm, electricity production, water consumption, hydropower
EC4	ANN-based forecasting of hydropower reservoir inflow	ANN, forecasting, hydropower, reservoir inflow
EC5	A water flow forecasting for dam using neural networks and regression models	Forecasting, Hydroelectric power generation, Neural Networks
EC6	Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm	Reservoir operation, Artificial intelligence, BP neural network, SVR, LSTM
EC7	Hydrological flow rate estimation using artificial neural networks Model development and potential applications	Flow rate, Artificial neural networks, Monte-Carlo simulation, Climatic scenario, Hydroenergy production

Nota: Elaboración Propia, 2020

Cadena de Búsqueda

Como punto inicial para esta actividad se definieron los términos representativos que servirán de base para la estructura de la cadena de búsqueda. La Tabla 3 presenta una síntesis jerárquica de términos (orden por frecuencia de apariciones).

Términos Representativos de la Cadena de Búsqueda. Como punto inicial para esta actividad se definieron los términos representativos que servirán de base para la estructura de la cadena de búsqueda. La Tabla 3 presenta una síntesis jerárquica de términos (orden por frecuencia de apariciones).

Tabla 3

Síntesis Jerárquica de términos

TÉRMINO REPRESENTATIVO	EC1	EC2	EC3	EC4	EC5	EC6	EC7	NÚMERO DE APARICIONES
Artificial	X	X	X	X	X	X	X	7
Neural		X	X	X	X	X	X	6
Network		X	X	X	X	X	X	6
Algorithm		X	X	X		X	X	5
Dam	X		X	X	X	X		5
Forecast			X	X	X		X	4
Inflow – flow				X	X	X	X	4
Hydropower		X	X	X			X	4
Prediction	X	X		X			X	4
Reservoir	X	X		X		X		4
schedul*		X		X		X		3
Plant	X		X				X	3

Nota: Elaboración Propia, 2020

Para la construcción de la cadena de búsqueda se utilizaron los términos representativos del grupo de control, los términos fueron relacionados entre sí a través de los operadores lógicos AND, OR y *. Adicional a éstos términos de acuerdo a los criterios de exclusión se añadieron los términos: short-term system (sistemas a corto

plazo) Mediante ejecución iterativa de búsquedas en el los repositorios digitales se definió la siguiente cadena de búsqueda:

((TITLE-ABS (hydropower AND reservoir) OR TITLE-ABS (hydropower AND plant)) AND (TITLE-ABS (artificial AND neural AND network) OR TITLE-ABS (ann*)) AND (TITLE-ABS (forecast*) OR TITLE-ABS (schedul*))) OR ((TITLE-ABS (artificial) AND TITLE-ABS (neural)) AND (TITLE-ABS (dam) OR TITLE-ABS (predict*)) AND TITLE-ABS (reservoir) AND TITLE-ABS (power)) OR ((TITLE-ABS (scheduling) AND TITLE-ABS (hydropower AND plant) AND TITLE-ABS (system) AND (TITLE-ABS (short-term) AND TITLE-ABS (algorithm))) OR (TITLE-ABS (forecasting) AND TITLE-ABS (streamflow) AND TITLE-ABS (neural) AND TITLE-ABS (algorithm*)))

Proceso de Selección

El resultado de este proceso fue alcanzando mediante la validación cruzada de los estudios arrojados por la cadena de búsqueda en el repositorio digital SCOPUS. Los estudios escogidos fueron sometidos a los filtros ya establecidos anteriormente en el presente documento y al análisis consensuado por parte de los revisores.

Síntesis y Resultados

A continuación se presentan la síntesis de los estudios que obtuvieron como resultado del proceso de selección, el resumen presentado tiene como objetivo resaltar las características del Estado del Arte.

Multistep-ahead Streamflow and Reservoir Level Prediction Using ANNs for Production Planning in Hydroelectric Stations (Hernandez, Asqui, Arellano, & Cunalata, 2017): Los autores nos presentan el uso de redes neuronales utilizando bucles abiertos y cerrados para realizar la predicción del caudal y del nivel del embalse de una presa en Ecuador. Para esto utilizan como variables de entrada inicial: caudales históricos, energía activa generada histórica y el nivel del embalse (histórico y pronóstico). Como resultado del entendimiento de los datos identifican 3 temporadas climáticas (verano,

invierno ligero e invierno fuerte) los cuales son determinantes para la configuración de los parámetros las capas y neuronas del modelo implementado. Los resultados de este trabajo muestran que los niveles de error entre los valores pronosticados con relación a los reales disminuyen de manera considerable con horizontes pequeños de pronóstico (3 horas y 6 horas).

Comparison of multiple linear regression, artificial neural network, extreme learning machine, and support vector machine in deriving operation rule of hydropower reservoir (Niu , y otros, 2019): En este estudio los autores presentan utilizan cuatro métodos para elaborar reglas de operación del reservorio de una hidroeléctrica (China, reservorio Honghiadu). Para presentar un punto de referencia a los resultados de esta comparativa, se utilizará el método convencional utilizado en la central; el cual consiste en ejecutar la operación de la misma con base en los gráficos de datos históricos y la experiencia del operador. Los estudios referenciados en este trabajo dan cuenta que las ANN proveen resultados razonable con respecto a problemáticas relacionadas con recursos hídricos. De acuerdo a sus resultados los métodos de pronóstico que hicieron uso de algoritmos de inteligencia artificial entregaron mejores resultados que los métodos convencionales como las regresiones lineales; esto sobre todo a que el problema de manejo de recursos hídricos y sus variables no tiene una relación linear entre ambas. Los resultados incluyen una comparación incluso con el método basado en los gráficos de datos históricos y la experiencia de los operadores de la hidroeléctrica, y la comparativa muestra que este método no toma en consideración la naturaleza no linear de la problemática a tratar.

Hybrid metaheuristic of artificial neural network — Bat algorithm in forecasting electricity production and water consumption at Sultan Azlan shah Hydropower plant

(Hussin, Malek, Jaddi, & Hamid, 2016): Este trabajo de investigación realza la importancia del uso de las Hidroeléctricas como una alternativa limpia para la generación de energía. A partir de este punto propone la creación de un sistema de computación, que exhiba muestras de inteligencia artificial, con el objetivo de usar conocimiento para realizar la calendarización de la generación hidroeléctrica. El propósito de dicha calendarización es estimar la cantidad óptima de energía producida por las unidades generadoras de una central, para las siguientes N horas (24 horas para este caso) en el futuro. El uso de redes neuronales para este caso produjo un mínimo error en comparación con el uso de lógica difusa y razonamiento basado en casos o en conocimiento. Para corroborar la capacidad del modelo para la predicción, los autores utilizan una evaluación del error de los pronósticos. De igual manera se establece un punto de referencia para poder realizar la comparación de resultados; en este caso se usan: un nuevo modelo híbrido propuesto, un modelo de redes neuronales ya existentes y los valores históricos de producción de energía de la central. En los resultados nos muestra que hay una mejora utilizando el modelo hibrido con respecto al modelo original de redes neuronales; sin embargo el autor no presenta el pseudocódigo del modelo mejorado por cuestiones de propiedad intelectual.

ANN-based forecasting of hydropower reservoir inflow (Sauhats, Petrichenko, Broka, Baltputnis, & Sobolevskis, 2016): Este estudio parte con la premisa de que los métodos convencionales de predicción para reservorios de plantas hidroeléctricas, como los modelos de series temporales consideran que el caudal que alimenta al reservorio está auto correlacionado y que se comporta con las mismas tendencias a través del tiempo. Como resultado del uso de estas técnicas no se puede describir de manera correcta la naturaleza no lineal del comportamiento de caudales; necesidad por

la cual surge el uso de técnicas de inteligencia artificial para ser aplicados como solución a esta problemática. Una de las técnicas son las redes neuronales. Para probar que los resultados de pronóstico de las redes neuronales describen de mejor manera el comportamiento no lineal de un caudal, se toma como punto de referencia y comparación se los compara con resultados de técnicas de series temporales como son ARMA y ARIMA. Los autores utilizaron 4 enfoques para desarrollar los modelos de predicción en los que se usaron fórmulas para definir los parámetros de configuración de las redes neuronales y a su vez distintos enfoques de los datos de entrada de acuerdo a las variables históricas disponibles. Como conclusión indican que el uso de redes neuronales mejora la precisión de pronóstico con respecto a métodos de series temporales y más aún con relación a métodos basados en cálculos empíricos.

A Multi-Objective Stochastic Approach to Hydroelectric Power Generation Scheduling (Sauhats, Petrichenko, Baltputnis, Broka, & Varfolomejeva, A multi-objective stochastic approach to hydroelectric power generation scheduling, 2016): El presente estudio recalca la naturaleza no lineal y estocástica de las variables relacionadas a la programación de la generación de la energía hidroeléctrica. Y a partir de este punto propone enfoques de múltiples objetivos para optimizar la generación hidroeléctrica. Y es así que propone el uso de varias técnicas de optimización como: linear determinística, no linear estocástica y determinística con programación dinámica. Los problemas a tratar dentro de los objetivos planteados incluyen pronósticos de: precios de energía generada, caudal de entrada, demanda del mercado eléctrico y estrategia de operación horaria de la central. Para abordar cada uno de estos problemas los autores analizan y explican la razón que sustenta la aplicación de cada una de las técnicas citadas anteriormente. Y es precisamente esta actividad que se considera como el

principal aporte del este trabajo. Como caso de estudio se aplica este trabajo en la empresa H-GENCO de Brasil que pertenece al Mercado Eléctrico Brasileño, y que depende de una programación de energía generada precisa para poder optimizar los beneficios resultantes de su operación.

A Generalized Decision Support System for Short-Term Scheduling of China's Big Hydropower Systems (Shen & Cheng, 2015): En este trabajo se analizan las ventajas de un Sistema de Soporte a decisiones para la programación a corto plazo de la generación de Sistemas Hidroeléctricos. Destaca estudios anteriores de cómo el aparecimiento y desarrollo de Sistemas DSS ayudan a los operadores e ingenieros de las centrales hidroeléctricas a reducir cargas de trabajo y obtener mejores prácticas en las políticas de operación. Sin embargo también detalle los principales retos para la implementación de este tipo de sistemas, sobre todo en lo relacionado a la naturaleza de la cantidad de variables que influyen en la programación de generación. Se menciona variables básicas (unidades generadores, características de embalse), condiciones de operación (reservorio inicial, caudal) y restricciones (máximos y mínimos de almacenamiento en embalse, demanda de energía, capacidad de turbinas). Sin embargo de esto el principal aporte del trabajo consiste en relacionar la implementación de este tipo de Sistemas DSS con el patrón MVC. Es así que propone un mapeo entre MVC y el DSS, el cual a breves rasgos se define así: Model - Data Management Subsystem, Controller - Operation Subsystem y View - Result Management Subsystem. Se puede hacer una equivalencia de los componentes del DSS con el proceso general de implementación de un modelo predictivo de la siguiente forma: Proceso ETL - Data Management Subsystem, Implementación del Modelo Predictivo - Operation Subsystem y Validación del Modelo - Result Management Subsystem.

Forecasting Daily Streamflow Discharges Using Various Neural Network Models (Nacar, Hinis, & Kankal, 2018): De acuerdo a los estudios referenciados en este trabajo de investigación el 90% de casos de aplicación de hidrología han usado redes neuronales convencionales (feedforward), a saber con el estándar de perceptrón multicapa. Con esta premisa los objetivos de los autores se enfocan en hacer un comparativo de modelos de predicción basados en redes neuronales con distintas variaciones en cuanto a los métodos y a los algoritmos de aprendizaje. Distintos aspectos son abarcados por los tres métodos que se usaron para la elaboración del estudio; estos son: MLP-NN (Multilayer Perceptron Neural Network) que proponen el uso de varias capas ocultas en la red, con el objetivo de intervenir entre las variables de entrada y los resultados de salida de la red. PCA-NN (Principal Component Analysis Neural Network) que es usada para resolver problemas con varios parámetros, para lo cual plantea escoger un mínimo de parámetros principales de entrada a la red con un mínimo de pérdida de información. TLR-NN (Time Lagged Recurrent Neural Network) es uno de los métodos más usados en hidrología debido a su facilidad de uso, además que introduce retrasos de tiempo en la estructura de la red para ajustar sus valores durante el entrenamiento. Cada modelo además se lo relacionó con un solo algoritmo de aprendizaje para poder establecer los distintos modelos a implementar. Con los resultados de la aplicación de estos modelos, los autores presentan una comparativa estadística del desempeño de los modelos, esto a través del análisis del Error Medio, Error Cuadrático Medio y el Coeficiente de Correlación de las variables.

Conclusión del Estado del Arte

Basados en la síntesis de los estudios primarios podemos observar que la generación hidroeléctrica cada vez gana más espacio a nivel mundial, sobre todo por

ser la principal fuente de energía limpia y renovable. Por tal motivo es de suma importancia, en las centrales hidroeléctricas, el lograr el mayor aprovechamiento posible de los recursos hídricos. La herramienta utilizada en la Operación de las Centrales Hidroeléctricas es la programación de energía que puede generar cada una de ellas. En cada una de las Centrales Hidroeléctricas un objetivo imperante es poder predecir con una gran precisión los recursos hídricos con los que se podrá contar, y los cuales se podrán traducir en energía generada.

Existen varios métodos utilizados para la elaboración de este pronóstico, desde métodos basados en la experiencia de los operadores, aproximaciones basadas en la observación de histogramas de tendencias, hasta métodos lineales simples. Cada uno de estos métodos con diferentes resultados. Sin embargo la naturaleza no lineal de las variables que intervienen en pronósticos relacionados a la programación de energía, resulta en que la estimación basada en los métodos mencionados tenga un alto porcentaje de imprecisión, incluso si se usan combinaciones de uso entre estos métodos.

Los estudios analizados muestran que problemas de predicción no lineales, como la programación de energía hidroeléctrica, encuentran soluciones más precisas al ser tratados mediantes técnicas de aprendizaje computacional como: Support Vector Machine, Redes Neuronales, Lógica Difusa, etc.

Los trabajos analizados reportan predicciones más precisas con la utilización, y sobre todo con la automatización, de modelos predictivos basados en aprendizaje automático. Vale la pena señalar que el Deep Learning (simulación de comportamiento del cerebro humano) es una de las técnicas más utilizadas en la gran mayoría de estudios encontrados, esto se sustenta sobre todo por los niveles cada vez más bajos

de error en las predicciones entregadas por ésta técnica.

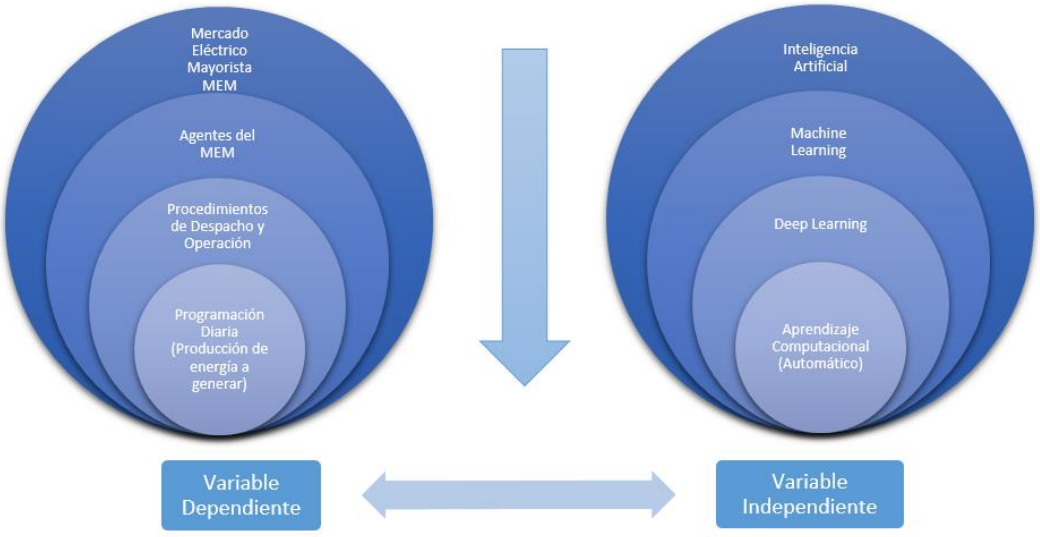
Finalmente el aporte del presente trabajo consiste en comprobar el uso de Modelos de Deep Learning (Redes Neuronales) para la predicción de series temporales univariantes; a diferencia de los estudios del grupo de control basados en modelos multivariantes. De igual manera es aporte se traduce en apoyo en los objetivos estratégicos de la CHCCS para la predicción y futuro aprovechamiento (umbral de 24 horas) de los recursos hídricos disponibles para la generación de energía eléctrica.

Marco Teórico

A través de la definición de las variables que intervienen en la problemática, se establece la jerarquía de estudio. Mediante ésta jerarquía se procede a desarrollar la categorización en la Figura 4, tanto de la variable dependiente como independiente.

Figura 4

Variable dependiente e independiente



Nota: Elaboración Propia, 2020

Categorización de la Variable Dependiente

Mercado Eléctrico Mayorista. En el Ecuador toda la energía eléctrica es comercializada a través del MEM (Mercado Eléctrico Mayorista), el mismo que se define como:

"El mercado integrado por generadores, transmisores, distribuidores y grandes consumidores, donde se realizan transacciones de grandes bloques de energía eléctrica. Así mismo incluye la exportación e importación de energía y potencias eléctricas"

Agentes del MEM. Los agentes que integran el MEM se definen así (CONELEC, 2016):

Generador. "Personal natural o jurídica, titular de una concesión, permiso o licencia para la explotación de una o varias centrales de generación eléctrica."

Transmisor. "Empresa titular de la concesión para la prestación del servicio de transmisión y la transformación de la tensión vinculada a la misma, desde el punto de entrega por un generador o auto productor, hasta el punto de recepción por distribuidor o gran consumidor."

Distribuidor. "Es la persona natural o jurídica titular de una concesión para la prestación del servicio público de distribución de energía eléctrica por virtud de la cual asume la obligación de prestar el suministro de electricidad a los consumidores finales ubicados dentro del área respecto de la cual goza de exclusividad regulada."

Gran Consumidor. "Agente de MEM, debidamente calificado por el CONELEC por sus características de consumo, que está facultado para acordar libremente con un generador o distribuidor el suministro y precio de la energía eléctrica."

Procedimientos de Despacho y Operación. Tomando en cuenta la base

conceptual de funcionamiento del MEM, el directorio del CONELEC ha establecido los Procedimientos de Despacho y Operación. Los cuales tienen como objetivo prioritario proporcionar una normativa y base metodológica para la correcta, oportuna y eficaz aplicación del Reglamento de Despacho y Operación. Dentro de los Procedimientos definidos se encuentra la Planificación de la Operación cuyo objetivo es “establecer una política óptima de la operación de los embalses y uso eficiente de los recursos disponibles de generación… teniendo en cuenta además, la previsión de las demandas y la aleatoriedad de la oferta y los caudales.” (ARCONEL, 2000).

Programación Diaria. Uno de los componentes de la Planificación Operativa es la Programación Diaria la cual se define como: “Proceso mediante el cual se obtiene para un período de 24 horas el programa horario de generación de los recursos del MEM…”. Esto se logra a través del cálculo del despacho horario de generación, que es elaborado por el CENACE y comunicado a los agentes del MEM.

Para poder realizar éste cálculo el CENACE estableció como responsabilidad de los agentes generadores la obligación de suministrarle de manera diaria y hasta las 10:00 horas toda la información necesaria a través de medios electrónicos. Ésta información debe incluir para el caso de Centrales Hidroeléctricas de Pasada (como lo es la CHCCS) la Curva de generación horaria prevista para el día siguiente. Esta curva es conocida también como la Programación de la Energía a producir por la Central, para nuestro caso de estudio programación diaria (ARCONEL, 2000).

Categorización de la Variable Independiente

Inteligencia Artificial. La inteligencia artificial (IA) no es un concepto nuevo ya que nace alrededor de los años 50 como una rama de la Informática. Su desarrollo estuvo enfocado en desarrollar computadores que posean la habilidad comportarse de

manera tal que éste mismo comportamiento lo realizara un ser humano (Torra, 2011). En un contexto más actual podemos colegir que la IA nos da las herramientas para poder llevar a cabo análisis sobre la gran cantidad de información, y poder utilizar el conocimiento extraído en diversos campos (Navarro, 2017).

De manera general podemos destacar, de acuerdo a sus objetivos, la siguiente clasificación de la IA:

- IA fuerte: Considera que un computador puede llegar a poseer una mente propia, e incluso que dicha mente pueda tener varios estados; lo que deriva en la posibilidad de construir un computador con las capacidades humanas en toda su extensión.
- IA débil: los adeptos a esta corriente sostienen que los computadores tan solo pueden simular un razonamiento, y que solo actúan de forma inteligente. Para poder realizar esta simulación los computadores hacen uso de algoritmos y métodos de aprendizaje como el Machine Learning por ejemplo (Rodríguez, 2020).

Machine Learning. Dentro del ámbito de la IA, Machine Learning (ML) es la disciplina que se encarga del desarrollo de sistemas que tengan la capacidad de aprendizaje automático. Es de esta forma que a través del auto aprendizaje un sistema es capaz de identificar: patrones complejos en millones de datos, predicción de comportamientos a través del uso de algoritmos en incluso tener la habilidad de mejorarse a sí mismo de forma independiente a través del tiempo (Castro, 2017).

Para el ML la función más básica consiste en el uso de algoritmos que nos permitan procesar datos y extraer patrones por medio de algoritmos, y entonces ser capaces de entregar predicciones o sugerencias sobre los mismos datos (Rodríguez, 2020).

Deep Learning. Dentro del aprendizaje automático, el cual forma parte del ML, el Deep Learning (DL) es la técnica que busca lograr aprendizaje a través del ejemplo. Esto es posible a través de modelos informáticos que pretenden funcionar de manera similar al comportamiento del cerebro humano.

Los modelos desarrollados son capaces de evaluar distintos ejemplos e instrucciones para modificarse a sí mismo en caso de encontrar errores en sus resultados. De esta forma los modelos desarrollados a través del aprendizaje buscan el poder anticiparse a los errores que puedan producirse por medio de la extracción de patrones (Navarro, 2017).

Aprendizaje Automático. Lograr que un computador, sistema o modelo obtenga la capacidad de predecir y tomar decisiones correctas en distintas circunstancias, basados en datos previos y sin programación explícita de programación humana es lo que definimos como aprendizaje automático (computacional) (Hewlett Packard, s.f.).

Existen tres técnicas principales de aprendizaje automático: (Kent, s.f.)

- Supervisado: dependemos de datos previamente etiquetados, se busca que el computador aprenda de los ejemplos ya resueltos.
- No supervisado: No existen etiquetas previas en los datos a analizar y es el sistema el que debe etiquetar los datos a partir de la información que recolecta.
- Por refuerzo: El sistema de forma progresiva es capaz de aprenden en base a pruebas y errores.

Redes Neuronales Artificiales

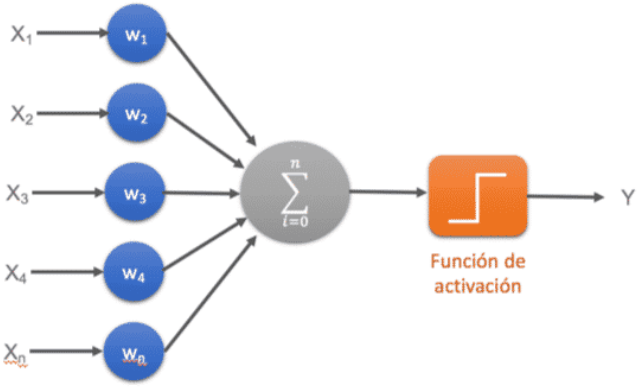
Entre las formas más comunes de implementar Deep Learning se encuentran las redes neuronales artificiales. Las redes neuronales se las define como un modelo

basado en el funcionamiento del cerebro humano. De manera general están formadas por un conjunto de nodos que se conocen como las neuronas, las cuales se conectan entre sí para trasmitir información; la información necesita tener un punto de entrada y entonces esta se procesa hasta generar el resultado de dicho procesamiento a través de una salida. Las conexiones entre las neuronas se forman a través de enlaces, los cuales entregan la información que reciben de la neurona predecesora a la sucesora multiplicada por un peso. También a la información que entrega una neurona a cada enlace puede ser modificada a través de las conocidas funciones de activación, la cual tiene como objetivo delimitar el rango que puede tomar el valor de salida entregado por la neurona. (Red neuronal artificial, 2020)

Funcionamiento De Las Neuronas. Podemos describir a la neurona como un repositorio al cual llegan datos junto a un peso definido, estos valores son usados por la neurona para hacer un cálculo matemático que pueden incluir también un tercer valor llamado sesgo como se ilustra en la Figura 5.

Figura 5

Diagrama de una Neurona Artificial



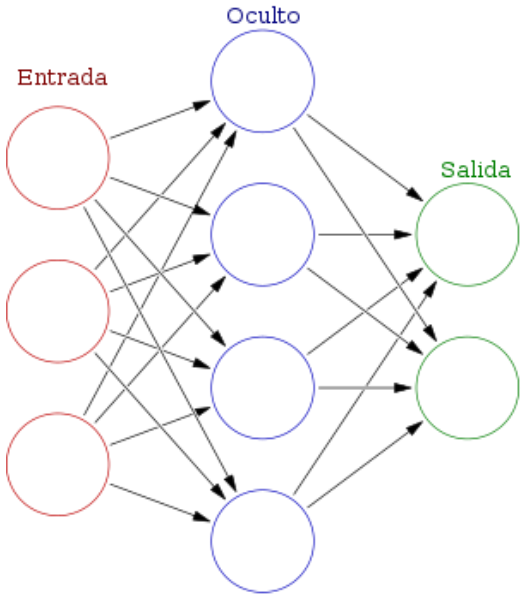
Nota: Tomado de (Calvo, 2017)

El resultado de este proceso se convierte en la salida que entregará la neurona, esta salida se une generalmente con salidas de más neuronas y se convierten en entradas para otra neurona que procesa estos datos y genera una nueva salida.

Capas de una Red Neuronal. Cuando hablamos de una red neuronal nos referimos a que existan interconexiones entre las neuronas pertenecientes a la red. Estas neuronas se ordenan a través de capas, en un modelo simple encontramos 3 capas como se puede ver en la Figura 6.

Figura 6

Modelo Simple de Red Neuronal



Nota: Tomado de (Red neuronal artificial, 2020)

Capa de Entrada. En esta capa se encuentran las neuronas que reciben los datos precedentes del entorno.

Capa Oculta. Recibe los datos que entregan las neuronas de otra capa oculta o de la capa de entrada.

Capa Salida. En esta capa se encuentran las neuronas que procesan la información por última vez para entregar el resultado final.

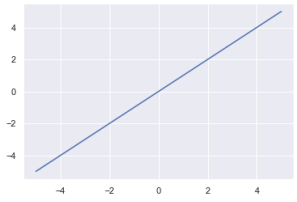
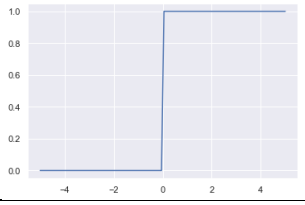
Dependiendo de la cantidad de neuronas y de las capas ocultas que contenga una red neuronal más esfuerzo tomará su procesamiento pero también aumentará la capacidad de resolver problemas más complejos.

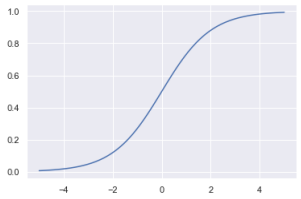
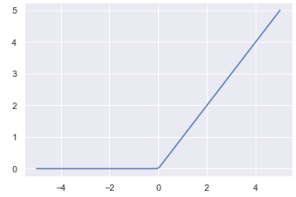
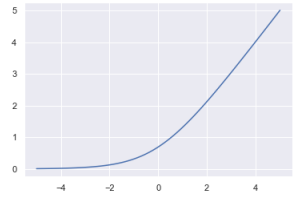
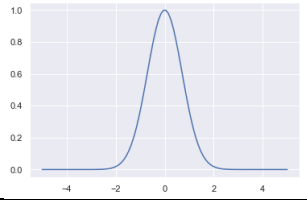
Función de Activación. Al igual que las neuronas biológicas las neuronas de una red artificial no transmiten los datos tal como los recibe. Para esto se utiliza la función de activación que se encarga de devolver una salida a partir de un valor de entrada. Se tiene una función de activación para cada capa que forma la red neuronal.

De manera general las funciones de activación se dividen en dos tipos: lineales, no lineales y de umbral (Google Developers, 2020), un resumen de las principales funciones de activación se muestran en la Tabla 4.

Tabla 4

Principales Funciones de Activación.

NOMBRE	TIPO	ECUACIÓN	GRÁFICA
Identidad	Lineal	$f(x) = x$	
Escalón	Umbral – No Lineal	$f(x) = \begin{cases} 0, para\ x < 0 \\ 1, para\ x \geq 0 \end{cases}$	

NOMBRE	TIPO	ECUACIÓN	GRÁFICA
Sigmoidea o Logística	No Lineal	$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$	
(ReLU) Rectified Linear Unit	No Lineal	$f(x) = \begin{cases} 0, para\ x < 0 \\ x, para\ x \geq 0 \end{cases}$	
SoftPlus	No Lineal	$f(x) = \ln(1 + e^{x})$	
Gausiana	No Lineal	$f(x) = e^{-x^2}$	

Nota: Elaboración Propia, 2020

Funciones Lineales. La salida que se entrega es proporcional a la entrada.

Funciones No Lineales. La salida no es proporcional a la entrada.

Funciones de Umbral. La salida se entrega a través de valores discretos, generalmente binario (uno o cero).

Aprendizaje de las Redes Neuronales. Al igual que su equivalente biológico las redes neuronales artificiales aprenden a través de la repetición, y para esto la cantidad de datos disponibles para entrenar la red y las veces que la red sea entrenada (épocas) son directamente proporcionales a la mejora de los resultados entregados por

la red. De acuerdo con esto para que una red neuronal pueda aprender es necesario que se le provea de una retroalimentación, para el caso de las redes neuronales este proceso se conoce como propagación hacia atrás.

El objetivo de la propagación hacia atrás es que la red sea capaz que optimizar los pesos definidos para cada conexión entre sus neuronas, y así las salidas que entregue la red sean lo más cercanas posibles a las salidas esperadas. (Backpropagation - El corazón de las redes neuronales, s.f.)

De manera general los pasos de la propagación hacia atrás, en cada época de entrenamiento, son los siguientes:

Pasos hacia adelante:

1. Se elige los datos que serán la entrada para la red neuronal.
2. Se ingresa la entrada elegida a la red neuronal y se calcula su salida.

Pasos hacia atrás:

3. Se calcula el error existente entre la salida deseada y la calculada por la red.
4. Se ajustan los pesos de las conexiones para que el error existente entra la salida obtenida versus la salida deseada disminuya.
5. Se repite el proceso para todos los datos que conforman el conjunto de valores con los cuales se va a entrenar la red neuronal hasta que el error obtenido se encuentre en un rango aceptable.

Función de Coste. Mientras entrenamos a la red tenemos los valores que se ingresan como entradas y a su vez las salidas que éstas generan y lo que buscamos es ser capaces de ajustar los pesos que utiliza la red; de esta manera la red aprende y las salidas se acercan lo más posible a los resultados deseados. La función de coste es que determina el error entre la salida obtenida con respecto al resultado deseado y a

partir de este cálculo optimizar los parámetros de la red neuronal, entre ellos lo pesos asignados.

Existen varias opciones para medir el error al que nos referimos anteriormente, los más utilizados, para modelos de predicción de valores son: (Data TechNotes, 2019)

- **Error Cuadrático Medio (Mean Squared Error - MSE)**: representa el promedio de las diferencias entre la salida obtenida y la salida deseada.*

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

- **Raíz del Error Cuadrático Medio (Root Mean Squared Error - RMSE)**: Simplemente es el resultado de extraer la raíz del MAE visto anteriormente.*

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- **Error Absoluto Medio (Mean Absolute Error MAE)**: Representa el promedio del valor absoluto de la diferencia entre la salida obtenida y la salida deseada.*

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

- **Error Absoluto Medio Porcentual (Mean Absolute Percentage Error - MAPE)**: Es el promedio de los errores porcentuales absolutos expresado en términos porcentuales.*

$$\text{MAPE} = \sum_{\text{t=1}}^{\text{n}}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

- **R Cuadrado**: También conocido como coeficiente de determinación representa

que tanto se ajustan los valores obtenidos con respecto a los esperados. Mientras más alto sea el valor mejor es el modelo.*

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

* $\hat{y}$ - Valor obtenido o predicho de y / $\bar{y}$ - Valor Medio de y

Optimizadores de Redes Neuronales. Cuando se entrena una red nos planteamos como objetivo encontrar los pesos apropiados para las conexiones de la red, esto se consigue minimizando la función de coste escogida. Los optimizadores nos ayudan a minimizar la función de coste, esto lo hace generando pesos cada vez más ajustados. De forma general los optimizadores basan su funcionamiento en el cálculo del gradiente de la función de coste (derivada parcial) para cada uno de los pesos, y con el objetivo de minimizar dicha función se modifican los pesos en la dirección negativa del gradiente. Es decir que el optimizador modifica los valores de los pesos de forma que la función de coste disminuya y se acerque al mínimo, la forma de saber que ha encontrado ese mínimo es con el gradiente (derivada parcial). (Velasco, 2020) A continuación se explica de forma resumida los optimizadores de la función de coste más utilizados.

- **SGD (Stochastic Gradient Descent):** Se escoge de manera aleatoria el conjunto de datos para calcular el gradiente y actualizar los pesos con lo que buscamos una convergencia mucho más rápida.
- **RMSprop (Root Mean Square prop):** Utiliza una media móvil de los cuadrados del gradiente para obtener el factor que influye en el cambio de los pesos.
- **Adaptive Gradient Algorithm (AdaGrad):** El algoritmo AdaGrad considera un

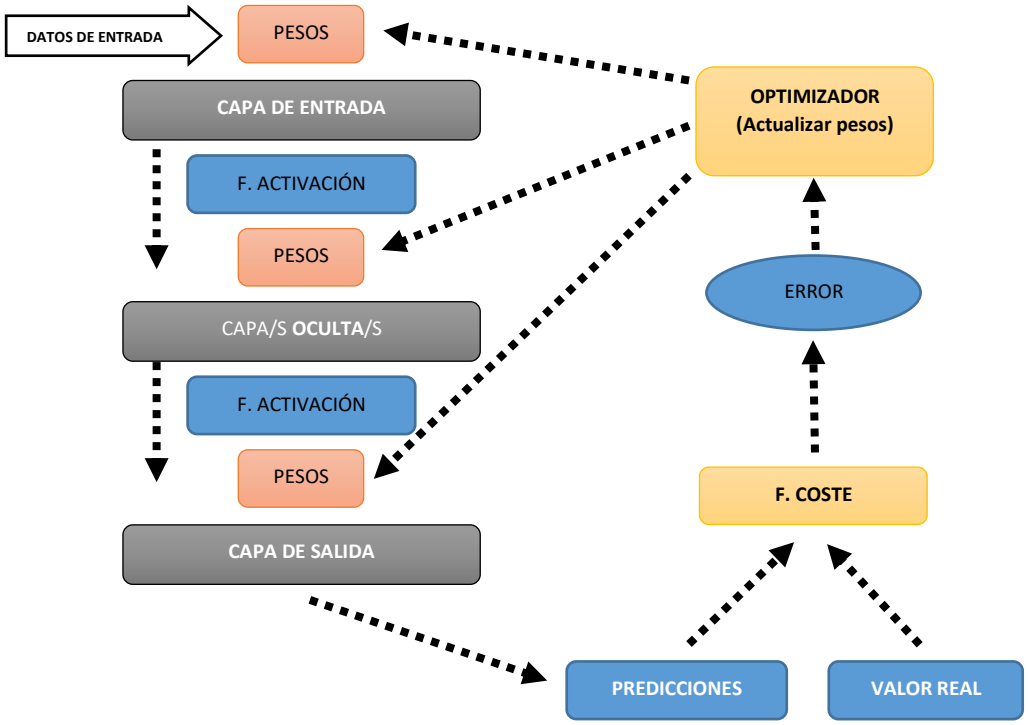
factor de cambio para cada uno de los pesos de la red y para ello parte de un factor inicial.

- **Adaptive Adam (Adaptive moment estimation):** Este algoritmo combina los conceptos de AdaGrad y RMSProp. Calculamos un factor de entrenamiento para cada peso y también utiliza el cálculo de RMSProp.

Utilizando los conceptos principales recogidos en el marco teórico podemos elaborar un esquema general del funcionamiento y aprendizaje de un modelo basado en redes neuronales, como se ilustra en el Figura 7; este esquema será el punto de partida para la implementación del modelo propuesto para la resolución de la problemática expuesta.

Figura 7

Esquema general de implementación y aprendizaje de una Red Neuronal



Nota: Elaboración Propia, 2020

Capítulo III: Modelo de Predicción de la Producción de Energía de la CHCCS, basado en Redes Neuronales

El Modelo a implementarse será resultado de la ejecución de la Metodología de Desarrollo Híbrida propuesta. Las fases que comprenden ésta Metodología son:

- Definición del Objetivo del Negocio
- Construcción del Modelo
- Implementación del Modelo

En el presente capítulo se abordará hasta la tarea Modelamientos Finales de la fase de Construcción del Modelo, y en el siguiente se detallará la tarea correspondiente a la Evaluación de los Modelos construidos y el despliegue del Modelo final elegido para dar solución a la problemática planteada.

Definición del Objetivo del Negocio

Comprensión del Negocio

Parte de la comprensión del negocio para el presente trabajo fue abordado en la sección “El Problema de Investigación”, la cual comprende el contexto y planteamiento del problema. Con esta base se pudo traducir el conocimiento adquirido en un problema a ser resuelto a través de Data Mining, en nuestro caso específico un Modelo Predictivo. Para cada una de las tareas de que componen la fase Comprensión del Negocio se realizó el respectivo análisis.

Determinación de los Objetivos del Negocio. El objetivo del negocio se centra en la generación eficiente de energía hidroeléctrica en la CHCCS, uno de los factores necesario para alcanzar este cometido es la optimización del uso de los recursos hídricos captados del Río Coca.

Esto está estrechamente relacionado con la capacidad de predecir a ciertos

umbrales futuros el caudal con el que la CHCCS puede contar para producir energía de acuerdo a los objetivos definidos para la central por parte de entes como el CENACE.

Evaluación de La Situación. Actualmente el personal técnico de la CHCCS cuenta con información limitada con respecto a las variables que intervienen en la predicción del caudal del Río Coca. La única fuente de información, de las variables mencionadas, está compuesta por un registro histórico desde el año 1972 del promedio diario del caudal del Río Coca. Con estos datos se realizan predicciones a través de métodos lineales (ARIMA) que se conjugan algunas ocasiones con valores resultantes de la experiencia previa del personal que ha operado tanto la CHCCS como otras centrales hidroeléctricas de CELEC EP.

Determinación de los Objetivos del Data Mining. Los objetivos de minería de datos propuestos son:

- Implementar opciones de Modelos Predictivos basados en redes neuronales.
- Entrenar los modelos para afinamiento de su arquitectura, función de activación, función de coste, etc.
- Evaluar los Modelos propuestos de acuerdo a la precisión de las predicciones con respecto a los valores reales.
- Definir el Modelo con mejores resultados para ser desplegado como solución propuesta al problema planteado.

Realizar el Plan de Proyecto. Esta tarea sufre cierta modificación debido al planteamiento de la Metodología Híbrida propuesta en la cual se definió que la ejecución del proyecto de Data Mining se lo ejecutará a través de sprints, siguiendo la lógica de la metodología SCRUM.

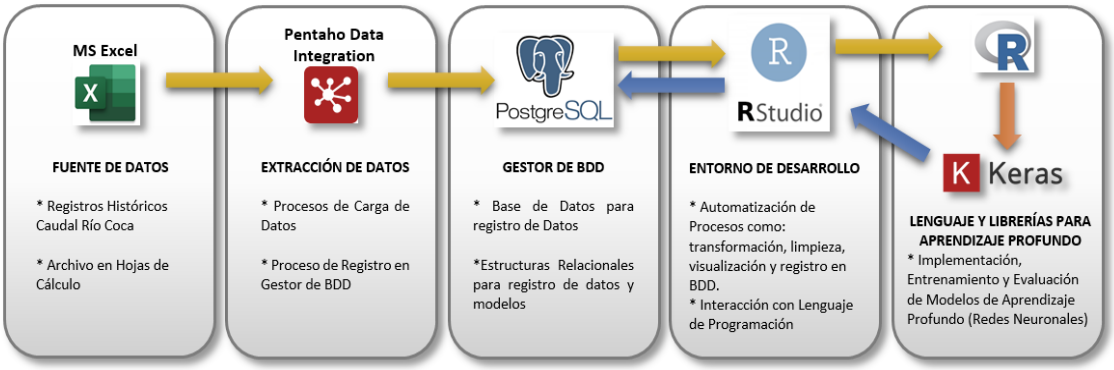
Estudio y Comprensión de los Datos

A través de contacto con personal de la CHCCS se pudo tener una idea inicial de la cantidad y calidad de los datos de la fuente de información mencionada en la fase anterior. De forma específica las tareas que comprenden la fase de comprensión de los datos fueron realizadas para el registro histórico del caudal promedio del Río Coca desde enero del 1972 a marzo del 2020, como se explica a continuación:

Implementación de la Arquitectura Técnica. El principal objetivo de esta tarea consistió en implementar la arquitectura técnica, que permitió realizar la primera iteración de modelamiento y las subsiguientes iteraciones de este proceso. Para lograr este cometido, y de manera complementaria a las herramientas ya descritas anteriormente, se realizó la instalación de la librería Keras para su ejecución en R. Adicional a esta librería se añadieron también otros paquetes adicionales de tratamiento y visualización de datos como tidyverse y ggplot2 respectivamente. En la Figura 8 se ilustra un diagrama de la arquitectura técnica implementada, donde se puede apreciar como interactuaron las herramientas tecnológicas descritas hasta este punto.

Figura 8

Diagrama de Arquitectura Técnica para Construcción de Modelos



Nota: Elaboración Propia, 2020

Recolección de los Datos. La información en la que se basa el modelo de predicción fue extraída de los registros históricos con los que cuenta el área de Operación de la CHCCS. Previo análisis y consulta al personal del área mencionada, se determinó que la principal fuente de datos, con la que es factible identificar patrones que nos permiten realizar una tarea predictiva, corresponde a los valores históricos del caudal del Río Coca. La factibilidad responde tanto a la cantidad como la calidad de los datos de la fuente identificada, características primordiales que están directamente ligadas a la probabilidad de que el o los modelos implementados sean más fiables y ajustados. (DECIDE SOLUCIONES S.L., 2019). Los registros solicitados de manera formal a CELEC EP UNIDAD DE NEGOCIO COCA CODO SINCLAIR corresponden a los valores del promedio diario del caudal del Río Coca, dentro del período comprendido entre enero del año 1972 hasta marzo del año 2020, los cuales se encuentran en formato de archivo Microsoft Excel.

Descripción de los Datos. Los valores que componen el archivo tomado como fuente de datos tienen las siguientes características:

- **Número de Datos:** 17623 mediciones.
- **Tipo de Dato:** Decimales con precisión 2.
- **Unidad de Medida:** m^3/s (metros cúbicos por segundos).
- **Periodicidad:** Se cuenta con el promedio diario del período enero 1972 hasta marzo 2020, el formato de la fuente para las mediciones está estructurado por años, meses y días. También se incluyen promedios mensuales y anuales.
- **Estructura:** Los datos tienen siguientes encabezados como se muestra en la Figura 9.

Figura 9

Ejemplo de estructura de Fuente de Datos de Caudal del Río Coca

RIO COCA EN SITIO DE TOMA SALADO

CAUDALES MEDIOS DIARIOS MENSUALES Y ANUALES (m^3/s)

AÑO: 1972

DIA	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1	503,5	161,9	160,8	483,8	217,5	311,7	747,0	713,3	478,4	229,1	218,3	280,4
2	673,0	205,3	163,5	413,6	257,8	329,1	973,3	475,1	492,6	215,9	188,7	264,0

Nota: Elaboración Propia, 2020

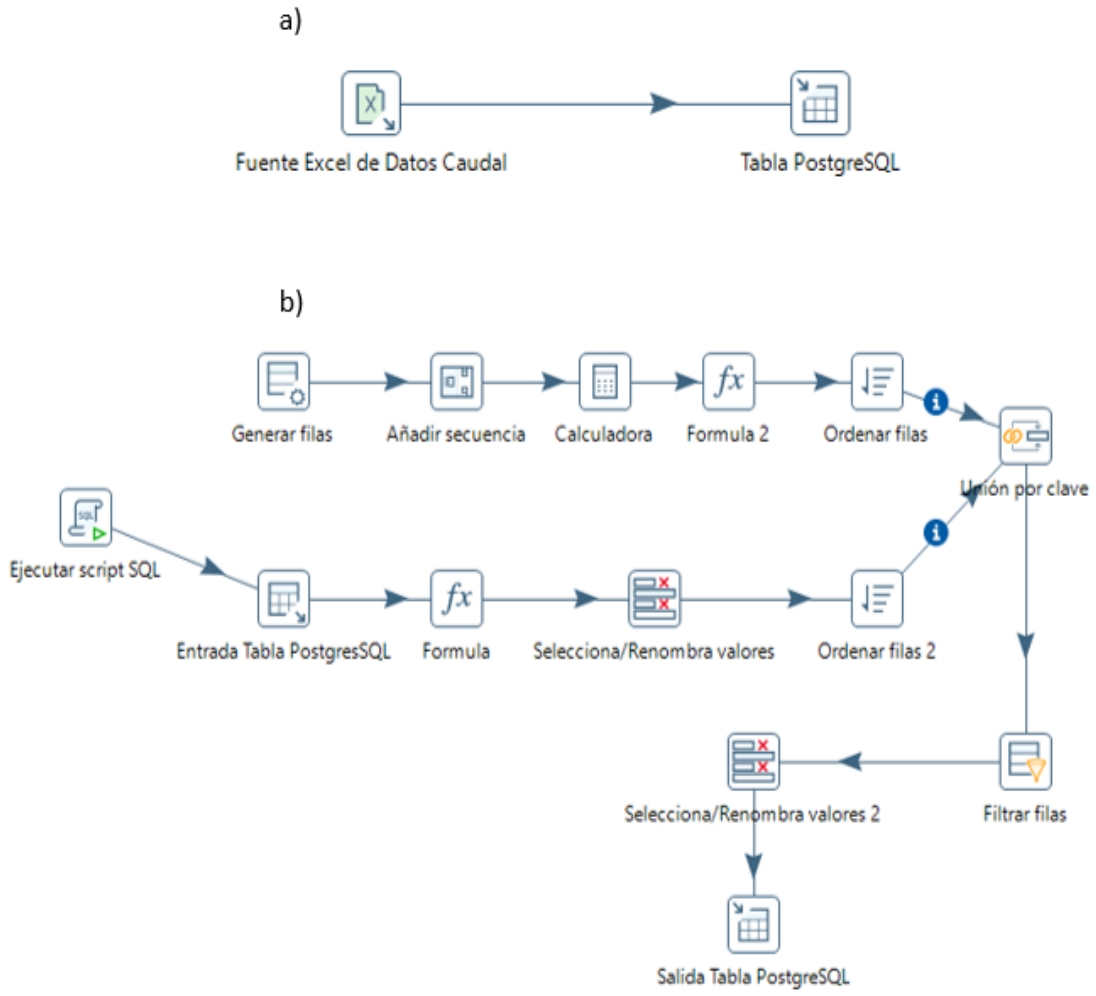
Exploración de Datos. Previo a la ejecución de la tarea de exploración se inició con la implementación de la arquitectura técnica que servirá de plataforma para la construcción del Modelo Predictivo final. En este punto se desarrollaron procesos ETL (extracción, transformación, carga) para poder migrar los registros hacia una base de datos relacional, esto con el objetivo de poder realizar la exploración de datos de una forma más eficiente y que nos permita registrar de manera permanente los resultados de las ejecuciones de las tareas futuras.

Tanto para los procesos de carga, así como para el repositorio relacional se utilizaron herramientas de acceso libre: Pentaho Data Integration 8.2 y PostgreSQL 12 respectivamente.

En la Figura 10 se muestran los dos trabajos de carga desarrollados en la herramienta Pentaho Data Integration: “CARGA TABLA CAUDALES_CCS” y “TRANSFORMACIÓN TABLA CAUDAL_PROM_DIA”, el esquema de los trabajos de migración se muestran.

Figura 10

Diagramas de Trabajos de Migración



Nota: a) “CARGA TABLA CAUDALES_CCS” y b) “TRANSFORMACIÓN TABLA CAUDAL_PROM_DIA”, Elaboración Propia (Pentaho Data Integration 8.2), 2020

Los trabajos de migración descritos registran la información de manera permanente en tablas relacionales implementadas sobre el Gestor de Base de Datos

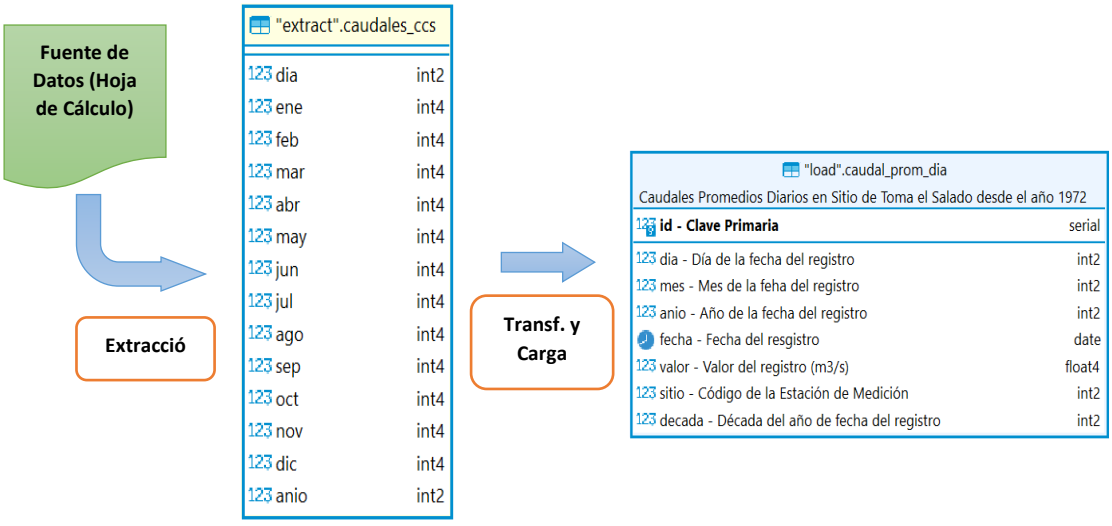
PostrgreSQL. El trabajo de migración “CARGA TABLA CAUDALES_CCS” realiza un proceso de carga manteniendo el mismo formato de dato de los registros del archivo Excel hacia la tabla: “extract.caudales_ccs”.

Mientras que el trabajo de migración *“TRANSFORMACIÓN TABLA CAUDAL_PROM_DIA”* toma los datos de la tabla “extract.caudales_ccs” y transforma los datos hacia una estructura de tabla relacional normalizada para registrarla en la tabla: “load.caudal_prom_dia”.

En la Figura 11 se muestra en diagrama conceptual de las tablas utilizadas para registrar los resultados de los trabajos de migración desarrollados.

Figura 11

Diagrama Conceptual de las Tablas del proceso de ETL de Datos



Nota: Tablas: “extract.caudales_ccs” y “load.caudal_prom_dia”, Elaboración Propia (DBeaver), 2020

Una vez disponibles los valores del caudal histórico, desde la base de datos de PostgreSQL, a través de las tablas descritas anteriormente se utilizaron el lenguaje de programación R y el entorno de desarrollo RStudio con el fin de realizar el análisis exploratorio de los datos.

Se obtuvo las medidas de estadística descriptiva de la variable correspondiente al caudal promedio diario histórico, los valores fueron calculados a través de la función summary(data) del software RStudio. Estos valores son recogidos en la Tabla 5.

Tabla 5

Estadística Descriptiva del Caudal promedio histórico del Río Coca

MEDIDA DESCRIPTIVA	VALOR
Mínimo	20,00
Primer Cuartil	180,90
Segundo Cuartil / Mediana	249,60
Media	291,60
Tercer Cuartil	350,90
Máximo	2329,70

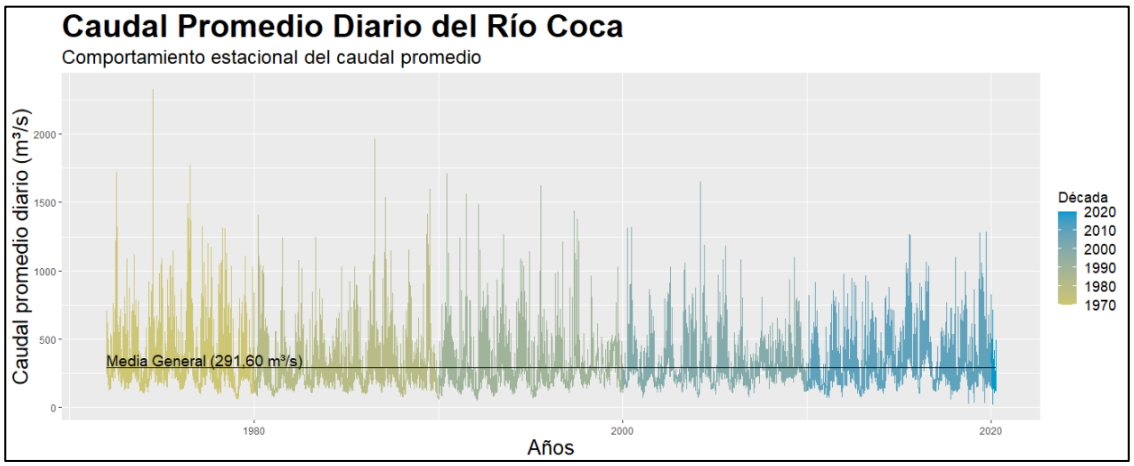
Nota: Elaboración Propia, 2020

Para poder establecer una idea inicial del comportamiento de los datos se utilizó un gráfico de línea, el cual relaciona el valor del caudal promedio histórico desde enero de 1972 hasta marzo del 2020 además el gráfico incluye como referencia el valor de la media de la serie de datos (291,60 m^3/s).

En la Figura 12 podemos observar un comportamiento del caudal histórico que presenta un comportamiento bastante idéntico en cada una de las décadas marcadas en la Figura 12.

Figura 12

Gráfico de Línea: Caudal Promedio Histórico del Río Coca



Nota: Período de los datos: (enero 1972 – marzo 2020), Elaboración Propia (RStudio), 2020

Verificación de la Calidad de los Datos. El principal objetivo de esta tarea es la de verificar la consistencia de los valores de cada dato, para esto se analiza la presencia de datos nulos o atípicos. La presencia de estas inconsistencias si no se tratan previamente son capaces de producir que el proceso predicción a través de cualquier modelo resulta fallido.

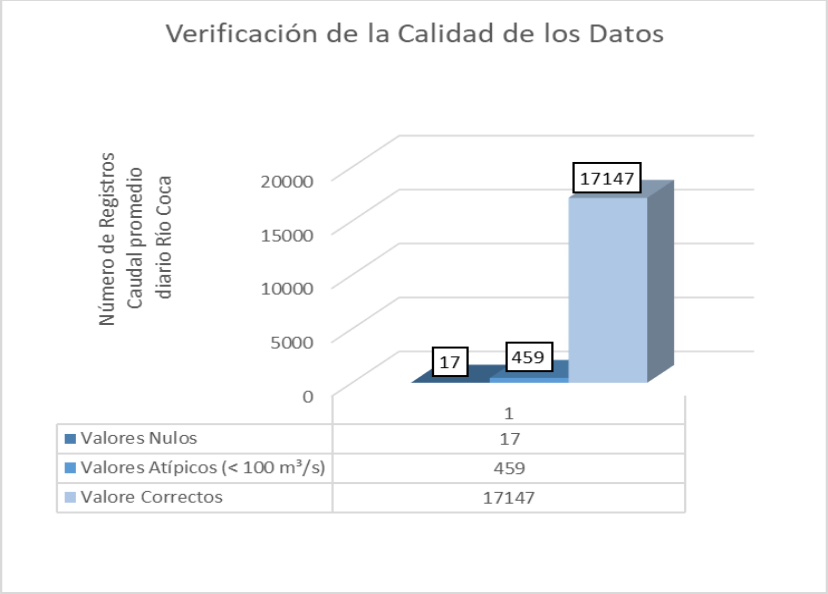
Con respecto a los valores nulos se identificaron 17 registros de un total de 17623. En cuanto a los valores atípicos, se pudo establecer que las mediciones menores a los 100 m³/s corresponderían a valores por debajo del caudal firme diario (caudal disponible incluso en la estación seca más crítica) del Rio Coca.

Este dato fue proporcionado por el personal de hidrología de la CHCCS basada en los estudios de factibilidad realizados previo a la construcción de la misma. En total

459 presentaban un valor menor a los 100 m³/s. En la Figura 13 se presenta el gráfico de la Verificación de Calidad de los datos.

Figura 13

Gráfico de Barra: Verificación de la Calidad de los Datos



Nota: Elaboración Propia, 2020

Preparación de los Datos

Para poder adaptar los datos recolectados a una determinada técnica de Data Mining es necesario realizar un análisis más específico a partir de los hallazgos de la sección: Exploración de Datos. Con este antecedente podemos evidenciar que los datos del caudal promedio diario histórico del Río Coca corresponden a una serie temporal. Cumplen con el concepto de serie temporal, al ser una colección de observaciones de una determinada variable que han sido obtenidas de forma secuencial y en períodos equidistantes a través del tiempo.

A partir de esta premisa realizaremos el análisis de los datos de acuerdo a los aspectos relevantes para las series temporales.

Una serie temporal se puede clasificar de manera general en:

- **Estacionarias:** este tipo de series temporales conservan estables sus propiedades estadísticas. Esto quiere decir que su media, varianza y covarianza permanecen constantes durante el tiempo.

- **No estacionarias:** contrario a lo anterior para este tipo de series temporales las propiedades estadísticas de la serie varían a través del tiempo. Es decir que encontramos una tendencia que hace que su media cambie a través del tiempo.

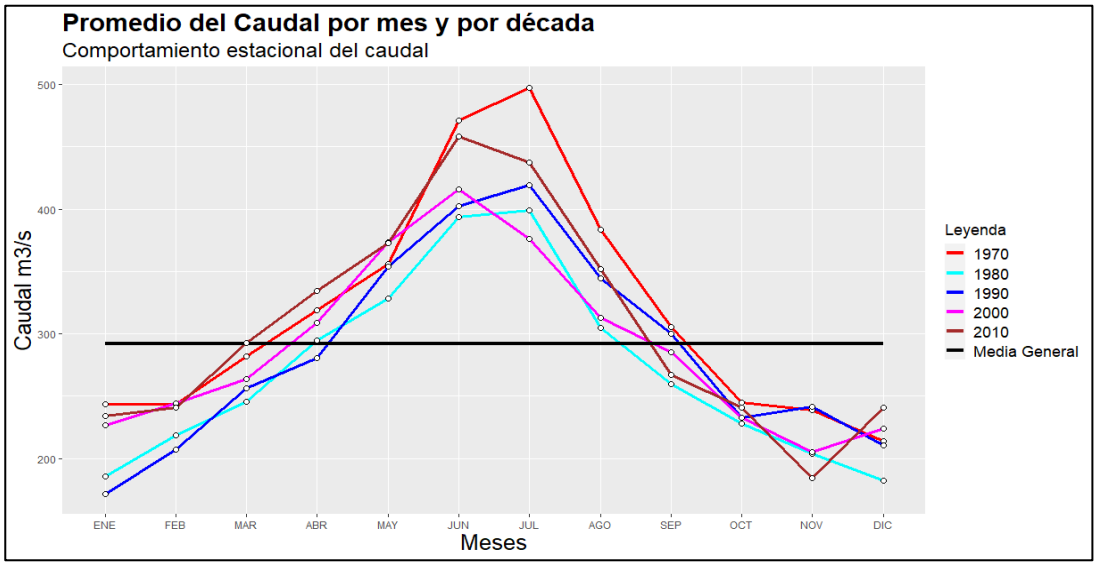
Para nuestro caso específico, en el que el fenómeno planteado como problemática corresponde a un proceso estocástico (predicción de caudales), nos encontramos en el segundo tipo de serie temporal es decir no estacionaria.

Esto se comprueba de manera directa al encontrar el componente de tendencia dentro de la serie temporal; la misma que causa que, dependiendo de ciertos períodos de tiempo la media de los datos varíe aumentando o disminuyendo su valor; lógicamente estos períodos corresponden a las temporadas con mayor o menor número de precipitaciones así como de la temperatura en el ambiente.

En la Figura 14 se observa a través de los valores promedios por mes en cada década la tendencia que presenta el comportamiento del caudal histórico; sobreponiendo todas las series de datos se comprueba que este patrón de aumento y disminución de caudal es similar para todas las décadas analizadas.

Figura 14

Promedio del Caudal del Río Coca por mes y por década



Nota: Elaboración Propia (RStudio), 2020

Selección de los Datos. La capacidad de implementar modelos más fiables está relacionado con una cantidad suficiente de datos a partir de los cuales se puedan entrenar estos modelos a través de técnicas de predicción como el aprendizaje profundo (redes neuronales).

Sin embargo estos datos deben ser de calidad lo que suele incluir que sean correctamente organizados y filtrados. (DECIDE SOLUCIONES S.L., 2019)

Para la construcción de nuestro Modelo de predicción, y basado en los resultados de la sección Verificación de los Datos, se seleccionaron el total de registros disponibles de la fuente de datos. En las tareas siguientes se ejecutaron procesos para asegurar la calidad, a través del tratamiento específico de los datos nulos y valores atípicos reportados.

Limpieza de los Datos. En esta tarea se realizó el tratamiento de los valores atípicos y nulos encontrados en el conjunto de datos seleccionados.

Para ello a través de Rstudio se desarrolló un script (tratamiento_datos.R) que de manera general realizan las siguientes acciones para el tratamiento de los grupos de datos mencionados:

- Tratamiento de valores atípicos: Ya identificado que los registros que muestran un valor de caudal menor a los 100 m^3/s, se optó por actualizar todos estos 459 registros al valor de 100 m^3/s.
- Tratamiento valores nulos: Tomando los 17 valores nulos identificados, se ordenan desde las fechas más antiguas a las más recientes. A partir de esto para cada valor nulo se extrae la fecha, y se lo actualiza con la media aritmética de todos los registros que coinciden con el mismo día y mes pero tomando sólo los años anteriores al año para el que se tiene el registro nulo. A través de un bucle se realizó este proceso para los 17 registros.

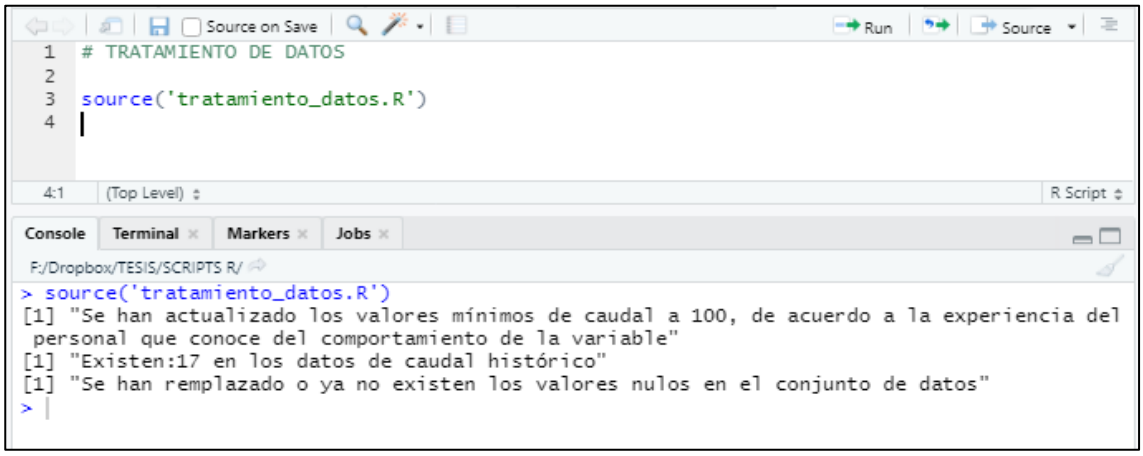
Es importante señalar que no existe una técnica única para realizar estos procesos de limpieza de datos, por lo cual la decisión de las acciones ejecutadas a través del script responde tanto al tipo de datos con el que se cuenta (serie temporal).

De igual manera se tomó el criterio del personal de la CHCCS que conoce por experiencia directa el comportamiento de la variable de la fuente de datos (caudal del Río Coca).

En la Figura 15 se muestra el resultado de la ejecución del script tratamiento_datos.R, que realiza la corrección de valores mínimos y máximos de caudal así como de tratamiento de valores nulos.

Figura 15

Resultado de la Limpieza de Datos



Nota: Resultado del proceso de ejecución del script: tratamiento_datos.R.

Elaboración Propia (RStudio), 2020

Estructuración de los Datos. Con el objetivo de que la estructura de los datos aporten a un mejor funcionamiento de los algoritmos usados en el Deep Learning, es necesario realizar procesos a los datos tales como: transformación, normalización **(Morante, 2018)**.En este sentido se detallan de manera general las acciones tomadas para cada uno de los tres procesos mencionados anteriormente.

- Transformación: Una transformación se puede resumir como una función matemática que se aplican a todas las observaciones de una determinada variable. Este proceso tiene como objetivo solucionar como asimetría, heterogeneidad, valores atípicos, etc. De forma específica se realizó una transformación logarítmica, la cual reemplaza los valores de cada observación con su respectivo logaritmo natural.
- Normalización: La normalización por otro lado toma cada observación de una variable y los ajusta dentro de un rango definido. Lo cual permite que se pueden

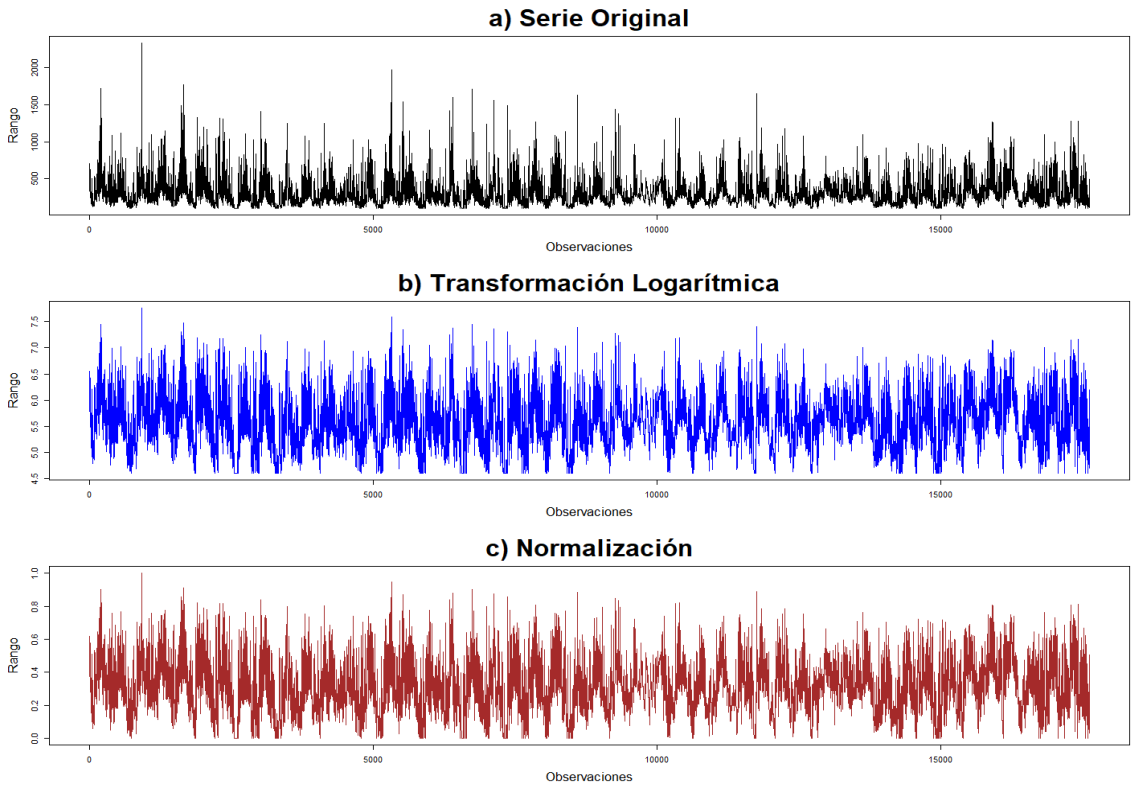
interrelacionar variables de diferentes órdenes de magnitud y mejorar de forma sustancial los resultados obtenidos a través de los modelos implementados.

Existen varios métodos para normalizar los datos, en nuestro caso se utilizó el método conocido como Escalado de variables (MinMax Scaler). Este método consiste en mantener los valores máximos u mínimos y comprime el resto de datos entre ellos.

En la Figura 16 se ilustra los gráficos de líneas de la serie original de los datos y a continuación los gráficos de los resultados de los procesos de transformación logarítmica y de normalización.

Figura 16

Resultados de los procesos de Transformación y Normalización



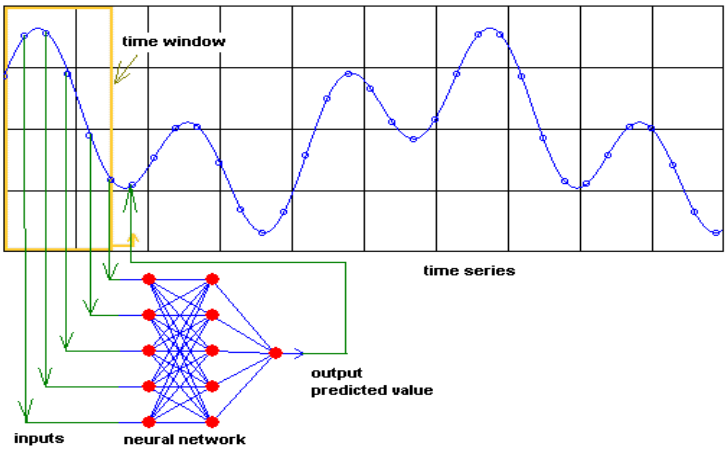
Nota: a) Serie Original, b) Transformación Logarítmica y c) Normalización, Elaboración Propia (RStudio), 2020

Integración De Los Datos Los modelos predictivos aplicados a las series temporales pueden ser univariantes o multivariantes. En el caso de tener un enfoque univariante el objetivo es realizar predicciones de valores futuros de una variable, para esto utilizo como insumo la información que puedo extraer de los valores pasado de la misma serie temporal. Por otro lado el enfoque multivariante se utiliza la información de otras variables adicionales, sin que esto excluya la información histórica de misma variable objeto estudio. De acuerdo a los datos disponibles nos encontramos en el caso de una serie temporal univariante, en este sentido de manera general se establece que los valores futuros tienen dependencia de los valores pasados.

Es decir nos basaremos en los valores de la serie en los tiempos *t-1*, *t-2*,…, *t-n* para poder predecir el valor de t para la serie del caudal promedio diario del Río Coca. Este concepto se conoce como ventada deslizante o de tiempo (Aragón, 2017). En la Figura 17 se ilustra como para una red neuronal se utilizan los valores previos una variable para su predicción de acuerdo al concepto de ventana de tiempo

Figura 17

Uso de "Ventana de Tiempo" en Redes Neuronales



Nota: Imagen obtenida de: (Aragón, 2017)

Con todo lo expuesto anteriormente, la tarea de integración de datos tuvo como objetivo el desarrollo de scripts en el lenguaje R que permitan crear y almacenar las ventanas de tiempo; las cuales que serán utilizadas como entradas para los modelos a implementarse. De manera específica se crearon 7 ventanas de tiempo, las cuales consisten en 3, 7, 15, 21, 30, 45 y 60 observaciones hacia atrás (rezagos/retardos).

Estas ventanas de tiempo aplicadas a los datos que fueron ya transformados y normalizados son registradas en tablas de la BDD creada en el Motor de Base de Datos Postgres. En la Figura 18 se ilustra la estructura de una de las tablas que almacena las ventanas de tiempo creadas para la serie temporal.

Figura 18

Tabla de la Serie Temporal de Datos con Ventana de Tiempo

dm_3_l_n

select * from "load".dm_3_l_n — Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text | Record

	id	anio	epoca	n_mes	t3	t2	t1	t
1	1	1.972	1	1	0,513420226	0,6055840448	0,6201133604	0,5709100461
2	2	1.972	1	1	0,6055840448	0,6201133604	0,5709100461	0,4211806596
3	3	1.972	1	1	0,6201133604	0,5709100461	0,4211806596	0,3260146541
4	4	1.972	1	1	0,5709100461	0,4211806596	0,3260146541	0,2811039379
5	5	1.972	1	1	0,4211806596	0,3260146541	0,2811039379	0,4570325476
6	6	1.972	1	1	0,3260146541	0,2811039379	0,4570325476	0,5348840603
7	7	1.972	1	1	0,2811039379	0,4570325476	0,5348840603	0,5060168381
8	8	1.972	1	1	0,4570325476	0,5348840603	0,5060168381	0,5367066488
9	9	1.972	1	1	0,5348840603	0,5060168381	0,5367066488	0,5788121962
10	10	1.972	1	1	0,5060168381	0,5367066488	0,5788121962	0,4255364033

Nota: Ventana de tiempo con 3 observaciones hacia atrás, Elaboración Propia (DBeaver), 2020

Tanto los procesos de transformación, normalización y creación de estructuras para ventana de tiempo fueron automatizadas a través del script: “crear_conjuntos_datos.R”.

Formateo De Los Datos. El proceso de formateo de datos responde a los requerimientos propios de la librería utilizada en las siguientes tareas para la creación de los modelos. La librería utilizada dentro del entorno RStudio fue Keras, la cual se define como una biblioteca, de código abierto, para la experimentación de Redes de Aprendizaje Profundo **(Keras, 2020)**.

Modelamiento Inicial

Implementación Del Modelo Inicial. Ya con los datos formateados y la Arquitectura Técnica implementada, se procedió a realizar el Modelamiento inicial para comprobar el correcto funcionamiento de las herramientas integradas.

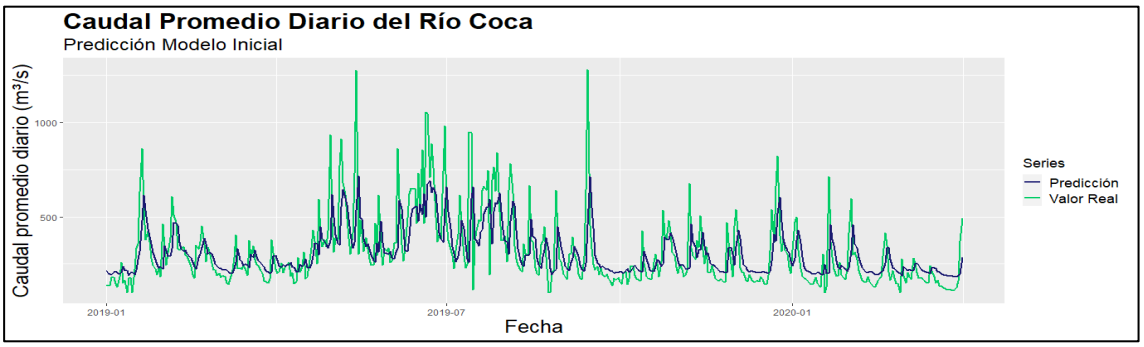
Para este primer ejercicio se eligió los siguientes parámetros para el modelo:

- Datos para el Modelo: Set completo de Datos con 3 observaciones hacia atrás.
- Particiones de Datos: 70% entrenamiento, 30% test.
- Estructuración de Datos: Datos Normalizados (rango de 0 a 1).
- Técnica de Modelado: Red Neuronal Multicapa
- Número de Capas / Neuronas: 3 (entrada, oculta, salida)
- Número de Neuronas: 3 capa entrada, 2 capa oculta, 1 capa salida; total 6.
- Función de Activación: Logística
- Función de Coste: Error Cuadrático Medio (Mean Squared Error - MSE)
- Optimizador: RMSprop (Root Mean Square prop)
- Épocas de Entrenamiento: 50

En la Figura 19 se ilustra el Gráfico de Línea de los resultados de predicción del Modelo Inicial, se incluye los valores reales y de predicción desde el 1 de enero del 2019 al 31/03/2020.

Figura 19

Modelo Inicial: Valores reales versus Predicciones



Nota: Período de Evaluación: enero 2019 a marzo 2020, Elaboración Propia (RStudio), 2020

Construcción del Modelo

Selección de la Técnica de Modelado

Debido a que para el desarrollo de nuestro modelo se aplicó una metodología híbrida que incluía un enfoque ágil, ya en la sección de modelamiento inicial se tuvo un primer acercamiento a la técnica de modelamiento utilizado. Nuestros datos tienen la estructura de una serie temporal con un enfoque univariante. La técnica de modelado base a utilizar corresponde a una Red Neuronal Artificial formada de múltiples capas, concepto conocido como perceptrón multicapa.

De manera más específica y acorde a la herramienta utilizada para la implementación, en la librería Keras las redes neuronales son definidas a través de una secuencia de capas. Esta secuencia de capas se crea a través de la clase *"Sequencial"*, a continuación tenemos un ejemplo de cómo se definiría un perceptrón multicapa en el lenguaje R, con dos entradas en la capa de entrada, cinco neuronas en la capa oculta y una neurona en la capa de salida (Allaire, 2017).

```
model <- keras_model_sequential()
model %>%
  layer_dense(units = 2,input_shape = 2) %>%
  layer_dense(units = 5) %>%
  layer_dense(units = 1)
```

Adicional a la definición de la estructura de la red neuronal la librería Keras también me permite parametrizar los aspectos detallados con respecto a las redes neuronales en el marco teórico, como son: función de activación, función de coste, optimizadores, épocas de entrenamiento, pesos, etc.

Preparación Detallada de los Datos

El proceso de preparación de los datos tiene como objetivo proveer a los modelos a construir de la información de entrada para el entrenamiento de dichos modelos. Encontrar un modelo final que brinde una respuesta satisfactoria a una problemática planteada requiere de varias iteraciones de prueba y error de varios modelos preliminares. Cada uno de estos modelos puede requerir un conjunto específico de datos, motivo por el cual se identificó la necesidad de un proceso automático que construya todos los conjuntos de datos requeridos.

Con base en la Arquitectura Técnica ya implementada, y tal como se indicó en la sección: Integración de los Datos, la creación de los conjuntos de datos ya se encuentra automatizada a través del script: “crear_conjuntos_datos.R”. El cual permite preparar distintos tipos de conjuntos de datos según las opciones enviadas como parámetros y que se aplican mediante combinaciones. A través de esta automatización se crearon un total de 60 conjuntos de datos, los cuales son registrados en forma de tablas en el gestor de Base de Datos. En la Tabla 6 se detalla a manera de ejemplo los conjuntos de datos (tablas) con 15 observaciones previas (rezagos), los cuales serán utilizados para los modelos que se construirán en las siguientes fases.

Tabla 6

Ejemplo de Conjuntos de Datos creados.

Conjunto de Datos - Tablas BDD	Parámetros / Opciones aplicadas
dm_15	Datos originales con 15 rezagos
dm_15_d	15 rezagos y diferenciación de datos
dm_15_d_n	15 rezagos, diferenciación y normalización de datos
dm_15_l	15 rezagos y transformación logarítmica de datos
dm_15_l_n	15 rezagos, transformación logarítmica y normalización de datos
dm_15_n	15 rezagos y normalización de datos

Nota: Elaboración Propia, 2020

Generación del Plan de Prueba

El plan de prueba escogido para validar los modelos construidos define dos aspectos que serán la base para escoger qué modelo presenta un mejor ajuste con respecto al resto. Estos aspectos corresponden a la división a realizarse en el conjunto de datos de entrada previa a la aplicación del modelo; y a las métricas para cuantificar el rendimiento del modelo posterior al proceso de predicción.

Para el primer aspecto cada conjunto de datos deberá ser divido en tres sub conjuntos de datos correspondientes a:

- Entrenamiento: Serán los datos con los cuales el modelo será entrenado.
- Validación: conjunto de datos embebidos dentro del conjunto de entrenamiento usado para prevenir que el modelo presente sobre o infra ajuste.
- Prueba: conjunto de datos sobre la que el modelo será evaluado en cuanto a su eficacia.

Los porcentajes asignados cada subconjunto varía en torno a valores usados de acuerdo a experiencias previas, de forma general el conjunto de entrenamiento suele comprender entre el 50% y el 80% de del conjunto, el conjunto de validación corresponde del 20% al 30% del conjunto de entrenamiento y finalmente el conjunto de prueba toma el porcentaje necesario para alcanzar el 100%. La Figura 20 ilustra de manera gráfica esta división del conjunto de datos, mostrando como el conjunto de validación forma parte del conjunto de entrenamiento.

Figura 20 División del Conjunto de Datos para construcción de Modelos

División del Conjunto de Datos para construcción de Modelos

Nota: Elaboración Propia, 2020

El segundo aspecto define las métricas que serán utilizadas para medir el rendimiento de los modelos, estas métricas se aplican sobre el conjunto de prueba; de manera general en modelos de predicción de valores continuos (regresión) las métricas tratan de representar la distancia existente entre los valores obtenidos de la predicción y el dato real.

Existe varias métricas utilizadas para el análisis de regresión, para el presente trabajo se escogió calcular y registrar para el ejercicio de evaluación posterior las siguientes:

- MAE: Error Absoluto Medio

- MAPE: Error Absoluto Medio Porcentual
- MSE: Error Cuadrático Medio
- R Cuadrado

Modelamiento Preparatorio

El objetivo de la tarea de Modelamiento preparatorio consiste en generar varios escenarios a través de modelos funcionales; los cuales serán evaluados de acuerdo a las métricas de desempeño definidas. Tanto el proceso de generación como de evaluación se lo ejecuta a través de continuas iteraciones de prueba, error y ajuste de los aspectos relacionados con la estructura de cada modelo (número de rezagos, capas ocultas, función de activación, etc.).

La construcción de todos los modelos de esta tarea fue optimizada a través de un script en lenguaje R: “modelamiento_preliminar.R”. Este script permitió generar 240 modelos correspondientes a distintas combinaciones de parámetros; en la Tabla 7 se detallan los parámetros con los respectivos valores utilizados para generar todos los modelos.

Tabla 7

Parámetros considerados para Modelamiento Preparatorio

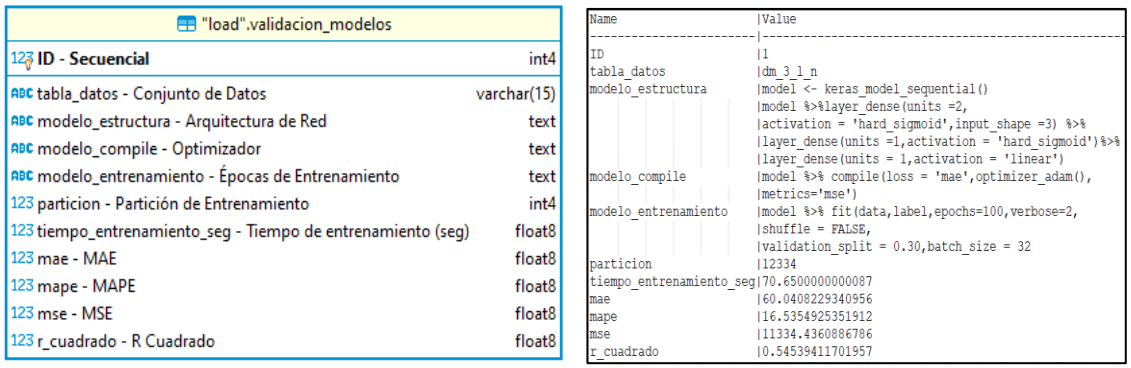
Parámetro	**Valores asignados**
Conjunto de Datos	Todos los generados en la fase de preparación (60)
Función de Activación	Linear, Logística, Logística Dura (variación Keras), Tangencial
Función de Coste	Pérdida: MAE, Métrica: MSE
Optimizador	RMSPROP y ADAM
Épocas de entrenamiento	100

Nota: Elaboración Propia, 2020

La automatización realiza el entrenamiento del modelo de acuerdo a la combinación de los parámetros y registra los resultados en el gestor de Base de Datos. Estos resultados permiten tener una trazabilidad de todo el proceso de modelamiento que sirvió para la tarea de evaluación ejecutada posteriormente. La Figura 21 ilustra la tabla de base de datos utilizada para registrar los resultados del modelamiento preliminar y un registro de la información recolectada.

Figura 21

Estructura para registro de Resultados del Modelamiento Preliminar



a) b)

Nota: a) Tabla de BDD para registrar la información, b) ejemplo de información recolectada para cada modelo generado, Elaboración Propia (DBeaver), 2020.

Es importante mencionar que la librería de Keras permite a través de sus funciones guardar el modelo generado como un objeto que incluye todas las configuraciones inherentes al modelo. La ventaja de este registro es que los pesos entrenados de cada modelo también se registran por lo cual se puede recuperar cualquier modelo generado y aplicarlo a futuro a nuevos datos.

En el siguiente capítulo se realizará la Evaluación de los modelos construidos con base a la información registrada, y de acuerdo a nuestra metodología planteada se procederá con el Modelamiento Final (Afinamiento) y Despliegue.

Capítulo IV: Evaluación e Implementación del Modelo

Una vez construidos los modelos que fueron resultado del Modelamiento Preparatorio, corresponde elegir, de acuerdo a los parámetros definidos en la sección: Generación del Plan de Prueba, el modelo que presentó el mejor rendimiento. Y a partir de la identificación de este modelo se realiza una iteración adicional para afinarlo (Modelamiento Final), y validar su efectividad final para responder a los objetivos del negocio planteados. Finalmente se ejecuta la fase de Implementación con su tarea de despliegue para poner a disposición del usuario final el Modelo Final construido.

Evaluación

El proceso de la evaluación se lo realizó bajo dos enfoques claramente identificados. Como primer enfoque se evaluaron los modelos preparatorios desde un punto vista intrínseco a su construcción; teniendo como resultado la identificación del modelo de mejor rendimiento. Con base en este modelo identificado se aborda el segundo enfoque orientado a evaluar si los resultados entregados por el modelo responden de manera eficiente como solución a la problemática. A continuación se detallan los resultados obtenidos de este proceso.

Evaluación del Modelo

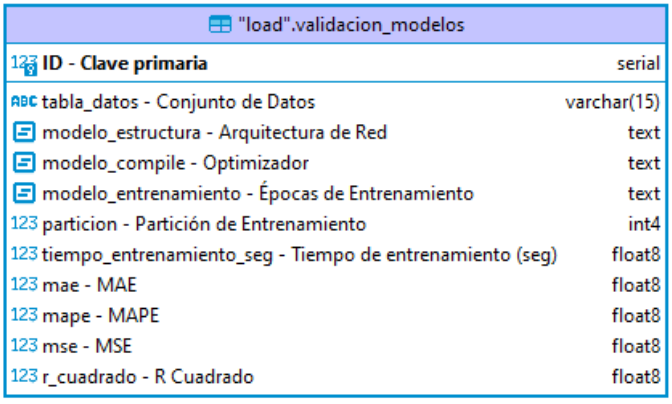
Las métricas calculadas para todos los modelos construidos en la tarea preparatoria incluyen las dos que son más utilizadas para su evaluación: MAE y RMSE. La decisión de una sobre la otra radica en un análisis y conocimiento de la calidad de los datos con respecto a la presencia y frecuencia de los datos atípicos, que dependiendo de su cantidad con respecto al total de datos pueden generar un sesgo. (Vandeput, 2019). Si se considera que existen gran cantidad de valores atípicos y que estos no inyectan sesgo en las predicciones que entregará el modelo se recomienda el

uso de MAE; por el contrario, si se considera que los valores altos o bajos no son atípicos y que por lo tanto forman parte de comportamiento posible de la variable el uso de RMSE se considera mejor (ZeroSpectrum, 2019).

Con este antecedente, y al considerar que los valores altos y bajos del conjunto de datos corresponden a valores reales de la variable estudiada (caudal promedio diario del Río Coca), el conjunto de datos fue utilizado para entrenar y probar todos los modelos generados. Durante este proceso se recolectaron, de cada entrenamiento realizado, los datos detallados en la Figura 22; que corresponden a las columnas creadas para la tabla: “load”.validacion_modelos almacenada en la Base de Datos creada en el servidor de PostgreSQL.

Figura 22

Datos almacenados y utilizados para la Evaluación de los Modelos



Nota: Elaboración Propia (DBeaver), 2020.

Con los datos recolectados se evaluó para cada modelo generado y entrenado el valor resultante para la métrica RMSE, como toda métrica relacionada al análisis de errores mientras más bajo sea el valor de RSME mejor será la capacidad de predicción.

En la Tabla 8 se detallan de forma ascendente los 5 valores de RSME más bajos identificados en la evaluación previa; como referencia también se incluyen los valores correspondientes a la métrica MAPE, la cual indica en términos porcentuales, y no en la magnitud de la variable como lo hace el RMSE, el promedio del error absoluto de los pronósticos.

Tabla 8

Rendimiento de los 5 mejores Modelos de acuerdo a la métrica RSME

Modelo	RMSE (m^3/s)	MAPE (%)	Función de Activación / Optimizador
dm_45_l_n	100,44	18,72 %	Linear / RMSPROP
dm_45_n	101,47	19,94 %	Sigmoid / ADAM
dm_60_n	101,71	19,62 %	Hard Sigmoid / ADAM
dm_60_n	101,79	20,04 %	Sigmoid / ADAM
dm_180_l_n	101,83	19,07 %	Linear / RMSPROP

Nota: Elaboración Propia, 2020

Evaluación de los Resultados

Para la evaluación con respecto a los resultados se usó el Coeficiente de Determinación R-cuadrado, que representa la bondad o aptitud del modelo con respecto a que tanto las variables independientes (rezagos), utilizadas como entradas del modelo, explican el comportamiento de la variable dependiente (pronóstico).

R-cuadrado se expresa en una escala intuitiva que desde valores negativos hasta 1, donde 0 o un valor negativo no mejora la predicción con respecto al modelo medio, de acuerdo a la media aritmética para todos los pronósticos; mientras que el valor de 1 indica una predicción perfecta.

Por tal motivo, es mejor el modelo cuando más se acercará este valor a 1, en la Tabla 9 se detalla los valores de R-cuadrado para los modelos evaluados en la sección

anterior; ordenados de forma descendente situando al inicio el modelo que explica mejor el comportamiento del caudal pronosticado.

Tabla 9 *Valor de R-cuadrado de los 5 Modelos con mejor rendimiento*

Valor de R-cuadrado de los 5 Modelos con mejor rendimiento

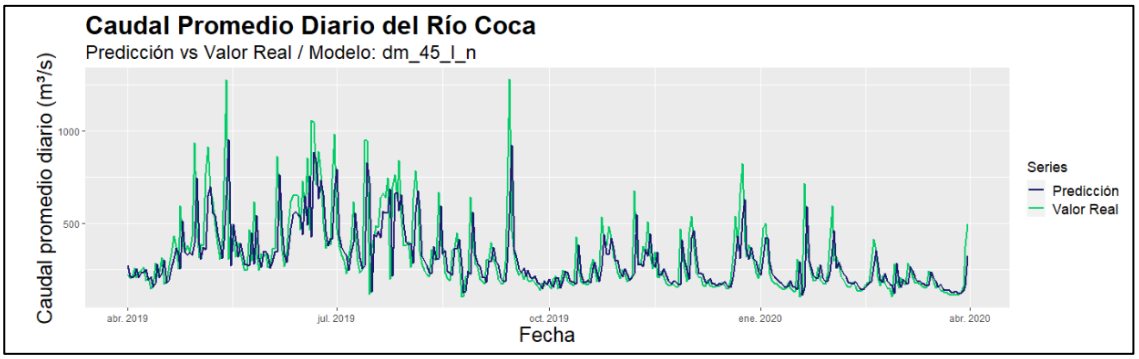
Modelo	R-cuadrado
dm_45_l_n	0,595
dm_45_n	0,587
dm_60_n	0,586
dm_60_n	0,585
dm_180_l_n	0,584

Nota: Elaboración Propia, 2020

En la Figura 23 se ilustra el gráfico de línea de las predicciones calculadas por el modelo con mejor rendimiento de acuerdo a la métrica RMSE y de R-cuadrado (dm_45_l_n).

Figura 23

Modelo mejor Evaluado: Valores reales versus Predicciones



Nota: Datos del período: abril 2019 - marzo 2020, Elaboración Propia (RStudio), 2020

Modelamiento Final

Una vez identificado el Modelo que presentó la mejor evaluación, tanto en su arquitectura como en los resultados entregados, se realiza un proceso de afinamiento final. Este afinamiento constituye el modelamiento final y tiene un enfoque específico sobre la arquitectura general del modelo, en este estudio en particular, del dm_45_l_n. El objetivo consiste en mejorar los indicadores utilizados previamente en la evaluación: RSME y R-cuadrado. El proceso de afinamiento incluyó una verificación adicional del conjunto de datos, donde se identificaron valores que, si bien son probables en el comportamiento de la variable estudiada, tienen una probabilidad muy baja de ser valores correspondientes a un promedio diario de la variable. En este caso, el valor máximo definido para un promedio diario del caudal del Río Coca se definió en los 1200 m^3/s. Se encontró un total de 42 registros que fueron tratados con este criterio.

A través del proceso iterativo final se logró una ligera mejora en las métricas del modelo, mediante el afinamiento de los parámetros de entrenamiento. La configuración del modelo afinado es la siguiente:

- Datos para el Modelo: Set completo de datos históricos con 45 observaciones hacia atrás.
- Particiones de Datos: 30% test, 70% entrenamiento, de este último 10% se usó para validación.
- Estructuración de Datos: Transformación logarítmica y normalización (rango de 0 a 1).
- Técnica de Modelado: Red neuronal multicapa
- Número de Capas: 4 (1 entrada, 2 ocultas, 1 salida)
- Número de Neuronas: 45 en la entrada, 18 en cada capa oculta, 1 en la salida.

- Función de Activación: Linear
- Función de Coste: Error Cuadrático Medio (Mean Squared Error - MSE)
- Optimizador: RMSprop (Root Mean Square prop)
- Épocas de Entrenamiento: 400

En la Tabla 10, se detallan los valores de las métricas: RMSE y R-cuadrado obtenido con los valores de parámetros detallados y como referencia los valores previos al afinamiento final. Adicionalmente en la Figura 24 se incluye el gráfico de línea de las predicciones entregadas por el modelo con respecto a los valores reales del caudal medio diario.

Tabla 10

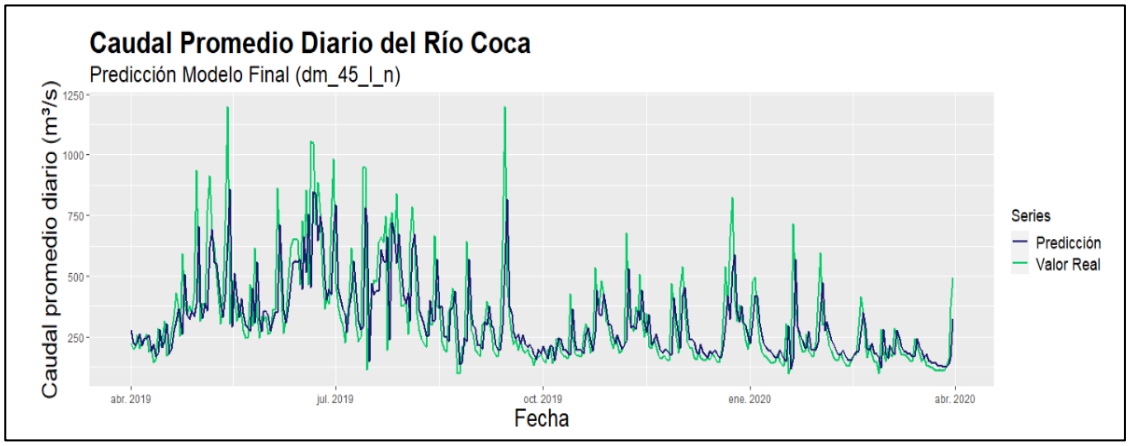
Valor de RMSE y R-cuadrado del Modelo Final elegido

Métrica	Modelamiento Preliminar	Modelamiento Final
RSME (m^3/s)	100,44	99,22
R- cuadrado	0,595	0,604

Nota: Elaboración Propia, 2020

Figura 24 Modelo Final: Valores reales versus Predicciones.

Modelo Final: Valores reales versus Predicciones.



Nota: Datos del período: abril 2019 - marzo 2020, Elaboración Propia (RStudio), 2020

Los resultados obtenidos para el Modelo Final arrojan un valor de R-cuadrado de 0,604, lo cual interpretado en porcentaje indica que el Modelo propuesto es capaz de explicar el comportamiento y las variaciones del Caudal medio diario del Río Coca en un 60,4%.

Este valor por sí solo no es capaz de justificar la aceptación o negación de nuestra hipótesis planteada. Motivo por el cual, es necesario realizar una evaluación comparativa con otro modelo que nos sirva de referencia. En este sentido el modelo de referencia será el que actualmente se utiliza en la CHCCS para realizar los pronósticos de la variable estudiada.

Evaluación Comparativa Entre el Modelo Actual y el Modelo Propuesto

De acuerdo al objetivo del negocio planteado en la sección: Determinación de los Objetivos del Negocio, lo que se pretende a través del presente trabajo es mejorar la precisión en la predicción del caudal medio diario del Río Coca. El caudal es una variable que tiene una relación directamente proporcional; por tal motivo al mejorar la predicción del caudal medio diario del Río Coca estaremos en la capacidad de mejorar la predicción de la energía que la CHCCS es capaz de producir.

La evaluación comparativa se efectúo entre los valores que el modelo utilizado en la CHCCS entregó para los períodos de enero a marzo del 2020, y los pronósticos que entrega el Modelo Final de la sección anterior. Mediante la Tabla 11 se detallan las principales métricas para evaluar el desempeño de un modelo de predicción, y a continuación se presenta, a través de la Figura 25, el gráfico de línea con los valores reales del caudal medio diario, junto a los pronósticos del modelo actual y el Modelo Final propuesto.

Tabla 11

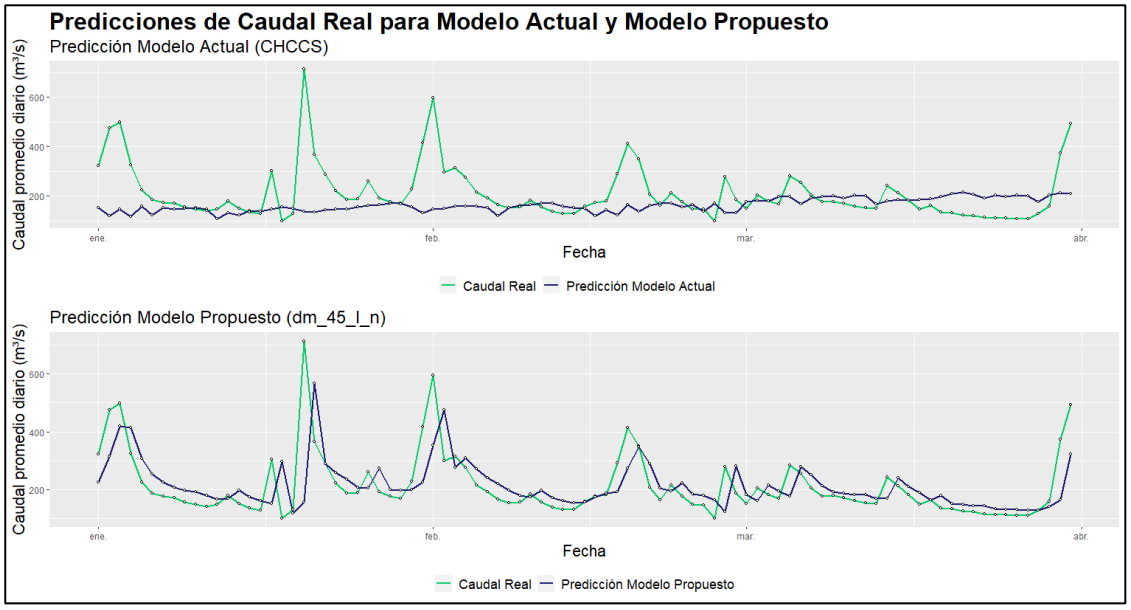
Valor de RMSE y R-cuadrado del Modelo Final elegido

Métrica	Modelo Actual CHCCS	Modelo Propuesto
MAE	81,26	60,03
MAPE	32,46	25,70
MSE	16749,38	9214,81
RSME	129,21	95,99

Nota: Elaboración Propia, 2020

Figura 25 Predicciones de Caudal Real en el Modelo Actual vs. Modelo Propuesto.

Predicciones de Caudal Real en el Modelo Actual versus Modelo Propuesto.



Nota: Datos del período: enero 2020 - marzo 2020, Elaboración Propia (RStudio), 2020

Según los resultados obtenidos del análisis comparativo, entre los pronósticos entregados por el Modelo de Predicción actual utilizado en la CHCCS y el Modelo Propuesto, se desprende que la precisión de pronóstico mejora en todas las métricas

utilizadas para evaluar el desempeño de los modelos. De igual manera, a través del gráfico de línea de las predicciones de ambos modelos se ilustra un mejor ajuste de los pronósticos del Modelo Propuesto con respecto a los valores reales del Caudal Medio Diario del Río Coca.

Implementación del Modelo

Con el Modelo Evaluado, tanto de forma intrínseca como en resultados, y realizado el Modelamiento Final (Afinamiento del Modelo mejor evaluado), permite dar paso a su implementación. Dentro de la Metodología CRISP-DM, que forma parte de la Metodología Híbrida, se considera que la implementación puede incluir la llamada puesta a Producción del Modelo construido; o sólo la elaboración del Plan para ejecutar este proceso a futuro.

Plan de Implementación

El Objetivo de este proceso tiene como fin que el modelo implementado pueda ser utilizado para la predicción de nuevos valores, ya sea para su uso en un ambiente de producción o para futuras pruebas y afinamientos.

Para lograr este cometido se desarrolló un script de R que lee los datos históricos necesarios para que el modelo realice la predicción; para esto los requisitos mínimos necesarios son:

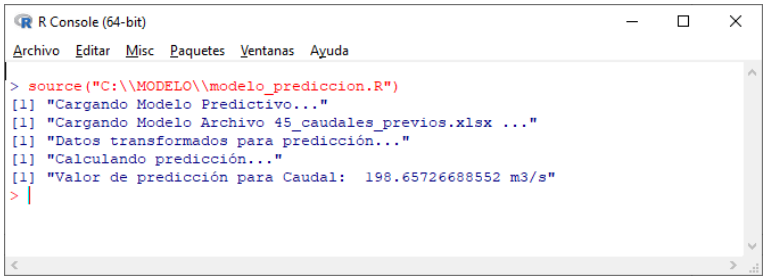
- Instalación de Lenguaje R (Consola)
- Archivo en formato Excel con las mediciones del caudal promedio, tomadas previamente para realizar la predicción
- Modelo Final

Con estos requisitos mínimos el Modelo Predictivo puede ser utilizado para obtener de manera sencilla predicciones de acuerdo a los datos proporcionados a

través del archivo Excel. En la Figura 26 se ilustra la interfaz en consola de R para realizar el cálculo de predicción del caudal con base a los requerimientos descritos anteriormente.

Figura 26 Interfaz de Consola del Modelo Predictivo

Interfaz de Consola del Modelo Predictivo



Nota: Elaboración Propia (R Console), 2020

Despliegue

Para poder desplegar nuestro modelo, y con miras a poder predecir valores con nuevos datos de entrada, se realiza un entrenamiento final del modelo, sin cambiar lo parámetros definidos, pero con el set completo de datos disponibles. Con esto se obtuvo un Modelo que puede ser puesto en Producción, una vez que se cumpla con el Plan de Implementación, para poder ser utilizado por el personal de la CHCCS para futuras predicciones.

Capítulo V: Conclusiones y Recomendaciones

Conclusiones

- De acuerdo a la revisión de literatura, se identificó que el uso de técnicas y modelos basados en Machine Learning, y de manera específica de las Redes Neuronales dentro del campo del Deep Learning, arrojan mejores resultados en comparación con técnicas tradicionales (Regresión Lineal, ARIMA) utilizadas para predicción de caudales y otros fenómenos estocásticos. Por ende se constituyen en los métodos a utilizar para definir la capacidad de energía a producir con base a los caudales predichos.
- Pese a que el caudal del Río Coca está directamente relacionados con múltiples variables como: precipitaciones, temperatura, estación del año, etc. Se identificó que de acuerdo a los datos disponibles, tanto por calidad como por cantidad, el caudal histórico del mismo río sería la variable para poder predecir su comportamiento a futuro.
- El tratamiento de los registros históricos del caudal como una serie temporal univariante, permitió la implementación de un Modelo de Aprendizaje Profundo, que logra identificar los patrones que sirven como base para realizar pronósticos del caudal disponible para la CHCCS. Al disponer del valor del caudal con la que contaría la CHCCS, se puede determinar la cantidad de energía que ésta puede producir.
- Con base en el análisis comparativo entre el Modelo Propuesto versus el Modelo Actual utilizado en al CHCCS, se acepta la hipótesis propuesta. La energía que es capaz de generar la CHCCS está directamente relacionada con el Caudal del río Coca; razón por la cual, al mejorar la predicción del caudal, que podrá ser aprovechado por la Central, se está mejorando de manera directa la predicción de la energía a producir.

- El uso de una metodología de desarrollo para la implementación del modelo, sirve de línea base para la ejecución lógica y ordenada de todo el proceso. En este caso particular la combinación de la Metodología CRISP-DM junto con SCRUM permitió obtener resultados funcionales (agregado de valor) desde las primeras iteraciones; es decir, contar con una arquitectura técnica y modelos preliminares funcionales; sobre los cuales, y a través de la ejecución del resto de “sprints”, se ejecutó el afinamiento y preparación para un despliegue del Modelo Final.
- Con el Modelo Final Propuesto se alcanzó un MAPE de 19,60%. Dicho de otra manera, el Modelo Propuesto tiene una precisión promedio del pronóstico del 80,4% (100% - MAPE) del valor real de caudal de todo el conjunto de datos (1972 - 2020). El Modelo desarrollado además presentó una mejora en la predicción, con los datos disponibles para para el período enero 2020 a marzo 2020, el análisis comparativo arrojó que el Modelo propuesto tiene una precisión del 74,3% del caudal real frente al 67,54% del Modelo actualmente utilizado en la CHCCS.
- La automatización de todo el proceso de generación de Modelos, que va desde la recolección y tratamiento de los datos, pasando por la construcción y entrenamiento de los modelos y finalizando con el cálculo de pronósticos del caudal, resultó de importante ayuda en la optimización del tiempo empleado. Sobre todo porque el proceso de encontrar una arquitectura de modelo, así como los parámetros de afinamiento constituye un ejercicio iterativo de prueba y error, es así que, la generación automática de escenarios (distintos modelos) preliminares permitió realizar un proceso gradual de discriminación de opciones hasta llegar al Modelo final planteado.
- La experiencia y conocimiento sobre el comportamiento de las variables involucradas en la implementación de Modelos de Aprendizaje automático son de vital

importancia y aporte para lograr que, tanto la cantidad como la calidad de información disponible, esté garantizada con el fin de poder alcanzar los objetivos planteados del negocio a través de las herramientas que nos ofrecen técnicas como el Machine Learning.

Recomendaciones

- Siempre que sea posible, la definición e implementación de una Arquitectura Técnica que apalanque todo el proceso de Modelamiento debe formar parte de las actividades incluidas en el alcance del trabajo propuesto. El registro y disponibilidad de los resultados obtenidos de manera preliminar al establecimiento de un Modelo a ser desplegado, servirán para orientar esfuerzos futuros, a optimizar soluciones que responden a los objetivos planteados, y a su vez descartar las que no han arrojado resultados satisfactorios.
- La inclusión a futuro de nuevas variables, además de los registros históricos del caudal del Río Coca, pueden ayudar a aumentar el porcentaje con el cual el Modelo es capaz de explicar el comportamiento del caudal medio diario del Río Coca. Valerse de información complementaria como: temperatura ambiental, precipitaciones, caudales de ríos que desembocan en el río Coca o sus afluentes, entre otras, pueden ser de gran aporte para robustecer el Modelo Propuesto o plantear un nuevo modelo de predicción.
- Con el auge que actualmente ha tomado temáticas como el Machine Learning, Data Science, Data Analytics, Big Data, etc. han promovido el desarrollo de herramientas cada más intuitivas para que los usuarios puedan experimentar con estos conceptos. Por lo cual, el uso de estas herramientas es totalmente recomendable para hacer más eficiente el proceso de implementación y uso de sus funcionalidades. Sin embargo, el uso de estas herramientas, debe venir siempre acompañado de la

comprensión de los conceptos teóricos utilizados, esto ayudará a que el afinamiento, que conlleva iteraciones de prueba y error, consiga alcanzar los objetivos planteados a través de esfuerzo y tiempos aceptables.

- Es importante establecer umbrales de precisión aceptable para las métricas planteadas para la evaluación de los resultados entregados por un Modelo como el propuesto a través del presente trabajo. La importancia de este punto radica en poder establecer límites para que el proceso de afinamiento y pruebas de distintas arquitecturas y parámetros de los modelos, a fin de que no resulte en un número iteraciones extremadamente altos y mejoren el desempeño objetivo del modelo, o por el contrario, un número corto de iteraciones no permita alcanzar el desempeño que el modelo puede entregar.
- Si bien es cierto el pronóstico del caudal medio diario del río Coca es de gran utilidad para el personal de la CHCCS, y de manera general para la predicción de la energía que se puede generar; el comportamiento del caudal en períodos más cortos o extensos que el estudiado son de vital importancia para otros frentes de la Central. Por tal motivo, se recomienda para trabajos futuros incursionar en la predicción del caudal por horas, semanas, meses, etc. Aspectos tan críticos como la erosión actual del río Coca justifican de sobremanera poder contar con la capacidad de adelantarnos a futuros fenómenos, aprovechando el potencial que entregan las herramientas de aprendizaje computacional como las Redes Neuronales.
- El proceso de recolección y comprensión de los datos que serán utilizados como entradas para los Modelos como el desarrollado, deben contar con el aporte del conocimiento y experiencia de las personas que trabajan de manera directa con estos registros. Esto evitará definiciones erróneas y reprocesos en el proceso de modelado.

Bibliografía

Allaire, J. (5 de Septiembre de 2017). *R Studio Blog*. Obtenido de https://blog.rstudio.com/2017/09/05/keras-for-r/

Aragón, F. (20 de Noviembre de 2017). *Series Temporales*. Obtenido de https://github.com/FrancisArgnR/SeriesTemporalesEnCastellano#predicci%C3%B3n-de-series-temporales-con-redes-neuronales

ARCONEL. (2000). Obtenido de Procedimientos de Despacho y Operación (Regulación No. CONELEC 006/00). Recuperado de: https://www.regulacionelectrica.gob.ec/wp-content/plugins/download-monitor/download.php?id=198

ARCONEL. (2019). *Reglamento de Despacho y Operación del Sistema Nacional Interconectado.* Obtenido de http://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2017/04/134_ReglamentoOperaci%C3%B3nSNI-Reforma-29.11.2016.doc

Backpropagation - El corazón de las redes neuronales. (s.f.). Obtenido de https://riptutorial.com/es/machine-learning/example/31623/backpropagation---el-corazon-de-las-redes-neuronales

Calvo, D. (12 de 07| de 2017). *Definición de red neuronal artificial*. Obtenido de https://www.diegocalvo.es/definicion-de-red-neuronal/

Castro, A. (11 de Agosto de 2017). *¿Qué es Machine Learning y para qué sirve?* Obtenido de https://www.inbest.cloud/comunidad/que-es-machine-learning-y-para-que-sirve

CONELEC. (2016). Obtenido de https://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2016/02/Regulacion-No.-CONELEC-007-00.pdf

Cross Industry Standard Process for Data Mining. (6 de Octubre de 2020). En Wikipeda. Obtenido de https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Data TechNotes. (14 de 02 de 2019). *Regression Model Accuracy (MAE, MSE, RMSE, R-squared) Check in R*. Obtenido de https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html

DECIDE SOLUCIONES S.L. (8 de 08 de 2019). *Calidad o cantidad de datos, qué es más importante para la IA*. Obtenido de https://decidesoluciones.es/calidad-o-cantidad-de-datos-para-ia/

Energía Estratégica. (23 de Mayo de 2019). Obtenido de https://www.energiaestrategica.com/en-2018-se-incorporaron-22-gw-hidroelectricos-en-el-mundo/

García., J. D. (s.f.). *Redes neuronales desde cero (I) – Introducción*. Obtenido de https://iartificial.net/redes-neuronales-desde-cero-i-introduccion/

Google Developers. (10 de 02 de 2020). *Introducción a las redes neuronales:*. Obtenido de https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/anatomy?hl=es-419

Hernandez, J., Asqui, G., Arellano, A., & Cunalata, C. (2017). Multistep-ahead Streamflow and Reservoir Level Prediction Using ANNs for Production Planning in Hydroelectric Stations. En IEEE (Ed.), *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).* Cancún. doi:10.1109/ICMLA.2017.0-115

Hewlett Packard. (s.f.). *¿QUÉ ES EL APRENDIZAJE AUTOMÁTICO?*

Hussin, S., Malek, M., Jaddi, N., & Hamid, Z. (2016). Hybrid metaheuristic of artificial neural network — Bat algorithm in forecasting electricity production and water consumption at Sultan Azlan shah Hydropower plant. En IEEE (Ed.), *2016 IEEE International Conference on Power and Energy (PECon).* doi:10.1109/PECON.2016.7951467

Kent, G. (s.f.). *¿Qué es machine learning?*

Keras. (28 de Abril de 2020). *Introduction to Keras for Researchers*. Obtenido de https://keras.io/getting_started/intro_to_keras_for_researchers/

Martinez, S. (s.f.). *SupeRRHHeroes*. Obtenido de https://superrhheroes.sesametime.com/la-metodologia-agil/

MEER. (2016). *Plan Maestro de Electricidad 2016-2025*. Obtenido de https://www.cnelep.gob.ec/plan-maestro-electricidad/

Mena, A. (4 de Agosto de 2017). *Corporación para la Investigación Energética*. Obtenido de http://energia.org.ec/cie/el-desarrollo-de-la-energia-renovable-en-el-ecuador/

Morante, S. (01 de Noviembre de 2018). *Precauciones a la hora de normalizar datos en Data Science*. Obtenido de https://empresas.blogthinkbig.com/precauciones-la-hora-de-normalizar/

Nacar, S., Hinis, M. A., & Kankal, M. (2018). Forecasting Daily Streamflow Discharges Using Various Neural Network Models and Training Algorithms. En Springer (Ed.), *KSCE J Civ Eng 22*, (págs. 3676–3685). doi:https://doi.org/10.1007/s12205-017-1933-7

Navarro, B. (30 de Septiembre de 2017). *Machine Learning y Deep Learning.* Obtenido de https://planetachatbot.com/claves-de-inteligencia-artificial-machine-learning-y-deep-learning-53a2032aaad

Niu , W.-J., Feng, Z.-K., Feng , B.-F., Min, Y.-W., Cheng, C.-T., & Zhou, J.-Z. (2019). Comparison of Multiple Linear Regression, Artificial Neural Network, Extreme Learning Machine, and Support Vector Machine in Deriving Operation Rule of Hydropower Reservoir. En Water (Ed.)., (pág. 88). doi:10.3390/w11010088

Proyectos Ágiles. (s.f.). *https://proyectosagiles.org/*. Obtenido de https://proyectosagiles.org/que-es-scrum/

proyectosagiles.com. (s.f.). Obtenido de https://proyectosagiles.org/que-es-scrum/

Red neuronal artificial. (24 de junio de 2020). Wikipedia. Obtenido de https://es.wikipedia.org/wiki/Red_neuronal_artificial

Rodríguez, T. (16 de Octubre de 2020). *Machine Learning y Deep Learning: cómo entender las claves del presente y futuro de la inteligencia artificial*. Obtenido de https://www.xataka.com/robotica-e-ia/machine-learning-y-deep-learning-como-entender-las-claves-del-presente-y-futuro-de-la-inteligencia-artificial

Rovira, I. (s.f.). *Estudio de caso: características, objetivos y metodología*. Obtenido de https://psicologiaymente.com/psicologia/estudio-de-caso

Sauhats, A., Petrichenko, R., Baltputnis, K., Broka, Z., & Varfolomejeva, R. (2016). A multi-objective stochastic approach to hydroelectric power generation scheduling. *2016 Power Systems Computation Conference (PSCC)*, (págs. 1-7). Genoa. doi:10.1109/PSCC.2016.7540821

Sauhats, A., Petrichenko, R., Baltputnis, K., Broka, Z., & Varfolomejeva, R. (2016). A multi-objective stochastic approach to hydroelectric power generation scheduling. En IEEE (Ed.), *2016 Power Systems Computation Conference (PSCC).* Genoa. doi:10.1109/PSCC.2016.7540821

Sauhats, A., Petrichenko, R., Broka, Z., Baltputnis, K., & Sobolevskis, D. (2016). ANN-based forecasting of hydropower reservoir inflow. *2016 57th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON).* Riga. doi:10.1109/RTUCON.2016.7763129

Shen, J., & Cheng, C. (2015). A Generalized Decision Support System for Short-Term Scheduling of China's Big Hydropower Systems. *World Environmental and Water Resources Congress 2015.* doi:10.1061/9780784479162.183

Torra, V. (Diciembre de 2011). *La inteligencia artificial*. Obtenido de http://www.fgcsic.es/lychnos/es_es/articulos/inteligencia_artificial

Vandeput, N. (5 de Julio de 2019). *Forecast KPI: RMSE, MAE, MAPE & Bias*. Obtenido de https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Velasco, L. (26 de 04 de 2020). *Optimizadores en redes neuronales profundas: un enfoque práctico*. Obtenido de https://medium.com/@velascoluis/optimizadores-en-redes-neuronales-profundas-un-enfoque-pr%C3%A1ctico-819b39a3eb5

Werick, S. (1 de Junio de 2017). *Developing Predictive Analytics Solutions Using Agile Techniques*. Obtenido de https://seanwerick.com/2017/06/01/developing-predictive-analytics-solutions-using-agile-techniques/

Wikipedia. (28 de 01 de 2020). *Cross Industry Standard Process for Data Mining*. Obtenido de https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

ZeroSpectrum. (2 de Junio de 2019). *MAE vs MSE vs RMSE*. Obtenido de http://zerospectrum.com/2019/06/02/mae-vs-mse-vs-rmse/#comments