UNIVERSIDAD TÉCNICA DEL NORTE

FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS

CARRERA DE INGENIERÍA INDUSTRIAL

TESIS DE GRADO PREVIA A LA OBTENCIÓN DEL TÍTULO DE INGENIERA INDUSTRIAL

"MODELO PARA EL PRONÓSTICO DE LA DEMANDA DE AGUA POTABLE APLICANDO MODELOS DE INTELIGENCIA ARTIFICIAL EN LA PORTADA DEL CANTÓN MIRA"

AUTOR: DÍAZ ENRÍQUEZ EMILIA CAROLINA

DIRECTOR: PhD. ROBERT VALENCIA CHAPI

IBARRA-ECUADOR

2023

UNIVERSIDAD TÉCNICA DEL NORTE

BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE

LA UNIVERSIDAD TÉCNICA DEL NORTE

IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	1003953286		
APELLIDOS Y NOMBRES:	Díaz Enríquez Emilia Carolina		
DIRECCIÓN:	Ibarra– Imbabura -Ecuador		
EMAIL:	ecdiaze@utn.edu.ec		
TELÉFONO FIJO:		**TELÉFONO MÓVIL:**	0987930660

DATOS DE LA OBRA		
TÍTULO:	Modelo para el pronóstico de la demanda de agua potable aplicando modelos de inteligencia artificial en La Portada del cantón Mira.	
AUTOR (ES):	Díaz Enríquez Emilia Carolina	
FECHA: DD/MM/AAAA	16/02/2023	
SOLO PARA TRABAJOS DE GRADO		
PROGRAMA:	■ PREGRADO	☐ POSGRADO
TITULO POR EL QUE OPTA:	Ingeniería Industrial	
ASESOR /DIRECTOR:	PhD. Robert Valencia Chapi	

CONSTANCIAS

La autora manifiesta que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de la misma y saldrá en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 16 días del mes de febrero del 2023

LA AUTORA:

.....

Díaz Enríquez Emilia Carolina

C. C. 1003953286

UNIVERSIDAD TÉCNICA DEL NORTE
FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS
CARRERA DE INGENIERÍA INDUSTRIAL
CERTIFICACIÓN DEL TUTOR

Yo PhD. Robert Valencia Chapi. Director del trabajo de grado desarrollado por la señorita estudiante: **EMILIA CAROLINA DÍAZ ENRÍQUEZ** para la obtención del título de Ingeniera Industrial.

CERTIFICA

Que, el Proyecto de Trabajo de Grado titulado**: "MODELO PARA EL PRONÓSTICO DE LA DEMANDA DE AGUA POTABLE APLICANDO MODELOS DE INTELIGENCIA ARTIFICIAL EN LA PORTADA DEL CANTÓN MIRA "** ha sido elaborado en su totalidad por la señorita estudiante Emilia Carolina Díaz Enríquez, bajo mi dirección, para la obtención del título de Ingeniera Industrial. Luego de ser revisado, considerando que se encuentra concluido y cumple con las exigencias y requisitos académicos de la Facultad de Ingeniería en Ciencias Aplicadas, Carrera de Ingeniería Industrial, autoriza la presentación y defensa para que pueda ser juzgado por el tribunal correspondiente

Ibarra, 16 días del mes de febrero del 2023

.....

PhD. Robert Valencia Chapi
DIRECTOR DE TRABAJO DE GRADO

DEDICATORIA

***A mis hijxs**, por ser mi motor, la razón por la que cada día me esfuerzo, ellxs son el pilar de mi vida el motivo por el que sigo luchando y por el que seguiré luchando cada día de mi vida. Sin ellxs nada tendría sentido, son mi prioridad y todos mis esfuerzos son para que tengan un mejor futuro y su calidad de vida cada día se incremente.*

***A mí**, porque a pesar de todo el dolor y sufrimiento que he vivido, he estado ahí para darme apoyo y demostrarme que soy extremadamente fuerte y el amor por mis hijxs es más grande que cualquier adversidad y ese es el motor para continuar esforzándome y ser mejor.*

.

EMILIA CAROLINA

AGRADECIMIENTO

A Universidad Técnica del Norte*, la Facultad de Ingenierías en Ciencias Aplicadas y la Carrera de Ingeniería Industrial, por darme la oportunidad de formarme académicamente, aprender y crecer profesionalmente.*

A Ing. Yakcleem Montero Santos MSc. *quien en calidad de tu director y docente ha sido mi mentor profesional y académico, guiándome hacia la excelencia académica.*

A Ing. Erik Orozco MSc. *por su constante apoyo en mi formación académica y personal, aconsejándome para mejorar cada día.*

A los señores *miembros de la Junta de Agua Potable de La Portada, quienes confiaron en mí permitiéndome realizar el trabajo de titulación.*

Y a todas aquellas personas que han sido parte directa o indirecta de todo este proceso de formación profesional y personal.

EMILIA CAROLINA

ÍNDICE

Resumen 7

Abstract 8

1. Capítulo I Generalidades 9

1.1. Problema 9

1.2. Objetivo General 10

1.3. Objetivos Específicos: 10

1.4. Justificación 10

1.5. Alcance 12

2. Capítulo II Marco Teórico 14

2.1. Introducción 14

2.2. Sistemas de Agua Potable 14

2.2.1. Captación 15

2.2.2. Producción 15

2.2.3. Reserva 15

2.2.4. Distribución 15

2.3. Pronóstico de demanda 16

2.3.1. Errores del Pronóstico 16

2.3.2. Redes Neuronales Artificiales 18

2.3.3. Máquinas de Soporte Vectorial 25

2.3.4. Bosques Aleatorios 31

2.3.5. Algoritmo K Vecinos más Cercanos 35

3. Capítulo III Diagnóstico de la situación actual 38

3.1. Descripción de la organización 38

3.2. Misión 38

3.3. Visión 38

3.4. Objetivos 39

3.5. Organigrama 39

3.6. Descripción del proceso de productivo 39

3.6.1. Captación y Conducción 39

3.6.2. Producción 40

3.6.3. Distribución 40

3.7. Producción y consumo anual 40

3.8. Volumen de almacenamiento 41

3.9. Datos de la población 41

3.10. Análisis de autocorrelación 42

3.10.1. Análisis de autocorrelación de la demanda mensual de agua potable 43

3.10.2. Análisis de autocorrelación de los clientes 43

3.11. Análisis de estacionalidad de la serie de tiempo 44

3.11.1. Análisis de estacionalidad de la demanda mensual de agua potable 44

3.11.2. Análisis de estacionalidad de los clientes 45

4. Capítulo IV Modelo de pronóstico 46
4.1. Selección de la variable 46
4.2. Obtención de datos 46
4.3. Estructuración de los modelos 47
4.3.1. Análisis la serie temporal 47
4.3.2. Análisis de tendencia y estacionalidad 48
4.4. Modelo de Redes Neuronales Artificiales 49
4.4.1. Pronóstico de la red neuronal óptima 49
4.4.2. Estructura de la RNA 50
4.4.3. Pronóstico mejorando el entrenamiento de la red 51
4.5. Modelo K Nearest Neighbor Regression 52
4.6. Modelo Support Vector Machine 54
4.7. Modelo Random Forest 55
4.8. Resultados de los modelos de pronóstico 56
4.8.1. Comparación de modelos 57
4.8.2. Resultados del pronóstico para los años 2022, 2023 y 2024 58
CONCLUSIONES 61
RECOMENDACIONES 63
REFERENCIAS 64
ANEXOS 71

Anexo 1. Codificación de la serie temporal de la demanda de agua potable..... 72

Anexo 2. Codificación del Modelo de Redes Neuronales Artificial 72

Anexo 3. Codificación del Modelo K Nearest Neighbor Regression..... 73

Anexo 4. Codificación del Modelo Support Vector Machine 74

Anexo 5. Codificación del Modelo Random Forest 75

ÍNDICE DE TABLAS

Tabla 1 Volumen $\boldsymbol{m3}$ de consumo mensual total desde el año 2016 hasta el 2021. 41

Tabla 2 Proyección poblacional de La Portada 2016-2022. 42

Tabla 3 Clientes mensuales desde el año 2016 hasta el 2021. 47

Tabla 4 Pronóstico de la demanda $\boldsymbol{m3}$ para el año 2022 de los diferentes modelos. 57

Tabla 5 Pronóstico de la demanda $\boldsymbol{m3}$ para el año 2022, 2023 y 2024..... 59

ÍNDICE DE FIGURAS

Figura 1 Sistema de agua potable 14

Figura 2 Célula nerviosa 21

Figura 3 Red neuronal totalmente conectada 23

Figura 4 Aprendizaje supervisado de un proceso 25

Figura 5 Hiperplano de separación óptimo desarrollado por las SVMs 27

Figura 6 Mapeo de datos con comportamiento no lineal 29

Figura 7 Arquitectura de Random Forest 31

Figura 8 Estructura Orgánica de la JAAP La Portada 39

Figura 9 Autocorrelación de la demanda de agua potable 43

Figura 10 Autocorrelación de los clientes 44

Figura 11 Análisis de estacionalidad de la demanda de agua potable 45

Figura 12 Análisis de estacionalidad de los clientes 45

Figura 13 Análisis de la serie temporal de la demanda de agua potable 48

Figura 14 Análisis de tendencia de la serie temporal de la demanda de agua potable 48

Figura 15 Análisis estacionalidad de la serie temporal de la demanda de agua potable 49

Figura 16 Codificación de la red neuronal óptima 50

Figura 17 Pronóstico de la red MLP 50

Figura 18 Estructura de las capas de la red 51

Figura 19 Codificación de la mejora de entrenamiento de la red neuronal 52

Figura 20 Pronóstico del modelo de Redes Neuronales Artificiales 52

Figura 21 Codificación del aprendizaje automático 53

Figura 22 Predicción del aprendizaje automático 53

Figura 23 Codificación de la mejora de entrenamiento del modelo 54

Figura 24 Pronóstico de la demanda utilizando el modelo K Nearest Neighbor Regression 54

Figura 25 Codificación del entrenamiento y mejora del modelo SVM 55

Figura 26 Pronóstico de la demanda utilizando el modelo SVM 55

Figura 27 Codificación del entrenamiento inicial del modelo Random Forest 56

Figura 28 Pronóstico de la demanda utilizando el modelo Random Forest 56

Figura 29 Comparación de los valores obtenidos como error de pronóstico de los modelos aplicados 58

Figura 30 Pronóstico de la demanda $\boldsymbol{m3}$ para el año 2022, 2023 y 2024. 59

Resumen

Este trabajo muestra la aplicación de los modelos de pronóstico de inteligencia artificial Redes Neuronales Artificiales, Máquinas de Soporte Vectorial, Random Forest y el Algoritmo KNN para el pronóstico de demanda de agua potable de La Portada, utilizando las herramientas de pronóstico del software RStudio.

Tomando en cuenta que, en todo proceso de planificación, la previsión cumple un rol esencial. Al ser uno de los datos necesarios para gestionar los recursos de la organización, y por lo tanto de las actividades y/o procesos de esta. Se debe considerar para la ejecución del modelo, el horizonte temporal de pronóstico evaluando las variables de entrada para el entrenamiento de los modelos.

Del desarrollo de los modelos de pronóstico se obtienen diferentes resultados de error de pronóstico. Estos serán los valores para comparar los modelos de inteligencia artificial específicamente MAE, RMSE y MAPE. Como resultado de esta comparación se identifica el modelo de pronóstico de redes neuronales artificiales como la mejor opción para el comportamiento de esta serie temporal, obteniendo como MAE 0.4033, 0.5464 de RMSE y MAPE de 0.0506.

Abstract

This paper presents an application of artificial intelligence forecasts models such as Artificial Neural Networks, Support Vector Machine, Random Forest, and K Nearest Neighbour Regression in the drinking water forecast demand of La Portada, using RStudio forecasting tools.

One must take into account that in any planning process, forecasts have a fundamental function. This is one of the preliminary data to manage the resources, therefore the activities or processes of the company. For the development of the model, the forecast time horizon must be considered, evaluating the variables that will be the inputs of the model. Model development gets different forecast error results. These will be the values to compare the intelligence artificial models specifically MAE, RMSE, and MAPE. As result is finding the forecasting artificial neural networks model, like the best choice for this temporal series, getting 0.4033 to MAE, 0.5464 in RMSE, and MAPE of 0.0506.

Capítulo I

Generalidades

1.1. Problema

En La Portada del Cantón Mira el servicio de agua potable y alcantarillado es brindado y administrado por La Junta Administradora de agua potable La Portada, constituida por una directiva de la comunidad y el operador. Al ser una zona en crecimiento, se considera que el problema en general consiste en la necesidad de anticiparse y proyectarse ante una demanda futura de agua potable que deriva del rápido crecimiento poblacional de la zona y al incremento del consumo que se registra a consecuencia de la pandemia, puesto que se utiliza más agua en la desinfección de los hogares. Lo que ocasiona la necesidad de previsión para abastecer adecuadamente del recurso hídrico a los usuarios, motivo del presente trabajo de investigación.

Parte del problema es también el abastecimiento, debido al crecimiento demográfico, el aumento del número de habitantes provoca una mayor demanda, cuando se habla de abastecimiento adecuado de agua se refiere a la cantidad y calidad de líquido disponible. Por ello, es importante un modelo para pronosticar la demanda de agua potable ya que se busca hacer coincidir la cantidad de suministro que la empresa ofrece con la cantidad requerida por los usuarios.

Teniendo en cuenta el crecimiento poblacional y las nuevas conductas de consumo producto de la pandemia, se ve la necesidad de incluir el pronóstico en la planificación de distribución del servicio para el desarrollo controlado y eficiente.

Sin la correcta planificación de la demanda de agua potable, la administración y distribución de los recursos sería inadecuada, no se priorizaría las zonas con mayor necesidad

producto de ello la insatisfacción de los clientes incrementaría. Hay que considerar el control continuo del servicio para evitar y atender a tiempo los fallos imprevistos en el sistema.

1.2. Objetivo General

Diseñar el modelo para el pronóstico de la demanda del Sistema de Agua Potable de La Portada aplicando modelos de inteligencia artificial para el aseguramiento de la planificación y la optimización en el uso de los recursos.

1.3. Objetivos Específicos:

- Establecer las bases teóricas y científicas de la investigación mediante la revisión del estado del arte para darle soporte a la investigación.
- Diagnosticar la situación actual de la organización y del consumo de agua potable aplicando el análisis de autocorrelación y estacionalidad para la validación de las series de tiempo que corresponden a la base de datos para la investigación.
- Pronosticar la demanda de agua potable en SAP la Portada, utilizando modelos de inteligencia artificial mediante la programación en el software R Studio, que permita determinar la previsión de la demanda en la empresa objeto de estudio.

1.4. Justificación

La Constitución de la República del Ecuador (2008), en el artículo 12 prescribe que: “El derecho humano al agua es fundamental e irrenunciable. El agua constituye patrimonio nacional estratégico de uso público, inalienable, imprescriptible, inembargable y esencial para la vida”.

La captación y distribución del agua han sido uno de los principales pilares para la supervivencia y desarrollo de los pueblos; en la actualidad las empresas encargadas del abastecimiento de este recurso hídrico de consumo humano, buscan estrategias para garantizar el provisionamiento de agua potable a toda la población (*Ley Orgánica de Recursos Hídricos, Usos*

y Aprovechamiento Del Agua, 2014), siendo prioridad para ellas la adaptabilidad y flexibilidad, en especial a los sectores en donde su población tiene un crecimiento considerable que influirá en el volumen de agua a potabilizar y que debe constar en la planificación para administrar y dotar efectivamente de recurso hídrico a sus habitantes.

La propuesta está alineada al Plan Nacional de Desarrollo "Toda una vida" (2017) a través de su Objetivo 5: "Impulsar la productividad y competitividad para el crecimiento económico sostenible de manera redistributiva y solidaria", y su política 5.6: Promover la investigación, la formación, la capacitación, el desarrollo y la transferencia tecnológica, la innovación y el emprendimiento, la protección de la propiedad intelectual, para impulsar el cambio de la matriz productiva mediante la vinculación entre el sector público, productivo y las universidades. (Consejo Nacional de Planificación, 2017)

Uno de los principales inconvenientes es la incertidumbre generada por la falta de una correcta planificación, el incremento en el consumo del agua potable debido a las necesidades de sanitización que trajo la pandemia del Covid 19 y el crecimiento demográfico del sector de La Portada. Generado así la necesidad de contar con una herramienta que permita identificar, pronosticar y evaluar la demanda del líquido vital, para la toma de decisiones en la prestación del servicio, por lo que se presenta el modelo de pronóstico adecuado generando predicciones precisas que permitan conocer la demanda de agua potable a mediano plazo.

La directiva del SAP La Portada para atender la creciente demanda de agua potable en el sector, propone incrementar su capacidad de captación, potabilización y reserva de agua; para ello juntamente con el GAD Cantonal MIRA, han ideado un proyecto de construcción de una nueva fuente de reserva con mayor capacidad, es así como este modelo de pronóstico de inteligencia artificial para la demanda permitirá predecir el comportamiento de la demanda y sus patrones de

variación en el tiempo. Además, traerá efectos importantes en la planificación; determinación de la expansión necesaria del sistema de agua potable; los costos de producción, en la reducción de falencias durante el servicio y en la satisfacción de la población expansión del sistema de agua potable, de acuerdo con la demanda que tendrá el sector.

Al no contar con un modelo de pronóstico para la demanda de agua potable del SAP La Portada, la planificación de la producción sería deficiente e inadecuada, debido a que se realizaría en base a juicios y razonamientos de acuerdo con el conocimiento empírico y la experiencia de quienes conforman la Junta Administradora del agua potable La Portada, provocando un bajo desempeño del sistema y el reducido aprovechamiento de los recursos, generando a su vez insatisfacción en los usuarios.

La investigación tendrá como resultado final el desarrollo de un modelo para el pronóstico de la demanda del servicio de agua potable y alcantarillado de La Portada del Cantón Mira, el mismo permitirá efectivizar la toma de decisiones en la red de distribución de la zona, optimizando los recursos de tal forma que logre suplir la demanda con un mínimo costo de operación. Es de total consideración que, en dependencia al campo de aplicación del modelo, se requiere ajustar ciertos parámetros en la estructura que se definirá en la programación de la red neuronal. Además, en las restricciones para el conjunto de variables conocidas en la resolución del sistema.

1.5. Alcance

La relevancia de la investigación se enfoca en la necesidad de realizar el pronóstico de la demanda de agua potable en La Portada, utilizando varios modelos de inteligencia artificial para obtener los pronósticos esperados en dependencia del crecimiento de la población y la demanda de obtener este recurso.

Para garantizar el correcto modelamiento del pronóstico, se observará el crecimiento poblacional de la zona, se recopilará los datos históricos de la demanda de consumo de al menos 36 meses, se establecerá el tiempo de la proyección esperada, la tendencia del consumo del recurso hídrico de la población, las estacionalidades que presentan en ciertas épocas del año y los errores de la predicción de la demanda.

El modelo de pronóstico como herramienta contribuirá al mejoramiento de la planificación de agua potable y apoyará a la correcta toma de decisiones de acuerdo con el crecimiento poblacional para el 2022 y los años venideros.

Capítulo II

Marco Teórico

2.1. Introducción

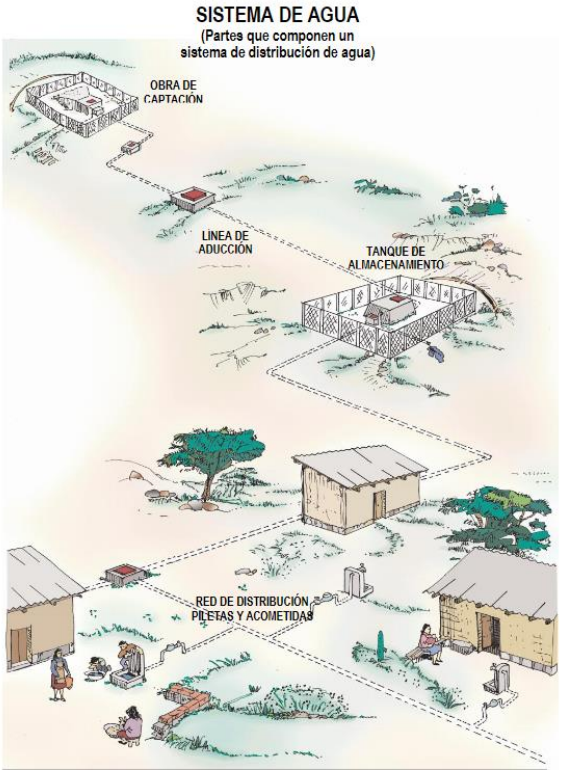
Se presenta la información relacionada con la investigación, así como los modelos de inteligencia artificial que se utilizan para el desarrollo del pronóstico.

2.2. Sistemas de Agua Potable

Un sistema de agua potable es el conjunto de todas las obras de ingeniería necesarias para la captación, conducción, tratamiento, almacenamiento y distribución del agua desde las fuentes naturales hasta las viviendas de los habitantes que consumirán este recurso. Para ilustrar la red del sistema a continuación se muestra la Figura 1.(Layme, 2020)

Figura 1

Sistema de agua potable



Nota. Adaptado de *Configuración típica de un sistema de abastecimiento de agua*, por Comisión Nacional del Agua, 2003

2.2.1. Captación

La etapa de captación consiste en la extracción de agua bruta desde una fuente natural, que puede provenir de distintos orígenes como: agua de lluvia almacenada, agua subterránea proveniente de manantiales; pozos; agua superficial de fuentes como: ríos, arroyos o lagos; e incluso agua proveniente del mar (que requiere de procesos adicionales de desalinización). (Lema, 2006)

2.2.2. Producción

La etapa de producción del agua potable está compuesta por todos los procedimientos necesarios para la potabilización. Estos pueden ser todos los tratamientos de desinfección y filtración utilizados para conseguir que el agua bruta sea apta para el consumo humano. Estos a su vez pueden tener subprocesos para el tratamiento de agua incluyendo etapas de retención de material grueso, de material fino en suspensión y de decantación de materiales muy finos; tratamientos químicos de desinfección en general. (Jiménez Terán, 2013)

2.2.3. Reserva

En la etapa de reserva en el sistema de agua potable se utiliza una bomba de extracción que suministra el agua tratada o potabilizada a un tanque de almacenamiento y regulación, éste tiene la función de mantener constante el flujo y caudal de salida a las tuberías de la red de distribución, independientemente de las variaciones de consumo. Es por esto por lo que la frecuencia de producción, que corresponde a la frecuencia de la bomba de extracción, es diferente de la frecuencia de consumo de agua. (Zhou et al., 2002)

2.2.4. Distribución

La red de distribución es el sistema de tuberías que va desde el tanque de regulación y llega a cada uno de los puntos de abastecimiento para los usuarios. Esta red está constituida por

estaciones de bombeo, tuberías, sistema de transporte de agua (aducción) y finalmente de dispositivos de medición de volumen de agua en los puntos de abastecimiento. (Comisión Nacional del Agua, 2003)

2.3. Pronóstico de demanda

El entorno en el que operan las organizaciones cambia constantemente, por ello existe la necesidad de los pronósticos. Las organizaciones que no reaccionen inmediatamente frente a las cambiantes condiciones del entorno y que no puedan prever a futuro con un grado de precisión aceptable, están condenadas a extinguirse (Hanke & Wichern, 2010).

Entre los modelos de pronóstico más utilizados están los modelos clásicos de media móvil simple, media móvil ponderada, suavizamiento exponencial y suavizamiento exponencial ajustado. No obstante, en la actualidad se utilizan modelos más sofisticados en la manipulación de datos como los de inteligencia artificial.

2.3.1. Errores del Pronóstico

Según Zavala (2015), para valorar el desempeño del pronóstico ($\hat{X}$) se utilizan medidas que comparan los resultados de la predicción con el valor real. Se puede utilizar estas medidas también para la evaluación de los datos.

Para una serie (X_n) = 1 de media $\overline{X}$, Hyndman (2014) describe las siguientes medidas de error cuyas ecuaciones se describen en cada caso:

2.3.1.1. Error Absoluto Medio (MAE):

El MAE es el promedio de la diferencia absoluta entre los datos de la base histórica y los valores obtenidos en el pronóstico. Valores más pequeños de MAE indican mejor ajuste del modelo al comportamiento de los datos.

$$MAE = \frac{1}{n}\sum_{t=1}^{T}\left|X_t - \hat{X}\right| \quad (1)$$

2.3.1.2. Error porcentual absoluto medio (MAPE):

Es el porcentaje el error absoluto en el pronóstico. Se utiliza para comparar como se ajustan diferentes modelos de pronóstico a la serie de tiempo. Menores valores obtenidos indican un mejor ajuste.

$$MAPE = \frac{1}{n}\sum_{t=1}^{T}\left|\frac{X_t - \hat{X}}{X_t}\right| \quad (2)$$

2.3.1.3. Error Porcentual Absoluto Medio Ponderado (Weighted MAPE):

En función al tamaño del error relativo al valor original se calcula un error MAPE ponderado.

El MAPE pondera con igualdad cada error en el pronóstico, independientemente del nivel que tengan los valores de la serie que se pronostica. Mientras que el WMAPE pondera cada error en dependencia del nivel de valores de la serie.

$$WMAPE = \frac{\frac{\sum_{t=1}^{T}|X_t - \bar{X}| * X_t * 100}{X_t}}{\sum_{T=1}^{T} X_t} \quad (3)$$

El MSE (Mean Square Error) se define como la media de e^2, o también el promedio de los errores entre el estimador y lo que se estima al cuadrado:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \quad (4)$$

Donde:

n: cantidad de muestras

$\hat{y}_t$: estimación de y_t

A partir de la ecuación anterior se deduce que el error medio al cuadrado es la función de pérdida de la medida.

Coeficiente de correlación (R^2):

$$R^2 = \frac{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2} \quad (5)$$

Donde:

$\hat{y}_t$: pronóstico de la demanda

y_t: demanda media

$\overline{y}$: media del pronóstico de la demanda

n: número de observaciones.

De acuerdo con Hyndman, (2014), el más recomendado indicador para comparar la efectividad de los modelos frente a una misma serie temporal es el MAE, porque se calcula e interpreta con facilidad. Es además una medida objetiva para medir el desempeño de los modelos. Su desventaja se nota al ser una medida escala-dependiente por lo que pierde sentido aplicarlo para comparar modelos para diferentes series de tiempo.

El MAPE por su parte es independiente de la escala, siendo una mejor alternativa para comparar modelos entre series. Pero, su efectividad se ve opacada cuando se busca pronosticar valores pequeños pues se producen errores pequeños que generan un gran MAPE porque significan un alto porcentaje del valor real. Para evitar la inclusión de errores en el desarrollo de la investigación, se utiliza el WMAPE, que pondera el error por el porcentaje que representa el valor de la serie en un punto en comparación con el valor real. (Herrera et al., 2010; Silver et al., 2017; Voß & Woodruff, 2006)

Una vez aclarado lo anterior, se determina utilizar los errores MAPE, MAE y RMSE como métricas para comparar el desempeño de los modelos.

2.3.2. Redes Neuronales Artificiales

Las redes neuronales son un modelo matemático, con la capacidad de aprender a través del entrenamiento o la "experiencia". El funcionamiento y comportamiento de la red se memoriza en

un gran número de nodos que finalmente, definen su función. Estos nodos, se encuentran dispuestos en capas y operan de forma paralela en el entrenamiento. (Herrera et al., 2010)

Las redes neuronales artificiales procuran imitar la estructura y funcionamiento de las redes neuronales biológicas para la construcción de sistemas que procesan información. Los componentes se distribuyen de manera jerárquica y pueden adaptarse a los objetos del mundo real, imitando el comportamiento del sistema nervioso biológico. (Rodríguez Aedo, 2016)

Las RNA's se crean con la intención de simular el comportamiento del cerebro humano a con el uso de diferentes softwares. Los ordenadores en la actualidad son capaces de resolver complejos cálculos matemáticos a una velocidad inimaginable para el ser humano. Sin embargo, hay muchas tareas que para el ser humano resultan sencillas, pero por sus características la computadora no es capaz de realizar. (García & Osella Massa, 2003)

2.3.2.1. Antecedentes de Redes Neuronales

Daza (2008) propone el diseño y aplicación de una red neuronal con resultados más ajustados al comportamiento de los datos en comparación con el uso de otros métodos. Según Villada et al., (2012) se identifica que la aplicación de RNA's genera resultados más confiables de pronóstico de series de tiempo que los modelos clásicos.

Toro Ocampo et al., (2004), comparan en su investigación la aplicación de los modelos tradicionales y las Redes Neuronales Artificiales.

Basándose en la información previa o la base de datos histórica para el pronóstico al tratarse de un problema de ventas estacionales, se aplica para analizar las condiciones de la empresa y elaborar un programa de producción estricto y flexible que cumpla con la creciente demanda. (Babel et al., 2006)

El requerimiento de una mayor variedad de productos implica la búsqueda permanente para maximizar el aprovechamiento de los recursos de la empresa, de esta forma se cumplirá eficientemente en los plazos de tiempo determinados, logrando las metas de producción en función a las ventas razón por lo cual las redes neuronales arrojan resultados más cercanos a la realidad. (Matich, 2001)

Sarmiento y Villa (2008), aplican RNA's en el pronóstico de la demanda de energía eléctrica en Colombia, usando redes *MultiLayerPerceptron* con los algoritmos de *Backpropagation* y *Radial Basic Function* para lograr el correcto entrenamiento de la red. Se plantea el pronóstico como un proceso de sistematización de información donde las anteriores redes muestran el desempeño basado en la demanda horaria en megavatios.

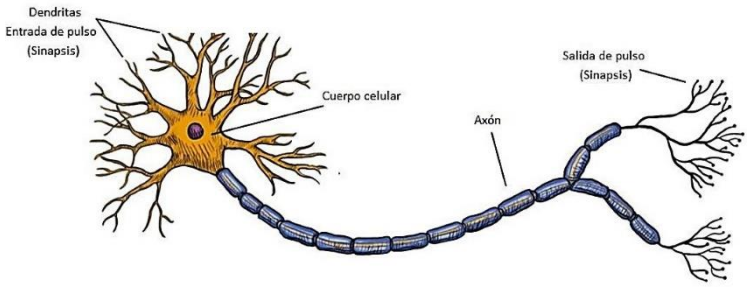
2.3.2.2. Redes Neuronales Biológicas

De acuerdo con Colina y Rivas (1998) las células nerviosas o neuronas conforman los fundamentales elementos del sistema nervioso central. Son capaces de comunicarse con otras neuronas. Este proceso inicia recibiendo mensajes de las neuronal con las que se mantiene conexión, se procesan los mensajes o señales que se transmiten, se generan pulsos nerviosos, se conducen estos pulsos, finalmente se transmiten a otras neuronas. Como se puede apreciar en figura 2 donde se muestra la estructura de una neurona.

El cerebro humano es capaz de procesar con gran rapidez una considerable cantidad de información proveniente de los sentidos a través de señales. Es capaz de compararlas o relacionarlas con información pasada que se guarda como aprendizaje de cada neurona y dar las mejores respuestas, aún en situaciones nuevas. Lo más destacable es la habilidad de aprender con lada estímulo o señal recibida, entregar información o respuestas sin la necesidad de instrucciones explícitas. (Rodríguez Aedo, 2016)

Figura 2

Célula nerviosa



Nota. Adaptado de *Célula nerviosa*, por Rodríguez Aedo, 2016

Biológicamente la red neuronal aprende en el momento en que un impulso o señal es aceptado o no por la neurona receptora, dicho de otro modo, la sinapsis neuronal da paso al aprendizaje de la red.

El aprendizaje de la red se va desarrollando cuando los neurotransmisores inhiben o excitan la neurona, cambiando el grado de influencia que ciertas neuronas tienen sobre otras. Consecuentemente, la estructura de la red y las conexiones neuronales son las que finalmente caracterizan el conocimiento de la red. (Rodríguez Aedo, 2016)

2.3.2.3. Características de las Redes Neuronales Artificiales

La naturaleza, estructura y las características de las redes neuronales artificiales mantienen una gran semejanza con las del cerebro. En ambos casos las redes neuronales tienen la capacidad de almacenar la experiencia, evaluar los casos anteriores con los nuevos, establecer patrones y tendencias. Por ello la importancia y aplicación del estudio de esta herramienta de inteligencia artificial. (Matich, 2001)

La aptitud de aprendizaje adaptativo es una de las características más llamativas que se aprecia en los modelos de redes neuronales. Así las redes aprenden a ejecutar tareas luego del

entrenamiento con ejemplos claros. La red es dinámica porque tiene la capacidad de ir adaptándose a las nuevas condiciones que se presenten. (Hilera Gonzáles & Martínez Hernando, 1995)

La red usa su capacidad adaptativa para autoorganizar la información que recibida durante el aprendizaje y/o la ejecución. La autoorganización implica modificar completamente la red neuronal. Cuando se usan las redes neuronales para el reconocimiento de patrones y tendencias, estas pueden ir autoorganizando la información que ya ha sido usada, por ejemplo, la red Backpropagation genera una representación de sí misma para reconocer los patrones existentes. (Yao, 1999)

De acuerdo con García y Osella (2003) los primeros modelos de programación con la habilidad de tolerancia a fallos son los de redes neuronales. En comparación con los modelos de pronóstico tradicionales que si tienen un pequeño error pierden completamente su utilidad, las redes neuronales si cometen un ligero error en la comunicación de sus neuronas no influyen considerablemente en la funcionalidad del sistema. Para Hyndman (2014) existen dos enfoques fundamentales en la tolerancia a fallos: el primero, la red aprende a reconocer factores respecto a los datos de patrones incompletos y/o distorsionados. Y, por otro lado, la red puede seguir funcionando, aunque una parte de esta se encuentre afectada.

La principal función en las aplicaciones es generalmente realizar procesos con datos de manera muy rápida. Así las redes neuronales se adaptan bien por su ejecución paralela. Los cambios deben ser mínimos en los pesos de las conexiones o entrenamientos para que las redes puedan operar con datos en tiempo real. (Vapnik, 1999)

La facilidad y rapidez con la que las redes neuronales pueden entrenar, comprobar, verificar y trasladar información permite que sean adecuadas para ejecutarse dentro de aplicaciones ya

existentes. Lo que da la oportunidad de usarlas para el mejoramiento de sistemas y evaluación continua de cada paso antes de ser aplicado. (Martí Pérez, 2009)

2.3.2.4. Modelo de una red neuronal

Yao (1999) menciona que, partiendo de la similitud de las redes neuronales al comportamiento biológico, es posible los modelos de la siguiente forma:

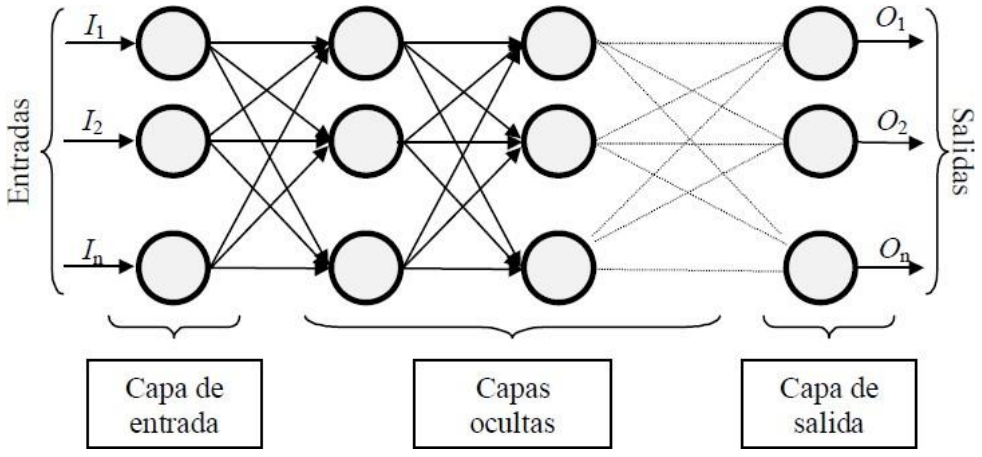
- Tipo biológico: pretende imitar los sistemas neuronales biológicos y simularlos, tomando prioritariamente las funciones auditivas y de visión para su funcionamiento.
- Dirigido a aplicación: considera conjugar el diseño de los sistemas neuronales y la necesidad de utilizar la inteligencia artificial para resolver problemas de simulación de la realidad.

2.3.2.4.1. Elementos básicos que componen una red

En la figura 3 se muestra una red neuronal:

Figura 3

Red neuronal totalmente conectada



Nota. Adaptado de *Ejemplo de una red neuronal totalmente conectada*, por Matich, 2001

Según Martí (2009) y Bonilla (2005) la red está conformada por neuronas que se interconectan y se distribuyen en tres capas (el número de capas puede variar), la información ingresa por la primera capa o capa de entrada, pasando a través de la/s capa/s oculta/s y los resultados salen por la capa de salida.

Las neuronas artificiales también son conocidas como unidades de proceso, y su funcionamiento consiste en recibir la información que entra de las neuronas vecinas y calcular un valor de salida, este es enviado a todas las neuronas restantes. (Hilera & Martínez, 1995)

García, Osella (2003), Sipper y Bulfin (1998) describen a detalle cada uno de los tipos de neuronas que conforman la red mencionando:

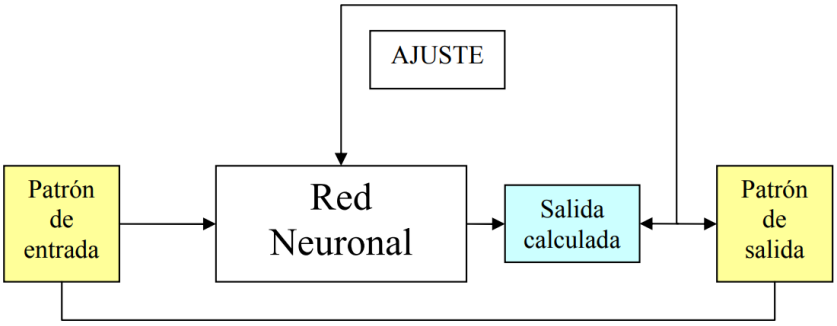
- Neuronas de entrada: reciben la información desde el entorno a través de señales; estas entradas (que a su vez son las entradas a la red neuronal) provienen por lo general de una serie de tiempo como una base de datos histórica con la que se trabaja para hacer predicciones.
- Neuronas de salida: envían un impulso o señal fuera de la red; en la práctica esto corresponde a la predicción o valor futuro estimado.
- Neuronas ocultas: sus entradas y salidas se encuentran dentro del sistema, están en medio de la red, así que no tienen contacto exterior. Las redes neuronales pueden aprender de experiencias que se producen por la entrada y la salida de datos de cada neurona oculta, estas interacciones generan el aprendizaje para la red, Preparándola para obtener la salida correcta cuando nuevas situaciones son encontradas.

2.3.2.4.2. Fase de entrenamiento

Patiño et al., (2020) dice en su trabajo que el aprendizaje o entrenamiento de la red busca determinar los pasos que facultan a la red para resolver adecuadamente una situación determinada. En la fase de aprendizaje se va desarrollando de manera iterativa la solución hasta conseguir el nivel operativo apropiado como se muestra en la figura 4.

Figura 4

Aprendizaje supervisado de un proceso



Nota. Adaptado de *Esquema de funcionamiento de un proceso de aprendizaje supervisado*, por Isasi & Galván, 2004

2.3.3. Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (*SVM*) es uno de los métodos más poderosos del aprendizaje automático, que a pesar de su simplicidad ha demostrado ser un algoritmo robusto, que sistematiza y afronta bien problemas de la vida real (Gala, 2013). Se considera una alternativa eficiente ante las limitaciones del modelo de Redes Neuronales Artificiales (*RNA*) cuando se presentan dificultad por la dimensionalidad y/o ruido en los datos. Hay estudios donde se determina que las SVM tienen mayor precisión a los modelos autorregresivos de medias móviles (Jaramillo, 2015).

Este modelo fue desarrollado por Vapnik (1999) y sus colaboradores en el marco de la Teoría de Aprendizaje Estadístico (*SLT*). Este método ha sido estudiado vigorosamente en los

últimos años y aplicado con éxito en una gran variedad de situaciones como: estimación de densidades probabilísticas, y la predicción de series de tiempo (Cuevas Alfaro, 2010).

Inicialmente las SVM se desarrollaron con la finalidad de ser utilizadas como clasificadores binarios. La tarea de clasificación está dividida en dos fases: la fase de aprendizaje automático y la fase de reconocimiento. (Vapnik, 1999)

Primero se selecciona los datos para el entrenamiento, se establecen los atributos y características del espacio de entrada y se entrena el clasificador. Este entrenamiento genera parámetros w que define al clasificador y un hiperplano de separación óptimo (*HSO*). En segunda fase, la etapa de reconocimiento el modelo ya entrenado asigna a los datos que ingresan una de las clases, de acuerdo con la región de clasificación en la se hayan mapeado los nuevos datos. (Cuevas Soto et al., 2019)

El aprendizaje según Farías (2011), se consigue buscando alguna dependencia funcional para un conjunto de vectores y los datos de entrada-salida, lo que permite encontrar el posible espacio más amplio así se puedan separar los datos en conformidad con clase que les corresponde. La particularidad de las SVM puede comprenderse sin el uso de fórmulas, pero es necesario el entendimiento de cuatro conceptos básicos: el hiperplano de separación, el hiperplano óptimo, el margen suave y la función núcleo o kernel.

2.3.3.1. SVM lineales.

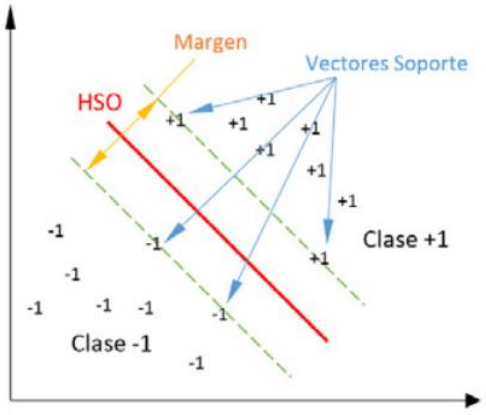
Hiperplano de separación óptimo y margen máximo según González et al., (2017). Las SVM son capaces de aprender partiendo de un conjunto de N muestras experimentales a lo que se le denomina conjunto de entrenamiento: $\{(x_1,\ y_1), (x_2,\ y_2), \dots, (x_n,\ y_m)\}$

Donde cada muestra $(x_i,\ y_i)$ para $i = 1, \dots, N$ está conformado por un vector de n características $x_i\ \epsilon\ R^n$ y a una etiqueta $y_i\ \epsilon\ R$ que indica la clase $\{\pm 1\}$ a la que pertenece cada

elemento muestral. La finalidad es localizar la función $f: R^n \rightarrow \{\pm 1\}$ que seleccione los datos en dos clases.

Figura 5

Hiperplano de separación óptimo desarrollado por las SVMs



Nota. Adaptado de *Hiperplano de Separación Óptimo implementado por las SVM*, por González et al., 2017

En problemas linealmente separables hay muchos hiperplanos en función del mapeo y clasificación de las observaciones, pero las SVM no hallan cualquier hiperplano sino únicamente el que maximiza la distancia entre el hiperplano y la observación más próxima de cada clase como se puede ver en la figura 5. Según Ing et al., (2019) el hiperplano de separación óptimo (HSO) se crea por el margen máximo de separación entre las dos clases. Se toma como referencia la denominación de la figura 5, hay dos hiperplanos paralelos al HSO ($w * x + b = 0$) que delimitan las muestras a los dos lados de cada clase: $w * x + b = +1$ y $w * x + b = -1$. la distancia entre los hiperplanos paralelos y el HSO establece el margen máximo cuyo resultado geométrico corresponde a $2/\|w\|$.

Como menciona Fernandes y sus colaboradores (2015) hallar el mejor hiperplano de separación es un clásico problema de maximización con restricciones lineales, quedando para su resolución planteadas las siguientes ecuaciones:

$$\text{min:} \quad \frac{\|w\|^2}{2} \tag{6}$$

$$\text{sujeto a:} \quad y_i[(w * x_i) + b] \geq 1$$

para resolverlo se puede aplicar los multiplicadores de Lagrange:

$$L(w, b, \propto) = \frac{\|w\|^2}{2} - \sum_{i=1}^{N} \propto_i \{y_i[(w * x_i) + b] - 1\} \tag{7}$$

Para encontrar la solución es preferible un espacio dual intentando que se dependa únicamente del producto escalar de los patrones de entrada, lo que permite simplificar (Fernandes et al., 2015):

$$\text{min:} \quad L_d(\propto) = \sum_{i=1}^{N} \propto_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \propto_i \propto_j y_i y_j \left(x_i^T * x_j\right) \tag{8}$$

$$\text{sujeto a:} \quad \sum_{i=1}^{N} \propto_i y_i = 0 \text{ y } \propto_i \geq 0, i \, \epsilon \{1, \dots, N\}$$

Este planteamiento del problema satisface las condiciones de Karush-Kuhn-Tucker (*KKT*) donde se tienen las condiciones suficientes para que un valor extremo exista según Mora (2001). De la expresión en forma matricial que se mencionan Statnikov y demás investigadores (2011) en su obra

$$\text{min:} \quad L_d(\propto) = \frac{1}{2} \propto^T H \propto - f^T \propto \tag{9}$$

$$\text{sujeto a:} \quad y^T \propto = 0, \propto \geq 0$$

en donde se usa el vector unitario $f = [11 \dots 1]^T$ y utilizando algún método de optimización se encuentra el vector de multiplicadores $(\propto_0 = \propto^0{}_1, \propto^0{}_2, \propto^0{}_3, \dots, \propto^0{}_N)$ y finalmente se determina el vector normal w_0 y el bias b_0 del HSO.

$$w_0 = \sum_{i=1}^{N} y_i \propto_i^0 x_i \text{ y } b_0 = -\frac{1}{2} w_0 * [x_T + x_N] \tag{10}$$

indicando que w_0 expresable como una combinación lineal de N vectores de entrada. De las condiciones KKT salen la mayoría de los multiplicadores de Lagrange $\propto_0$, una cantidad inferior de vectores del conjunto de entrada N_{sv} intervienen en la combinación lineal que genera los vectores de soporte. En la expresión anterior *Xr* y *Xs* son vectores soporte, uno de cada clase (Cuevas Alfaro, 2010). La expresión final del clasificador buscado quedaría como

$$f(x) = sign(w_0 * x + b_0) = sign[\sum_{i\epsilon SV} y_i \propto_i^0 (x_i x) + b_0] \quad (11)$$

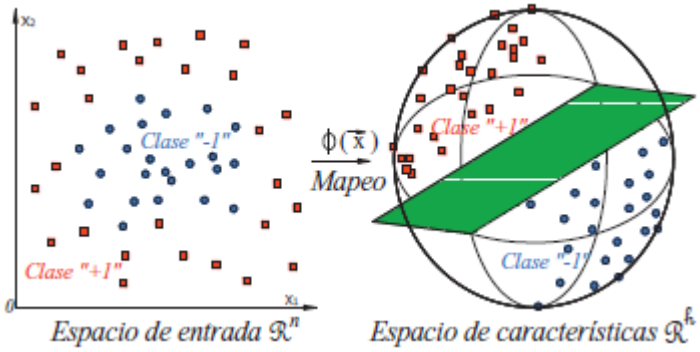
según Nievas (2016) donde el signo que resulte de la función determinará la clase a la que pertenece un dato determinado. Hay que tomar en cuenta que la sumatoria no se da sobre todos los puntos de entrenamiento N, sino sólo sobre los vectores soporte (SV) siendo la cantidad de puntos evaluados generalmente menor.

2.3.3.1. SVM no lineales con Kernel

Generalmente los eventos reales tienen un comportamiento no separable linealmente lo que se dificulta la definición del HSO. La figura 6 muestra un grupo de datos que no pueden ser separados linealmente por un hiperplano en R^n, pero si en un espacio de mayor dimensión R^h. (Statnikov et al., 2011)

Figura 6

Mapeo de datos con comportamiento no lineal



Nota. Adaptado de *Mapeo de datos a una mayor dimensión y separación lineal de las clases en el nuevo espacio*, por Cuevas Soto et al., 2019

Para Yeh y Lien (2009) cuando la superficie lineal de decisión dentro del espacio original de los datos no es apropiada, se mapea el vector de entrada en un espacio más amplio R^h o también denominado espacio de características. Se realiza la transformación de $R^n \rightarrow R^h$ realizando el mapeo y buscando el HSO, de acuerdo con la metodología ya descrita, que será lineal dentro de R^h, pero que representaría un espacio no lineal en R^n. Para establecer este tipo de proyección hacia un espacio de característica se utilizan las funciones llamadas kernels: $K(x_i, x_k) = {}^{\Phi}(x_i) * {}^{\Phi}(x_k)$. Las funciones kernel o núcleo permiten desarrollar las operaciones algebraicas en R^h. Si se considera así cualquier técnica de análisis multivariado para datos $x \epsilon R^n$ que se permita reformular un algoritmo computacional en productos escalares, se puede utilizar genéricamente en los datos transformados con las funciones kernel (Farías, 2011).

Existen varias funciones núcleos o kernel, según Velásquez y otros autores (2010) de donde se pueden destacar las siguientes consideradas básicas:

Kernel lineal:

$$K(x_i, x_j) = {x_i}^T * x_j \tag{12}$$

Kernel polinomial:

$$K(x_i, x_j) = \left(p + Y{x_i}^T * x_j\right)^d; Y > 0 \tag{13}$$

Kernel gaussiano RBF:

$$K(x_i, x_j) = exp\left(-Y\left\|x_i - x_j\right\|^2\right); Y > 0 \tag{14}$$

donde *Y, d y* p son elementos de producto escalar. En el caso no lineal se debe utilizar un kernel para buscar el clasificador buscado, cuya expresión seria

$$f(x) = sign[\sum_{i \epsilon SV} y_i \propto_i K(x, x_i) + b] \quad (15)$$

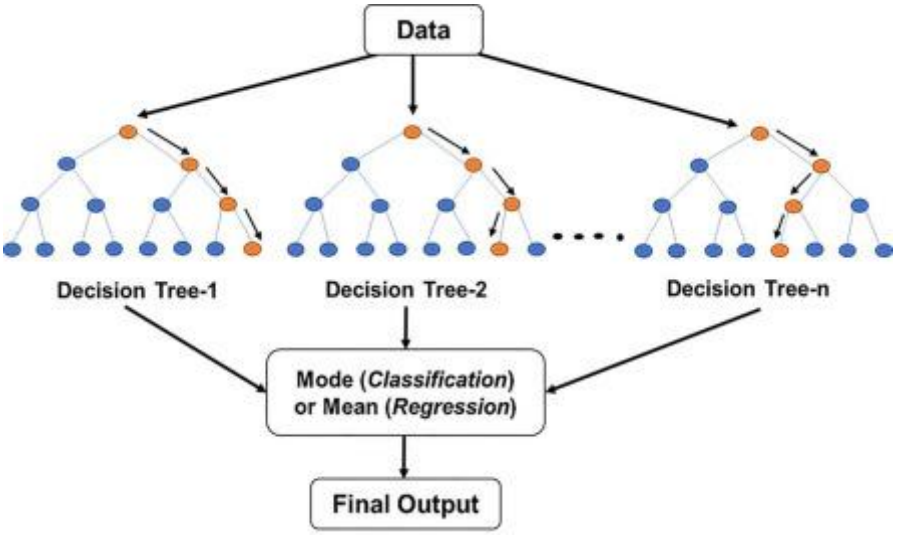
En general, la función de kernel gaussiano RBF es una buena primera elección. Esta función no lineal permite mapear las muestras a un espacio de mayor dimensión, y a diferencia del kernel lineal puede ayudar cuando la relación entre las clases y las características es no lineal (Nievas Lio, 2016).

2.3.4. Bosques Aleatorios

Bosques Aleatorios o Random Forest son un clasificador que se conforma por un grupo de árboles de decisión repartidos de manera idéntica. Sus aplicaciones pueden ser tanto para tareas de clasificación como regresión usando voto mayoritario y ponderación respectivamente (Patiño Pérez et al., 2020). Para Sharma y Kumar (2016) la combinación de dichos árboles formando un bosque dentro de ciertas condiciones genera un mejor resultado siendo un método más preciso, estable, dinámico que busca el equilibrio entre la varianza y el sesgo del bosque.

Figura 7

Arquitectura de Random Forest



Nota. Adaptado de *A general random forest architecture*, por Lindner, 2017

De acuerdo con trabajo de Glocker y sus colaboradores (2016) un modelo de bosques aleatorios está constituido por un grupo de árboles de decisión (*ensemble*), cada uno entrenado con una muestra aleatoria de los datos de entrenamiento originales mediante *bootstrapping*. Se plantea que cada árbol usa datos ligeramente distintos para el entrenamiento. En cada árbol, las entradas se van ordenando por bifurcaciones (nodos) generando la estructura del árbol hasta obtener un nodo terminal. La predicción de un nuevo dato se obtiene agregando las predicciones de todos los árboles que forman el modelo, conformando el bosque aleatorio (Bravo Sanzana et al., 2015).

Para comprender el funcionamiento de los Bosques Aleatorios se necesita comprender primero los conceptos de ensemble y bagging.

2.3.4.1. Métodos de ensemble

Todos los modelos de aprendizaje de tipo *machine learning* tienen la dificultad de conseguir el equilibrio entre el sesgo estadístico y la varianza. Fontanazza y los demás autores (2014) del trabajo describen que:

- El sesgo estadístico se refiere a la media de cuánto se alejan las predicciones de un modelo en comparación con los valores reales. Indica cuan efectivo es el modelo de aprender la relación entre los resultados y los datos de la muestra.
- En cuanto a la varianza permite medir cuánto cambia el modelo en función de los datos utilizados para su entrenamiento. Lo ideal sería que un modelo no se modifique mucho por las pequeñas variaciones en los datos de entrenamiento, lo que indica que el modelo está memorizando los datos no aprendiendo la relación que tienen los predictores y la variable respuesta.

Adamowski y Chan (2011) mantienen que la complejidad de un modelo es directamente proporcional a la flexibilidad de este para adaptarse a las observaciones. Así cuando estos se

incrementan se reduce el sesgo y se mejora la capacidad de predicción. Sin embargo, cuando se alcanza un cierto grado de flexibilidad, aparece el problema de *overfitting*. En cuyo caso el modelo se adapta tanto a los datos de entrenamiento que no puede predecir con eficiencia nuevos datos. El mejor modelo es aquel que logra un equilibro óptimo entre bias y varianza.

Jere y sus colaboradores (2017) estiman que generalmente los árboles con pocas ramificaciones poseen pequeña varianza, pero no son capaces de representar apropiadamente la relación entre las variables, porque tienen el sesgo alto. Por el contrario, el comportamiento del modelo cuando los árboles tienen muchas ramificaciones se ajusta mucho a los datos de entrenamiento, presentando reducido sesgo pero abundante varianza. Una manera de dar solución a este problema son los métodos de *ensemble*.

Para Schonlau y Zou (2020) los métodos de *ensemble* acoplan varios modelos en uno nuevo con la finalidad de conseguir un equilibrio entre sesgo y varianza, obteniendo predicciones más acertadas en comparación con los modelos individuales. Dos de los más utilizados tipos de *ensemble* son:

- Bagging: según Amat (2017) este tipo de *ensemble* se ajusta a diversos modelos, cada uno con una muestra distinta de observaciones para el entrenamiento. Para conseguir el pronóstico, se forma un agregado donde todos los modelos aportan con su propia predicción. Finalmente, como resultado, se considera la media de todas las predicciones cuando son variables continuas o la moda si las variables son categóricas. Los modelos de bosques aleatorios están inmersos de esta clasificación.
- Boosting: Se acoplan secuencialmente modelos sencillos, de tal manera que cada nuevo modelo aprende de los errores del modelo anterior. Como resultado, igual

que en bagging, se toma la media de todas las predicciones cuando las variables son continuas o la clase más frecuente para variables cualitativas (Amat Rodrigo, 2017).

Para que los métodos de *ensemble* logren mejores resultados en comparación a los modelos individuales que los conforman, estos deben ser lo más diversos posibles de tal forma que sus errores no estén correlacionados.

A continuación, se detalla la estrategia de bagging en donde se fundamenta el modelo Random Forest.

2.3.4.2. Bagging

El término *bagging* es el diminutivo de *bootstrap aggregation*, y se refiere al uso del muestreo repetido con reposición con el objetivo de reducir la varianza de algunos modelos de aprendizaje estadístico, entre ellos los que utilizan árboles de decisión. (Patiño Pérez et al., 2020)

De acuerdo con lo que mencionan Schonlau y Zou, (2020) se presentan n muestras de datos independientes $Z_1, \dots, Z_n$ cada una con varianza σ^2, la varianza de la media de las observaciones es $\bar{Z}$ igual a σ^2/n. Dicho de otra forma, se promedia un grupo de observaciones para reducir la varianza. Con base en esta idea, una manera de reducir la varianza e incrementar la precisión de un modelo predictivo es obtener varias muestras de la población, ajustando un modelo distinto con cada una de las muestras, y hacer la media o la moda si se trabaja con variables cualitativas, de las predicciones obtenidas (Amat Rodrigo, 2017). En la aplicación de este tipo de modelos no es común tener acceso a varias muestras, se puede optar por simular el proceso utilizando a bootstrapping, generando *pseudo-muestras* para ajustar diferentes modelos y después agregarlo los resultados de cada uno. A este procedimiento se le conoce como *bagging* y se puede aplicar a gran variedad de métodos de regresión.

Para el caso específico de los árboles de decisión, por su resultado de bajo sesgo y alta varianza, el *ensemble bagging* tiene muy buenos efectos. Según Arroyo (2008) la forma de aplicarlo es:

1. Generar *B pseudo-training sets* mediante *bootstrapping* partiendo de la muestra de entrenamiento inicial.
2. Entrenar un árbol con cada *B* muestra obtenida en paso 1. Cada árbol se genera y no se somete a *pruning*, por ende, tiene alto valor de varianza y poco sesgo. Generalmente la única restricción es el número mínimo de datos que deben tener los nodos terminales. El valor óptimo se puede obtener comparando el error de predicción del modelo.
3. Para cada nuevo dato, se obtiene la predicción de cada *B* árbol. El resultado de la predicción se obtiene calculando la media de las *B* predicciones para las variables cuantitativas y como la moda para variables cualitativas.

Para el *bagging*, cuando se consigue un cierto número de árboles, la reducción de la prueba de error se estabiliza. Por lo que es recomendable generar solamente los necesarios.

2.3.5. Algoritmo K Vecinos más Cercanos

El algoritmo k-NN es un método de aprendizaje automático basado en peticiones (Aha et al., 1991). Lo que indica que las predicciones que genera el algoritmo provienen directamente del propio entrenamiento sin estimar ningún modelo. González y sus colegas (2016) consideran que en su mayoría la información se almacena en memoria, y el proceso de predicción analiza todas las peticiones conocidas. Por ende, generar la predicción puede resultar costoso hablando en términos de memoria y ejecución, puesto que ambos crecen de forma lineal con la cantidad de datos a analizar. (Tao et al., 2022)

No se genera ningún modelo durante el entrenamiento. Esto porque se realiza una aproximación parcial de forma localizada, en lugar de aproximar una función por completo, De esta manera, la obtención de los resultados a través de los cálculos se delega al último momento donde se asigna el resultado. Este tipo de modelos se conocen como de aprendizaje perezoso (Arroyo Gallardo, 2008; Wettschereck et al., 1997).

2.3.5.1. Descripción del k-NN para series temporales

En un resumen breve, el método del k-NN consiste en buscar dentro de un grupo de observaciones, cuyos componentes son similares, que se busca predecir usando una cierta medida de distancia. Se escogen de entre aquellos elementos los k elementos que conservan mayor similitud con el elemento a predecir. Donde finalmente, se realiza la predicción en función del valor siguiente de dichos elementos. (Sahoo et al., 2009)

El algoritmo k-NN es un modelo versátil que puede usarse para problemas de clasificación y de regresión. En el caso de regresión se puede utilizar para predecir series temporales. La particularización del algoritmo para series de tiempo se describirse de la siguiente forma según Arroyo (2008):

1. La serie temporal debe estar definida como $\{y_1, y_2, \dots, y_{n-1}, y_n\}$ donde n es la longitud que pasa a ser elementos de longitud d. De tal manera que se tiene una serie definida como $y_t^d = \{y_{(t+1)-d}, y_{(t+2)-d}, \dots, y_{t-1}, y_t\}$.
2. Se calcula las distancias entre el elemento que se quiere predecir $y_n^d = \{y_{(n+1)-d}, y_{(n+2)-d}, \dots, y_{n-1}, y_n\}$ y todos los elementos previos a este en la serie.
3. Se ordena los elementos de conformidad con la distancia obtenida y se seleccionan los k más cercanos. Estos k elementos se denotan como $y_{t_1}^d, y_{t_2}^d, \dots, y_{t_{k-1}}^d, y_{t_{k1}}^d$.

4. Se calcula los valores sucesivos a cada uno de los k seleccionados y se obtiene la predicción como la media ponderada de esos valores

$$\hat{y}_{n+1} = \frac{\sum_{i=1}^{k} w_i * y_{t_i}^{d}}{\sum_{i=1}^{k} w_i}$$

siendo $\hat{y}_j$ la predicción en el $j - ésimo$ instante de tiempo y w_i el peso vinculado al siguiente valor del $i - ésimo$ vecino.

2.3.5.2. Ensamblado de k-NN para series temporales

Consiste en el uso de varios predictores que conjuntamente prometen conseguir mejor desempeño del modelo que haciéndolo de forma individual. De manera que la predicción obtenida individualmente con cada método haga aportes al resultado final, datos que los otros modelos no consiguen modelar. El número de métodos presentes en el ensamble es inversamente proporcional al ruido presente en el resultado final. No obstante, incrementar la cantidad de modelos también supondrá que las predicciones obtenidas sean menos heterogéneas. Por consiguiente, la cantidad de métodos elaborados y utilizados durante el ensamble es un factor que permite el suavizado de las predicciones.

Por lo simple que es el método, se sugiere agrupar diferentes combinaciones de $k - d$. Para conseguirlo, se inicia generando dichas combinaciones, lo que se desarrollará mediante diferentes estrategias planteadas. Cuando ya se ha seleccionado el conjunto de combinaciones que formarán el conjunto de datos predictores, se realiza la predicción usando ensamble. Entonces, se crearán las predicciones de cada modelo. Cuando se han calculadas todas ellas, se realiza una media ponderada de los resultados y esa es la predicción mediante ensamble. (Habadi & Tsokos, 2017; Tashman, 2000)

Capítulo III

Diagnóstico de la situación actual

3.1. Descripción de la organización

La Junta Administradora de Agua Potable (JAAP) "La Portada" tiene bajo su responsabilidad la administración, operación y servicio eficiente del sistema de agua potable. Fue conformada en el año 2008 luego que el entonces Municipio del Cantón Mira realizara la obra civil del sistema de agua potable "La Portada" que beneficiaría a ochenta moradores de las comunidades La Portada y San Marcos. La JAAP se conforma por seis miembros que se eligen por voto mayoritario entre los moradores de la comunidad que cumplan los requisitos estipulados en el Reglamento Interno de la Junta Administradora de Agua Potable de "La Portada" (Reglamento Interno de La Junta Administradora de Agua Potable de "La Portada," 2010)

Luego de 14 años de operación la JAAP La Portada suministra el servicio a los sectores de La Portada, San Marcos, Cooperativa, San Nicolás y debido al crecimiento de la comunidad La Playita se analiza la factibilidad de ampliar la red de distribución hasta este sector.

3.2. Misión

Estamos comprometidos a generar bienestar a nuestra comunidad, ofreciendo un servicio de agua potable con eficiente y sostenible. Elevando la calidad de vida de la comunidad, mediante la provisión, en calidad y cantidad de agua potable preservando las fuentes naturales y el medio ambiente cumpliendo las normas del GAD Mira.

3.3. Visión

Ser la organización líder en gestión sostenible y responsable del recurso hídrico apto para el consumo humano indispensable para la vida en el cantón Mira, a través de la prestación de servicios oportunos, continuos y de calidad en beneficio de todos los habitantes de la comunidad.

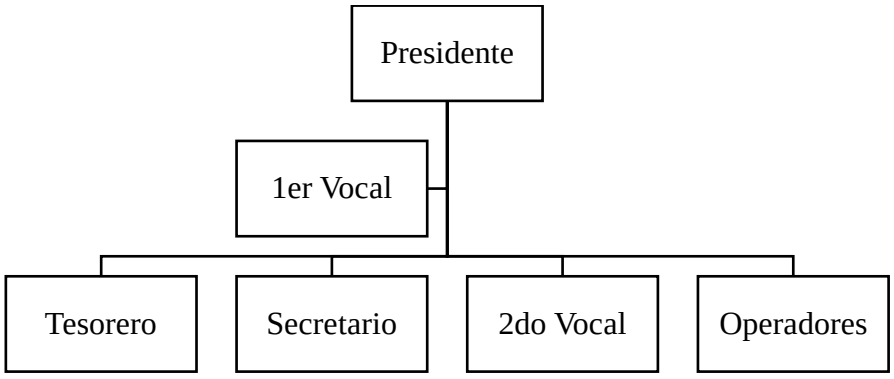
3.4. Objetivos

- Aportar a la salud y bienestar de la comunidad, suministrando con calidad y eficiencia el servicio de agua potable. Formando un equipo humano capaz, comprometido y solvente.
- Colaborar con el desarrollo, la calidad de vida y el cuidado del entorno ecológico de las fuentes naturales de agua, a través del desempeño integral de nuestro personal.
- Lograr el abastecimiento a todos los rincones del sector, con enfoque en la satisfacción de las necesidades de los moradores, responsabilidad social y ambiental.
- Garantizar el servicio de suministro de agua potable en los años venideros.

3.5. Organigrama

Figura 8

Estructura Orgánica de la JAAP La Portada



Nota. El organigrama de la organización se elabora a partir del Reglamento Interno de La Junta Administradora de Agua Potable de “La Portada” 2010) donde se mencionan los miembros que se debe elegir para la conformación de la junta administradora de de agua potable por la autora

3.6. Descripción del proceso de productivo

3.6.1. Captación y Conducción

La captación del agua bruta se realiza en una fuente natural de agua ubicada en el sector de La Calera entre El Garrapatal y La Cocha. Se entuba el agua con la ayuda de un canal y la caída

natural del terreno. Pasa hacia una bomba de 200 HP que tiene una capacidad de bombeo de 150 litros/s, controlando la variación de frecuencia del caudal que llega al tanque te tratamiento. La extensión de la tubería de conducción es de 12 Km y contiene 5 filtros a lo largo de toda su trayectoria para separar las impurezas gruesas y finas existentes en el recurso hídrico.

3.6.2. Producción

El agua que llega al tanque de recolección y tratamiento donde el agua pasa a por un proceso de cloración que realiza el operador del sistema, así el agua cumple con los requisitos para el consumo humano. El agua potabilizada se conduce con la ayuda de una bomba de 150 HP hacia el tanque de almacenamiento para su posterior distribución. La tubería de conducción del tanque de tratamiento al tanque de almacenamiento es de 20 metros.

3.6.3. Distribución

La red de distribución es un sistema de tuberías desde el tanque de almacenamiento hasta cada uno de los consumidores. Para esta red se utiliza una bomba de 100 HP y la caída natural del terreno que impulsa el caudal de agua hacia cada uno de los puntos de conexión y medición para los consumidores.

3.7. Producción y consumo anual

La JAAP La Portada cuanta con un solo sistema de agua potable que capta un caudal de 150 litros/s de la fuente y se distribuye de acuerdo con las necesidades y consumos de los moradores. Para la correcta administración y control de la información, se lleva a cabo un registro mensual, desde el inicio de las funciones de la JAAP, del consumo del recurso hídrico de cada uno de los usuarios. Esto ayuda a identificar el comportamiento que tiene el consumo de agua potable y del crecimiento del sistema por el incremento de consumidores. A continuación, se muestra la tabla 1 que presenta los consumos mensuales del recurso hídrico en m^3.

Tabla 1

Volumen $(\boldsymbol{m^3})$ *de consumo mensual total desde el año 2016 hasta el 2021.*

Meses	***2016***	***2017***	***2018***	***2019***	***2020***	***2021***
Enero	*502*	*639*	*726*	*861*	*981*	*1095*
Febrero	*508*	*645*	*711*	*868*	*984*	*1113*
Marzo	*536*	*642*	*748*	*882*	*1018*	*1128*
Abril	*497*	*607*	*829*	*907*	*1056*	*1143*
Mayo	*506*	*661*	*815*	*916*	*1067*	*1179*
Junio	*569*	*688*	*807*	*923*	*1094*	*1187*
Julio	*587*	*719*	*843*	*946*	*1085*	*1182*
Agosto	*612*	*723*	*859*	*994*	*1084*	*1164*
Septiembre	*594*	*705*	*842*	*972*	*1093*	*1193*
Octubre	*603*	*697*	*861*	*975*	*1104*	*1185*
Noviembre	*629*	*708*	*867*	*989*	*1116*	*1218*
Diciembre	*647*	*716*	*882*	*993*	*1128*	*1207*
Total	*6790*	*8150*	*9790*	*11226*	*12810*	*13994*

3.8. Volumen de almacenamiento

El tanque recolección y tratamiento tiene capacidad de 50 m^3, el agua que ha tenido un proceso de potabilización pasa al tanque de almacenamiento que tiene una capacidad de 100 m^3. Desde este punto el agua potable es impulsada por una bomba hacia la red de distribución que tiene tres brazos principales destinados a las comunidades de La Portada, San Marcos, Cooperativa y San Nicolas donde está ubicada la Planta de Producción de Uyamafarms, allí existe uno de los mayores consumos de agua potable que tiene el sistema.

3.9. Datos de la población

La población de La Portada con el pasar de los años ha ido incrementando, según las proyecciones del equipo técnico de análisis del Instituto Nacional de Estadísticas y Censos, considerando que para el año 2022 la localidad contará con 442 habitantes, que corresponden al 3.51% de los habitantes de todo el cantón Mira que se prevé sean 12611 para el mismo año; la

tabla 2 presenta la proyección de la cantidad de habitantes clasificados por edad desde el 2016 hasta el 2022.

Tabla 2

Proyección poblacional de La Portada 2016-2022.

Grupo de habitantes por edad	***2016***	***2017***	***2018***	***2019***	***2020***	***2021***	***2022***
0-4	*33*	*35*	*41*	*42*	*40*	*41*	*43*
5-9	*36*	*38*	*44*	*46*	*43*	*45*	*47*
10-14	*37*	*40*	*46*	*48*	*45*	*47*	*49*
15-19	*32*	*36*	*40*	*43*	*40*	*41*	*43*
20-24	*26*	*30*	*34*	*36*	*34*	*35*	*36*
25-29	*24*	*28*	*31*	*33*	*31*	*32*	*33*
30-34	*23*	*26*	*29*	*31*	*29*	*30*	*31*
35-39	*20*	*25*	*27*	*29*	*27*	*28*	*29*
40-44	*17*	*22*	*23*	*26*	*23*	*25*	*25*
45-49	*13*	*19*	*19*	*22*	*20*	*21*	*21*
50-54	*12*	*15*	*16*	*18*	*16*	*17*	*18*
55-59	*10*	*13*	*14*	*15*	*14*	*15*	*15*
60-64	*10*	*12*	*13*	*14*	*13*	*13*	*14*
65-69	*8*	*10*	*11*	*12*	*11*	*12*	*12*
70-74	*6*	*9*	*9*	*10*	*9*	*9*	*10*
75-79	*5*	*6*	*7*	*7*	*7*	*7*	*7*
80 y más	*7*	*7*	*8*	*9*	*8*	*8*	*9*
Total	*319*	*371*	*412*	*441*	*410*	*426*	*442*

Nota. Adaptado de *Análisis Información Censal*, por INEC, 2022

3.10. Análisis de autocorrelación

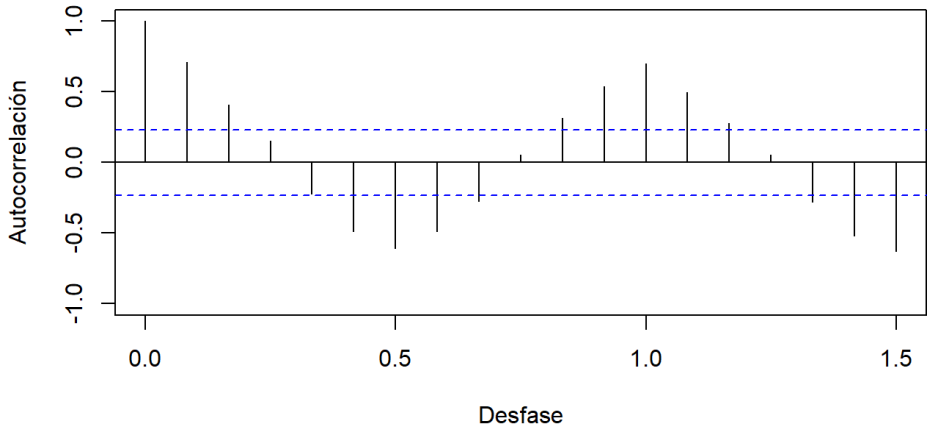
Se revisan los patrones existentes en función de dos: demanda mensual y la cantidad de clientes mensual. Se realiza este análisis con el fin de verificar si el volumen de consumo del recurso hídrico por parte de la población se ha incrementado o reducido.

3.10.1. Análisis de autocorrelación de la demanda mensual de agua potable

En el análisis realizado a la demanda de agua potable se identifica que la serie de tiempo muestra una tendencia creciente, como se muestra por lo lejanos que son los valores resultantes de los primeros tres retardos de la serie, igual como se observa en la figura 9.

Figura 9

Autocorrelación de la demanda de agua potable

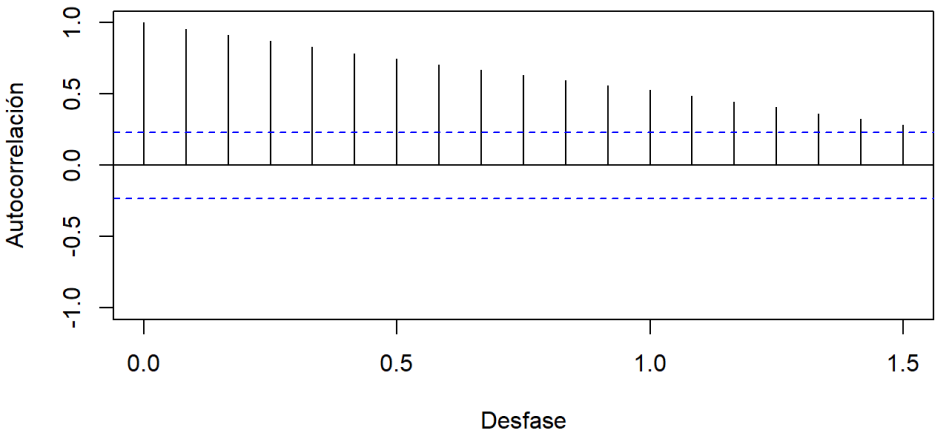


3.10.2. Análisis de autocorrelación de los clientes

El análisis de los clientes mensuales de los últimos seis años proporciona un resultado que evidencia la existencia de tendencia creciente, que como en el análisis anterior los valores obtenidos de la autocorrelación son diferentes de cero y se alejan en los primeros retardos de la serie. Mostrando que efectivamente los clientes que utilizan este recurso hídrico se han incrementado año tras año. Ver figura 10.

Figura 10

Autocorrelación de los clientes



3.11. Análisis de estacionalidad de la serie de tiempo

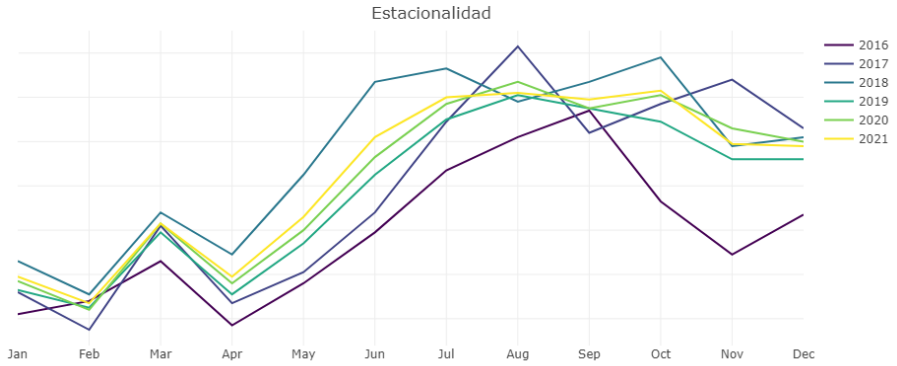
El análisis de estacionalidad que se realiza a la demanda de agua potable y a la cantidad de clientes existentes en los seis últimos años, permite observar el comportamiento de la serie de tiempo y detectar patrones en la misma.

3.11.1. Análisis de estacionalidad de la demanda mensual de agua potable

Para la base de datos históricos de la demanda de agua potable que corresponde al consumo de los usuarios de los últimos seis años, se realiza el análisis de estacionalidad donde se muestra estacionalidad en los periodos de tiempo, debido a que el comportamiento de la demanda ha sido similar en los primeros cuatro meses del año, luego se observa un incremento progresivo y se identifica un segundo patrón para los meses de agosto y septiembre, véase en la figura 11 para tener información más ilustrativa.

Figura 11

Análisis de estacionalidad de la demanda de agua potable

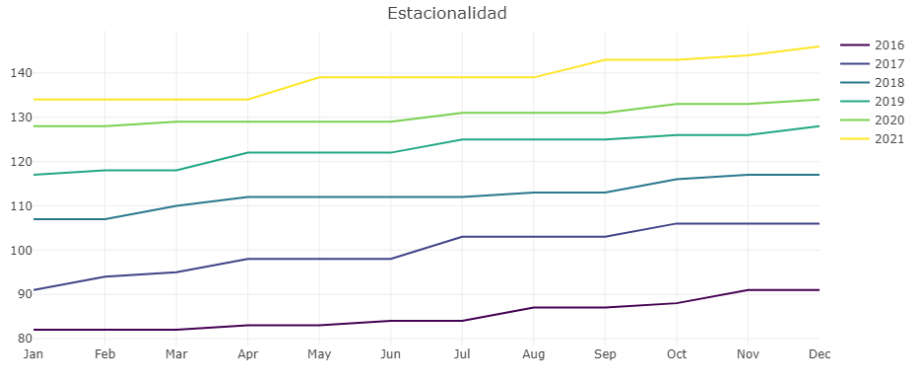


3.11.2. Análisis de estacionalidad de los clientes

En la base de datos históricos de la cantidad de clientes de los últimos seis años, se realiza el análisis de estacionalidad donde no se identifica estacionalidad en los periodos de tiempo, ya que los clientes se han incrementado progresivamente como se puede apreciar en la figura 12.

Figura 12

Análisis de estacionalidad de los clientes



Capítulo IV

Modelo de pronóstico

Para la selección del modelo de pronóstico se recopilan los datos de la serie temporal que se utilizará para alimentar y entrenar los modelos de inteligencia artificial, en este caso corresponden al consumo mensual de agua desde el 2016 hasta el 2021. Antes de ejecutar los modelos de pronóstico se analiza la autocorrelación y estacionalidad de la serie de tiempo.

Los modelos de pronóstico de inteligencia artificial que se emplean son: Redes Neuronales Artificiales, K Nearest Neighbor Regression, Support Vector Machine y Random Forest.

4.1. Selección de la variable

Para garantizar la correcta ejecución de los modelos y obtención de los pronósticos, se definen los datos de la demanda de agua potable, puesto que estos valores alimentan los modelos y se trabaja sobre estos en los diferentes modelos.

- Variable para pronosticar: demanda de agua potable de La Portada
- Tipo de Variable: numérica o cuantitativa
- Unidad de medida: metro cúbico *(m^3)*

4.2. Obtención de datos

La Junta Administradora de Agua Potable de La Portada posee el registro de los consumos mensuales de cada uno de los clientes. Así se puede recopilar los volúmenes de consumo individual y posteriormente el volumen mensual total consumido y a la para la cantidad total de clientes. Los registros se llenan manualmente en físicos, para poder usar estos datos se elaboró una base de datos en Excel que alimenta los modelos de pronóstico. A continuación, en la tabla 3 se muestran los clientes y en la tabla 1 se muestran los consumos desde el 2016 hasta diciembre del 2021.

Tabla 3

Clientes mensuales desde el año 2016 hasta el 2021.

Meses	***2016***	***2017***	***2018***	***2019***	***2020***	***2021***
Enero	*82*	*91*	*107*	*117*	*128*	*134*
Febrero	*82*	*94*	*107*	*118*	*128*	*134*
Marzo	*82*	*95*	*110*	*118*	*129*	*134*
Abril	*83*	*98*	*112*	*122*	*129*	*134*
Mayo	*83*	*98*	*112*	*122*	*129*	*139*
Junio	*84*	*98*	*112*	*122*	*129*	*139*
Julio	*84*	*103*	*112*	*125*	*131*	*139*
Agosto	*87*	*103*	*113*	*125*	*131*	*139*
Septiembre	*87*	*103*	*113*	*125*	*131*	*143*
Octubre	*88*	*106*	*116*	*126*	*133*	*143*
Noviembre	*91*	*106*	*117*	*126*	*133*	*144*
Diciembre	*91*	*106*	*117*	*128*	*134*	*146*

4.3. Estructuración de los modelos

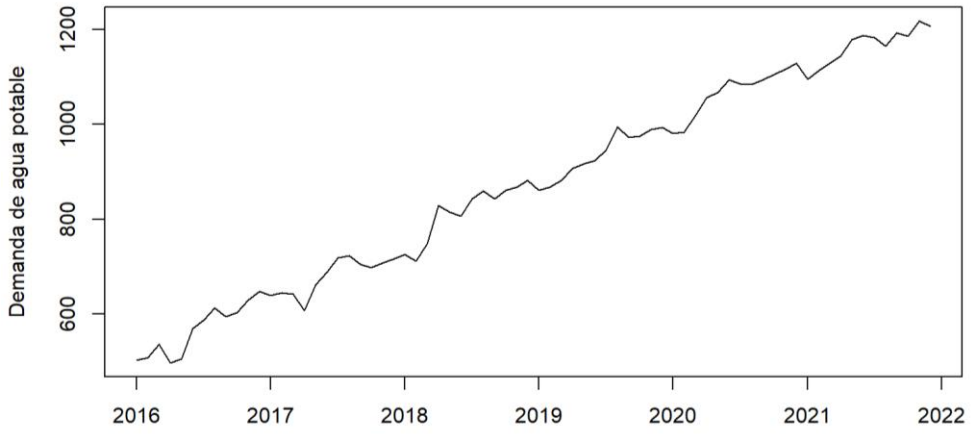
4.3.1. Análisis la serie temporal

La serie de tiempo se utiliza para analizar el comportamiento de las variables en el transcurso del tiempo, como se ha indicado con anterioridad la serie de tiempo corresponde a la demanda de agua potable desde 2016 hasta el 2021, desagradados en meses y cuya unidad de medida son los metros cúbicos. En la figura 13 se observa el consumo de agua potable a través a los años.

En el anexo 1 se encuentra la codificación para el análisis de la base de datos histórica en el mismo software RStudio.

Figura 13

Análisis de la serie temporal de la demanda de agua potable

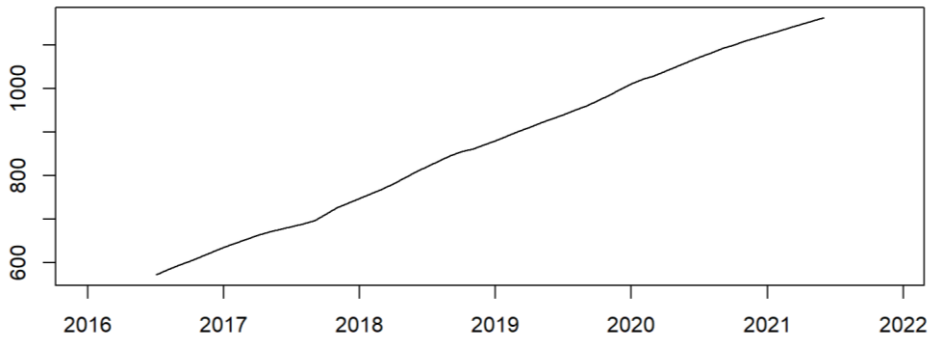


4.3.2. Análisis de tendencia y estacionalidad

La base de datos históricos provenientes del volumen de consumo de agua potable, que se ha incrementado en el transcurso de los años a causa del incremento de clientes y consumo de agua potable de estos. Se realiza un análisis de tendencia y se observa que la serie de tiempo tiene tendencia positiva desde el 2016 hasta el 2021. Ver figura 14.

Figura 14

Análisis de tendencia de la serie temporal de la demanda de agua potable

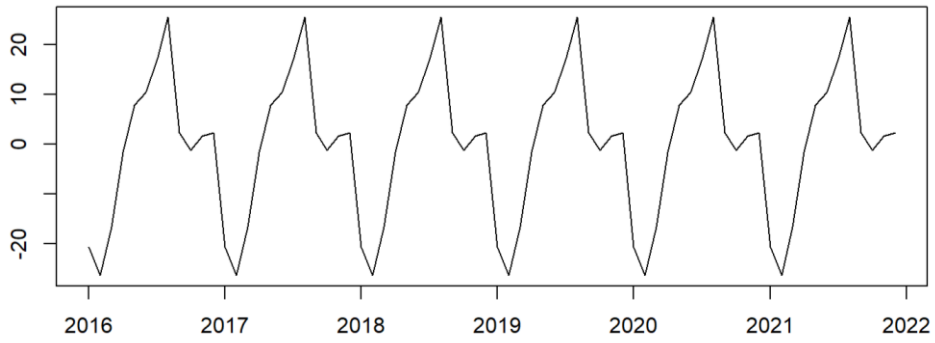


Utilizando la descomposición de la serie temporal se identifica como a lo largo de los seis años analizados, existen patrones en la demanda de agua potable, es decir el comportamiento es

similar en ciertos períodos del año. En el resultado es una estacionalidad significativa como se puede comprobar en la figura 15.

Figura 15

Análisis estacionalidad de la serie temporal de la demanda de agua potable



4.4. Modelo de Redes Neuronales Artificiales

4.4.1. Pronóstico de la red neuronal óptima

Se busca realizar el pronóstico de la demanda de agua potable, para ello se ingresa como base de datos la cantidad en metros cúbicos consumidos históricamente desde el 2016 hasta el 2021. Las variables de entrada son doce correspondientes a los meses del año.

La red neuronal óptima es con la que comienza el modelo, donde se determina las repeticiones que se va a realizar para reducir el MSE.

Se comienza con la primera iteración (fit1) y se identifica el pronóstico obtenido, la gráfica de la red MLP, la estructura de las paras y nodos; y los errores de pronóstico. Puede verse el código en la figura 16 y en la figura 17 el pronóstico obtenido para esta red MLP, y para mayor comprensión del código puede verse el anexo 2.

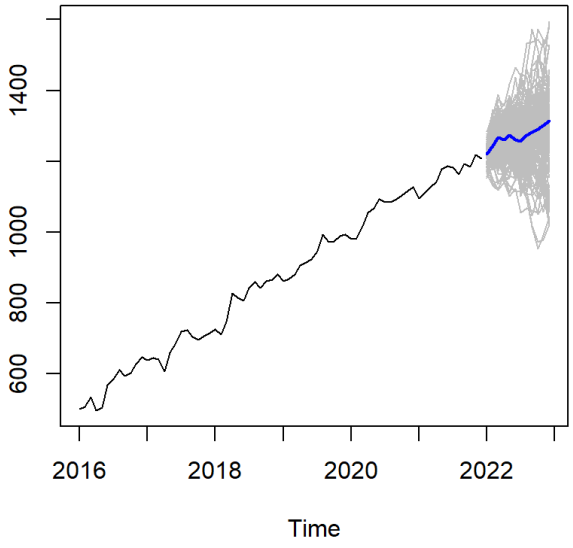
Figura 16

Codificación de la red neuronal óptima

```
#Red neuronal optima
Fit1<- mlp(y, reps = 200, lags =NULL,difforder = NULL,hd.max = NULL)
plot(Fit1)
forecast(Fit1)
print(Fit1)
plot(forecast(Fit1))
```

Figura 17

Pronóstico de la red MLP



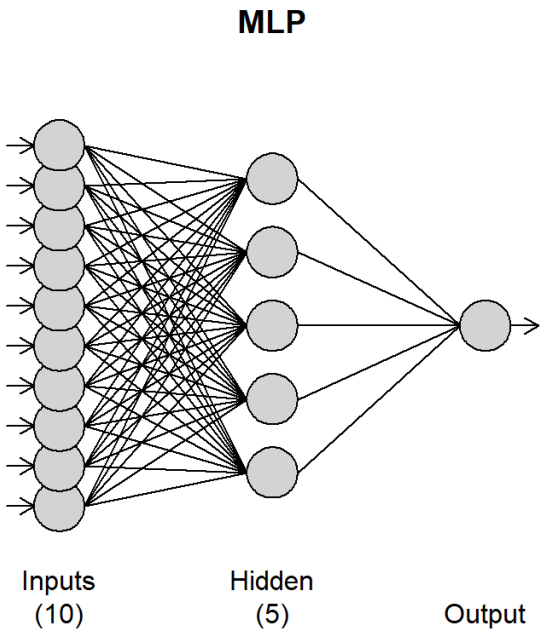
4.4.2. Estructura de la RNA

El perceptrón multicapa para el desarrollo y ejecución del modelo está conectado en su totalidad, está conformado por una capa de entrada, una capa oculta y una capa de salida. Compuestas por doce nodos circulares, 5 nodos y un solo nodo respectivamente como se muestra en la figura 18.

Cabe recalcar que los nodos circulares corresponden a una neurona artificial y las líneas son cada conexión que vincula la salida de una neurona con la entrada de otra, de donde se genera el pronóstico de la demanda de agua potable con la información procesada.

Figura 18

Estructura de las capas de la red



4.4.3. Pronóstico mejorando el entrenamiento de la red

Para lograr disminuir el error obtenido durante el entrenamiento inicial y la primera red ejecutada, se entrena la red nuevamente. Se codifica un Fit2 correspondiente a las mejoras partiendo de los resultados obtenidos Fit1. Donde se obtiene el menor error posible con un MSE de 0.3025. Para mayor comprensión del mejoramiento de la red se presenta la codificación en la figura 19 y el gráfico del pronóstico en la figura 20.

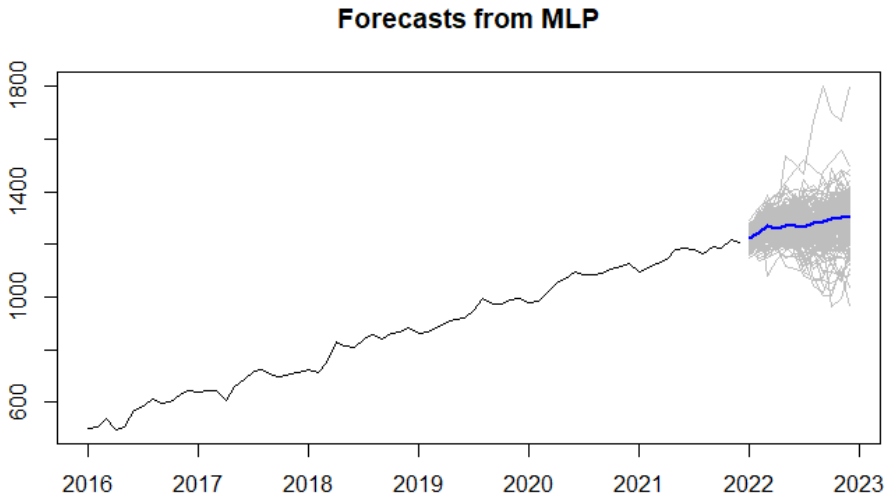
Figura 19

Codificación de la mejora de entrenamiento de la red neuronal

```
##Mejora de entrenamiento
Fit2<-mlp(y, model=Fit1,retrain=TRUE)
print(Fit2)
plot(Fit2)
plot(forecast(Fit2,h=h))
summary(forecast(Fit2,h=h))
```

Figura 20

Pronóstico del modelo de Redes Neuronales Artificiales



4.5. Modelo K Nearest Neighbor Regression

Utilizando las funciones en el software se realiza una primera predicción para el aprendizaje automático, ver figura 21 para comprender el código utilizado, donde se obtiene un pronóstico que se puede observar en la figura 22.

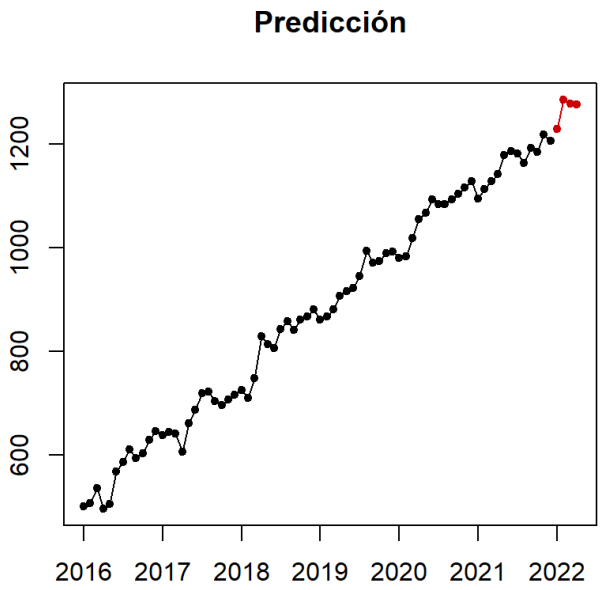
Figura 21

Codificación del aprendizaje automático

```
# Time Series Forecasting Using Nearest Neighbors
pred <- knn_forecasting(y, h = 12, lags = NULL, k = 2)
autoplot(pred)
autoplot(pred, highlight = "neighbors")
```

Figura 22

Predicción del aprendizaje automático



Respecto a la parametrización de las funciones para generar los parámetros del método de k-NN, se codifica una predicción basándose en la primera predicción obtenida como entrenamiento, para el cálculo de la predicción y los errores de pronóstico.

Si se desea comprender cada fase de la configuración del algoritmo a lo largo del código en R, se puede analizar el anexo 3. Y para analizar la última fase código y la gráfica del pronóstico utilizando este método véase la figura 23 y figura 24 respectivamente.

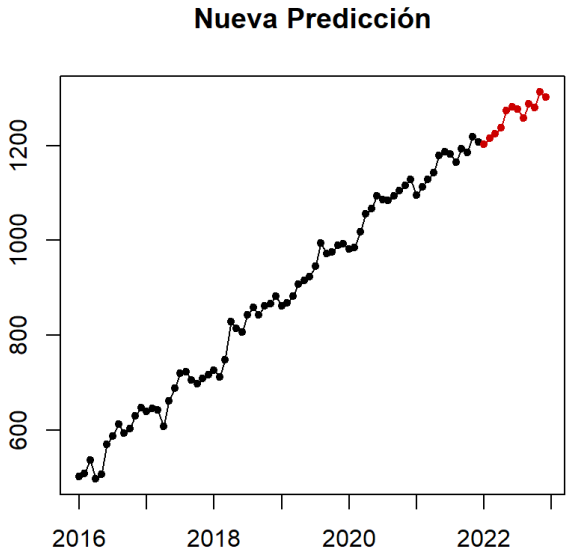
Figura 23

Codificación de la mejora de entrenamiento del modelo

```
pred <- knn_forecasting(y, h = 4, k = 1, msas = "recursive")
new_pred <- predict(pred, h = 12)
print(new_pred$prediction)
plot(new_pred) # To see a plot with the forecast
```

Figura 24

Pronóstico de la demanda utilizando el modelo K Nearest Neighbor Regression



4.6. Modelo Support Vector Machine

Utilizando el modelo SVM primero se define el entrenamiento automático inicial. Y luego con base en esos resultados se trabaja en el mejoramiento de la predicción. Para ello se utilizan librerías específicas en el software que ya tienen predefinidas las etapas de aplicación del modelo, en la figura 25 se puede visualizar y analizar el código utilizado, de donde se obtiene un pronóstico que se puede observar en la figura 26. Si se desea comprenden o analizar la totalidad del modelo la codificación se encuentra en el anexo 4.

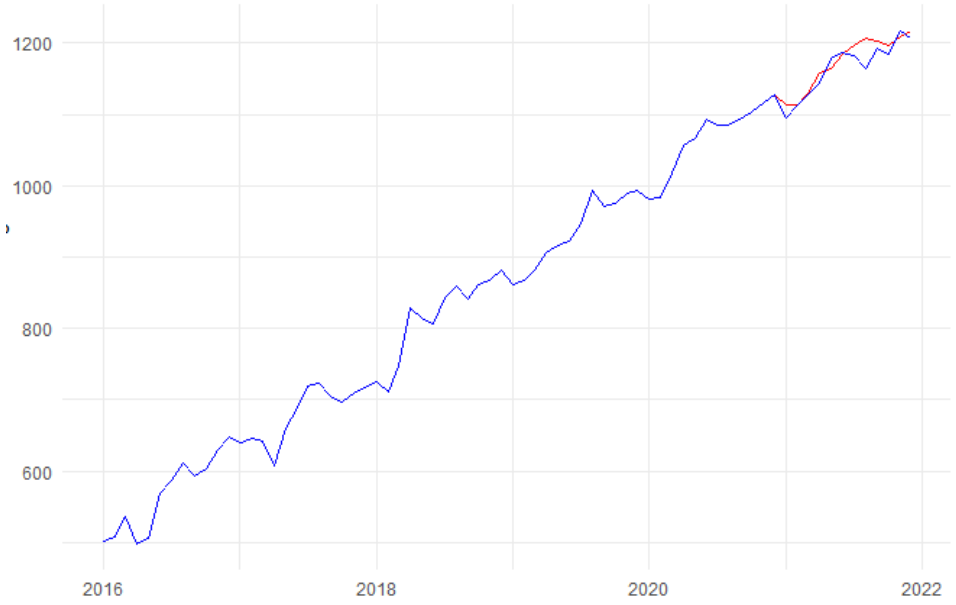
Figura 25

Codificación del entrenamiento y mejora del modelo SVM

```
# train an svm model, consider further tuning parameters for lower MSE

training_data <- window(Datosts, end = c(2020, 12))
testing_data <- window(Datosts, start = c(2021, 1))

svmodel <- ARml(training_data, maxlag = 5, caret_method = "svmLinear",
     lambda = 0.9,cost=11000000)

data <- caretForecast::forecast(svmodel, h = 12, level = NULL)-> fc

accuracy(fc, testing_data)

date <-seq(as.Date("2016/1/1"), as.Date("2021/12/1"), "months")
data
df2021 <- fc$mean
df2016 <- fc[[1]]
c(unclass(df2016),unclass(df2021))
```

Figura 26

Pronóstico de la demanda utilizando el modelo SVM



4.7. Modelo Random Forest

Para este modelo de pronóstico se sigue una serie de pasos que ayudándose de la codificación en R se va ejecutando progresivamente. Primero se obtienen las diferenciales, se ordenan los datos en forma de matriz, se realiza un primer entrenamiento automático (ver figura 27) que es la base con la se trabaja a lo largo del desarrollo del modelo. Se calcula una a una las

predicciones mensuales y finalmente se pronostica toda la demanda de agua potable para el año 2022 (ver figura 28). Para analizar cada fase o la totalidad del modelo se puede visitar el anexo 5.

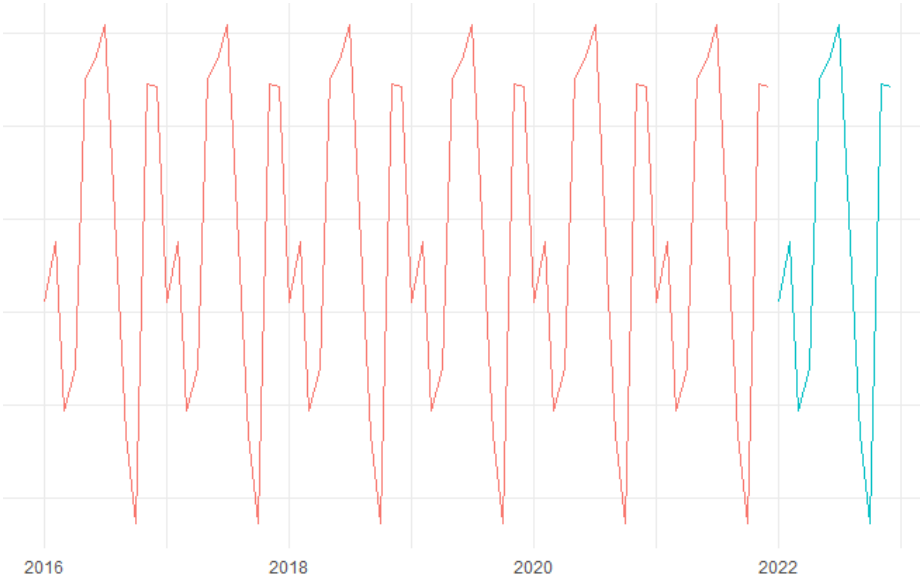
Figura 27

Codificación del entrenamiento inicial del modelo Random Forest

```
# Se obtiene le test y train
y_train <- tax_ts_mbd[, 1] # the target
X_train <- tax_ts_mbd[, -1] # everything but the target
#La parte test
y_test <- window(Datosts, start = c(2020, 1),end=c(2021,12))
X_test <- tax_ts_mbd[nrow(tax_ts_mbd), c(1:2)] # the test set consisting
```

Figura 28

Pronóstico de la demanda utilizando el modelo Random Forest



4.8. Resultados de los modelos de pronóstico

Una vez que se han ejecutado los cuatro modelos de inteligencia artificial se elabora la tabla 4 con los resultados del pronóstico para el año 2022 de cada uno de los modelos.

Donde se puede comprobar que existe diferencia entre todos y cada uno de los resultados obtenidos para cada mes del año.

Tabla 4

Pronóstico de la demanda $(\boldsymbol{m^3})$ *para el año 2022 de los diferentes modelos.*

Meses	***Redes Neuronales Artificiales***	***K Nearest Neighbor Regression***	***Support Vector Machine***	***Random Forest***
Enero	*1221*	*1202*	*1197*	*1197*
Febrero	*1244*	*1215*	*1201*	*1199*
Marzo	*1267*	*1223*	*1216*	*1194*
Abril	*1261*	*1237*	*1242*	*1196*
Mayo	*1274*	*1274*	*1249*	*1203*
Junio	*1261*	*1280*	*1267*	*1204*
Julio	*1276*	*1276*	*1274*	*1205*
Agosto	*1274*	*1257*	*1275*	*1199*
Septiembre	*1282*	*1287*	*1274*	*1194*
Octubre	*1290*	*1279*	*1275*	*1191*
Noviembre	*1302*	*1313*	*1275*	*1203*
Diciembre	*1315*	*1302*	*1288*	*1203*

4.8.1. Comparación de modelos

Se comparan los errores más relevantes de los modelos de pronóstico aplicados específicamente la raíz del error cuadrático medio (*RMSE*), error absoluto medio (*MAE*) y error porcentual absoluto medio (*MAPE*). Y en dependencia del menor error se selecciona el modelo idóneo para aplicarlo en esta serie de tiempo.

El RMSE es una medida útil para identificar que tanto se ajusta la base de batos histórica con el modelo. El valor del RMSE es directamente proporcional con la diferencia entre los valores predichos y los datos observados. Mientras más valor calculado de RMSE se obtiene, se determina que peor se ajusta el modelo de regresión con los datos.

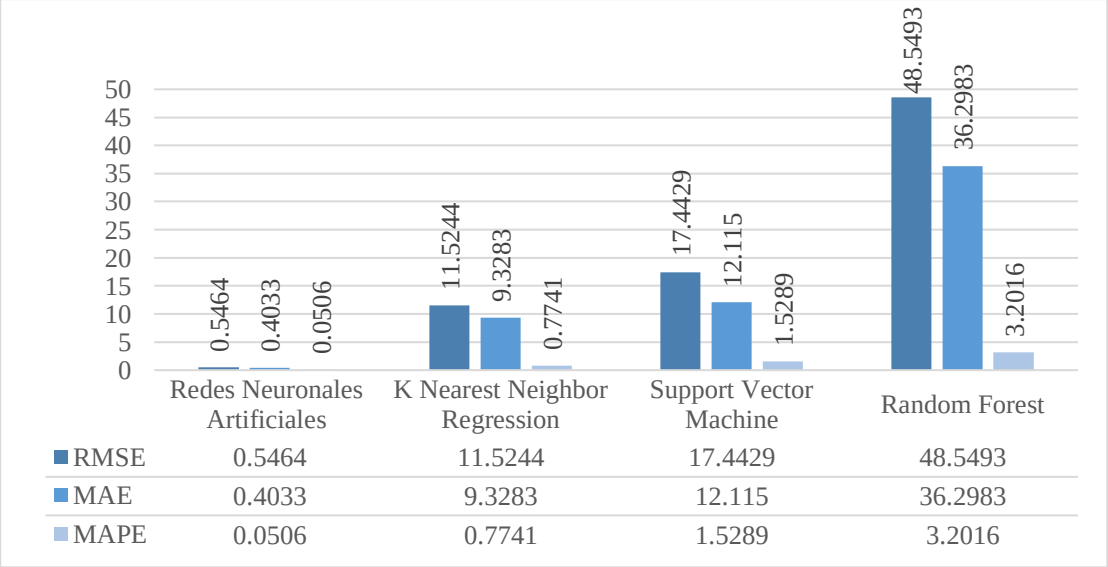
El MAE corresponde a la media de la diferencia absoluta entre los datos de la base histórica y los valores del pronóstico. Esta medida de error es más robusta que RMSE ya que no les da tanta importancia a los valores atípicos. Cuando se obtiene un valor más pequeño de MAE esto indica un mejor ajuste del modelo al comportamiento de los datos.

Por su parte el MAPE indica el desempeño del modelo permitiendo medir la precisión del pronóstico de cada uno. Semejante a la interpretación de los resultados obtenidos para las medidas de error previamente mencionadas en este caso también entre menor sea el valor mayor exactitud tendrá la predicción.

Como se observa en la figura 29 de acuerdo con los valores de error obtenidos el modelo de Redes Neuronales Artificiales es el que mejor se ajusta al comportamiento de esta serie de tiempo, por lo que se obtiene un menor error de pronóstico comparado con los otros modelos.

Figura 29

Comparación de los valores obtenidos como error de pronóstico de los modelos aplicados



4.8.2. Resultados del pronóstico para los años 2022, 2023 y 2024

Una vez que se ha determinado a través de la comparación de las medidas de error que el modelo de Redes Neuronales Artificiales es el mejor para trabajar con esta serie de tiempo se elabora la tabla 5 y la figura 30 con los resultados del pronóstico para el año 2022, 2023 y 2024 haciendo una nueva corrida del código para obtener los de estos últimos años.

Figura 30

Pronóstico de la demanda $(\boldsymbol{m^3})$ *para el año 2022, 2023 y 2024.*

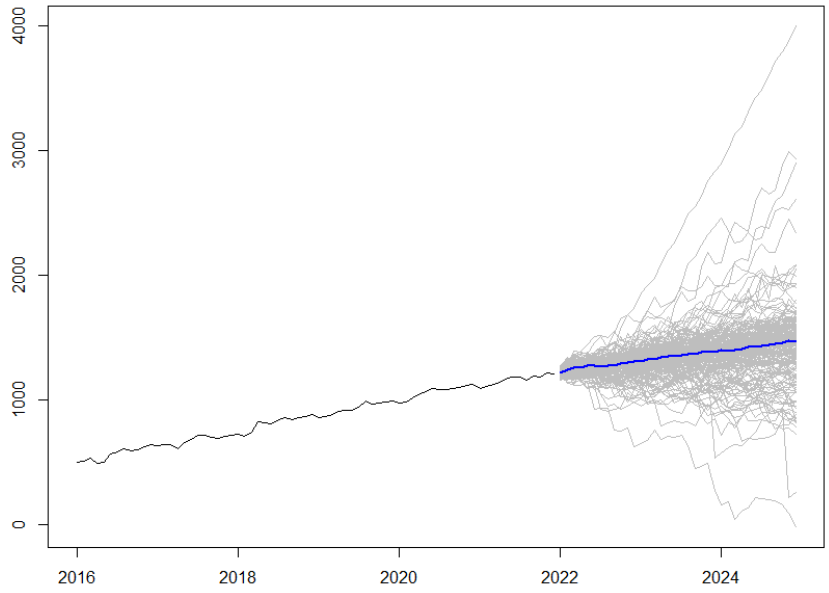


Tabla 5

Pronóstico de la demanda $(\boldsymbol{m^3})$ *para el año 2022, 2023 y 2024.*

Meses	***2022***	***2023***	***2024***
Enero	*1221*	*1312*	*1386*
Febrero	*1244*	*1318*	*1395*
Marzo	*1267*	*1336*	*1403*
Abril	*1261*	*1337*	*1415*
Mayo	*1274*	*1347*	*1422*
Junio	*1261*	*1354*	*1428*
Julio	*1276*	*1358*	*1442*
Agosto	*1274*	*1360*	*1440*
Septiembre	*1282*	*1361*	*1446*
Octubre	*1290*	*1365*	*1447*
Noviembre	*1302*	*1371*	*1456*
Diciembre	*1315*	*1372*	*1464*

Analizando los resultados obtenidos se considera que la demanda de agua potable en función del tiempo se ha incrementado y continuará con ese, es decir la tendencia de consumo es creciente. La organización como medida para continuar brindando el servicio y a su vez lograr

mantener el equilibrio entre oferta y demanda correspondientes al volumen de suministro que la empresa produce y el volumen que consumen por los usuarios, debe considerar alternativas para aumentar su capacidad de captación, potabilización y reserva de agua.

Se pueden considerar diversas opciones entre las que se puede mencionar a continuación. La construcción de un depósito con mayor capacidad de almacenamiento con el fin de incrementar la reserva de agua previa a la distribución. La identificación y uso de fuentes externas de agua para que alimenten la red existente. Y/o la construcción de una nueva red de agua potable diseñada para abastecer del líquido vital a las zonas que se vayan poblando debido al crecimiento demográfico.

CONCLUSIONES

Realizando el marco teórico referencial se determinó los modelos de inteligencia artificial para obtener el pronóstico de agua potable definiendo las técnicas y algoritmos empleados por los modelos en la codificación, predicción y análisis de los resultados. Logrando de esta manera resaltar a importancia y precisión de cada uno de estos.

En el análisis de la situación actual de la organización se obtuvieron los datos de consumo mensual que corresponde a la base de datos sobre la que se trabaja, la cantidad de clientes y los datos de la población del sector. Se realizó un análisis estadístico de la base de datos validando y comprobando que los datos históricos cumplen con los requisitos de una serie de tiempo.

Se emplearon cuatro modelos de inteligencia artificial; Redes Neuronales Artificiales, K Nearest Neighbor Regression, Random Forest y Support Vector Machine; para elaborar la previsión de la demanda de agua potable, utilizando el software RStudio se codificó los modelos para obtener la mejor solución en cada caso y los errores de pronóstico de cada modelo.

Una vez se obtuvo los resultados de cada modelo se comparó los valores obtenidos de los diferentes errores de pronóstico y se seleccionó como la mejor opción para aplicar en esta serie temporal el modelo de Redes Neuronales Artificiales que presenta menor error en la predicción de la demanda como se detalla en la figura 29. Obteniendo un RMSE de 0.5464, MAE de 0.4033 y MAPE de 0.0506.

Finalmente se realizó la adaptación del código para modelo de Redes Neuronales Artificiales ya que en un inicio se definió pronosticar un año según se muestra en el código la variable h que es la cantidad de años a pronosticar. Que para este caso se establece pronosticar 36 meses correspondientes a los años 2022, 2023 y 2024. Se ejecuta la corrida del software con el

objetivo de obtener el pronóstico de la demanda correspondiente a los años 2023 y 2024 que son los que faltaban por incluir en los resultados.

Teniendo en cuenta los resultados obtenidos se determina que la demanda se va incrementando con el pasar del tiempo, por ende, la tendencia de consumo es creciente. Motivo por lo que la organización debe considerar alternativas para incrementar la capacidad de captación, potabilización y reserva de agua, con la finalidad de continuar satisfaciendo la demanda del recurso vital. Entre las opciones se pueden considerar la construcción de una nueva unidad de reserva con mayor capacidad, buscar nuevas fuentes de agua bruta que se integren a la red existente de agua potable y/o construir una nueva red que esté destinada a abastecer de agua potable a los sectores que aún no tienen asentamientos de moradores.

RECOMENDACIONES

Con base en el modelo de pronóstico de la demanda la empresa o un consultor pueda elaborar un plan maestro de producción donde se contemplan las necesidades de la organización para satisfacer la demanda además que permita llevar un mejor control de la producción de agua potable reduciendo los riesgos de pérdida.

Para un próximo estudio sería pertinente que se tomen para la base de datos todas las plantas de captación y distribución del cantón que son manejadas por las diferentes juntas administradoras de agua potable.

Sería apropiado considerar la evaluación de la base de datos como una serie tipo panel considerando el incremento de clientes para el pronóstico de la demanda de agua potable y futura planificación de la producción y distribución del recurso hídrico.

REFERENCIAS

Adamowski, J., & Chan, H. F. (2011). A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology*, *407*(1–4), 28–40. https://doi.org/10.1016/J.JHYDROL.2011.06.013

Aha, D., Kibler, D., & Albert, M. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–66.

Amat Rodrigo, J. (2017). *Árboles de decisión, random forest, gradient boosting y C5.0*.

Arroyo Gallardo, J. (2008). *Métodos de Predicción para Series Temporales de Intervalos e Histogramas*. Universidad Pontificia Comillas.

Babel, M., das Gupta, A., & Pradhan, P. (2006). A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal. *Water Resources Management*, *21*, 573–589.

Bonilla Martínez, E. G. (2005). *Reconocimiento de caracteres mediante redes neuronales con MATLAB*. Escuela Politécnica Nacional.

Bravo Sanzana, M., Salvo Garrido, S., & Muñoz Poblete, C. (2015). Profiles of Chilean students according to academic performance in mathematics: An exploratory study using classification trees and random forests. *Studies in Educational Evaluation*, *44*, 50–59. https://doi.org/10.1016/J.STUEDUC.2015.01.002

Colina, E., & Rivas, F. (1998). Introducción a la inteligencia artificial. *Cuadernos de Control. Postgrado En Ingeniería de Control. Universidad de Los Andes*.

Comisión Nacional del Agua. (2003). Diseño de Redes de Distribución de Agua Potable. In *Manual de agua potable, alcantarillado y saneamiento.*

Consejo Nacional de Planificación. (2017). *Plan Nacional de Desarrollo 2017-2021 Toda una Vida*.

Constitución de la República del Ecuador. (2008). In *Registro Oficial* (Vol. 449, Issue 20). www.lexis.com.ec

Cuevas Alfaro, E. A. (2010). *MÁQUINAS DE SOPORTE VECTORIAL CON ALGORITMOS BASADOS EN POBLACIONES PARA EL PRONÓSTICO DEL PRECIO DE ACCIONES LAN CHILE*. PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO.

Cuevas Soto, V. M., Alvares Iriarte, S., Azcona Romero, M., & Rodríguez Rogert, I. (2019). Predictive power of the Support Vector Machine. An application to the financial planning. *Revista Cubana de Ciencias Informáticas*, *13*(3), 59–75. http://rcci.uci.cu

Daza Sánchez, Francisca. (2008). *Demanda de agua en zonas urbanas en Andalucía* [Tesis Doctoral]. Universidad de Córdoba.

Farías Concha, M. N. (2011). *MÁQUINAS VECTORIALES HÍBRIDAS PARA CLASIFICAR ACCIDENTES DE TRÁNSITO EN LA REGIÓN METROPOLITANA*. Pontificia Universidad Católica de Valparaíso.

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. *Progress in Artificial Intelligence*, *9273*, 535–546.

Fontanazza, C. M., Notaro, V., Puleo, V., & Freni, G. (2014). Multivariate Statistical Analysis for Water Demand Modeling. *Procedia Engineering*, *89*, 901–908. https://doi.org/10.1016/J.PROENG.2014.11.523

Gala García, Y. (2013). *Algoritmos SVM para problemas sobre big data*. Universidad Autónoma de Madrid.

García, E. A., & Osella Massa, G. L. (2003). *Evolución de Redes Neuronales mediante Sistemas de Reescritura* [Tesis Pregrado]. Universidad Nacional de La Plata.

Glocker, B., Konukoglu, E., & Haynor, D. R. (2016). Random Forests for Localization of Spinal Anatomy. *Medical Image Recognition, Segmentation and Parsing*, 93–110. https://doi.org/10.1016/B978-0-12-802581-9.00005-6

González, H., Santos, G., Campos, F., & Morell Pérez, C. (2016). Evaluación del algoritmo KNN-SP para problemas de predicción con salidas compuestas. *Revista Cubana de Ciencias Informáticas*, *10*(3).

González, R., Barrientos, A., Toapanta, M., & del Cerro, J. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, *14*(4), 394–405. https://doi.org/10.1016/J.RIAI.2017.07.005

Habadi, M., & Tsokos, C. (2017). Statistical Forecasting Models of Atmospheric Carbon Dioxide and Temperature in the Middle East. *Journal of Geoscience and Environment Protection*, *5*(10).

Hanke, J. E., & Wichern, D. W. (2010). *Pronósticos en los negocios* (9th ed.). Pearson Educación.

Herrera, M., Torgo, L., Izquierdo, J., & Pérez-García, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, *387*(1–2), 141–150. https://doi.org/10.1016/J.JHYDROL.2010.04.005

Hilera Gonzáles, J. R., & Martínez Hernando, V. J. (1995). *Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones* (1st ed.).

Hyndman, R. J. (2014). *Forecasting: Principles & Practice*.

INEC. (2022). *FASCÍCULO PROVINCIAL CARCHI*.

Ing, E., Su, W., Schonlau, M., & Torun, N. (2019). Support Vector Machines and logistic regression to predict temporal artery biopsy outcomes. *Canadian Journal of Ophthalmology*, *54*(1), 116–118. https://doi.org/10.1016/J.JCJO.2018.05.006

Isasi Viñuela, P., & Galván León, I. (2004). *Redes Neuronales Artificiales: Un enfoque práctico*.

Jaramillo, J. (2015). *Tendencias Recientes en el Pronóstico de Series de Tempo Financieras usando Máquinas de Vectores de Soporte*. Universidad Nacional de Colombia.

Jere, S., Kasense, B., & Chilyabanyama, O. (2017). Forecasting Foreign Direct Investment to Zambia: A Time Series Analysis. *Open Journal of Statistics*, *4*(8).

Jiménez Terán, J. M. (2013). *MANUAL PARA EL DISEÑO DE SISTEMAS DE AGUA POTABLE Y ALCANTARILLADO SANITARIO*.

Reglamento Interno de la Junta Administradora de Agua Potable de "La Portada," (2010).

Layme, M. (2020). *Manual de Operación y Mantenimiento*.

Lema, M. F. (2006). *Diseño del Sistema de Agua Potable a Bombeo para la Comunidad de Cochaloma del Cantón Colta de la Provincia de Chimborazo*.

Ley Orgánica de Recursos Hídricos, Usos y Aprovechamiento del Agua. (2014). www.lexis.com.ec

Lindner, C. (2017). Automated Image Interpretation Using Statistical Shape Models. *Statistical Shape and Deformation Analysis: Methods, Implementation and Applications*, 3–32. https://doi.org/10.1016/B978-0-12-810493-4.00002-X

Martí Pérez, P. C. (2009). *Aplicación de redes neuronales artificiales para predicción de variables en ingeniería del riego: evapotranspiración de referencia y pérdidas de carga localizadas en emisores integrados* [Tesis Doctoral]. Universitat Politècnica de València.

Matich, D. J. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universidad Tecnológica Nacional.

Mora Escobar, H. M. (2001). *OPTIMIZACION NO LINEAL Y DINAMICA* (2nd ed.).

Nievas Lio, E. (2016). *Aplicando máquinas de soporte vectorial al análisis de pérdidas no técnicas de energía eléctrica*. Universidad Nacional de Córdoba.

Patiño Pérez, D., Silva Bustillos, R., Munive Mora, C., & Botto Tobar, M. (2020). Predicción de Covid19 con el uso del Algoritmo Random Forest y Redes Neuronales Artificiales. *Ecuadorian Science Journal*, *4*(2), 101–110. https://doi.org/10.46480/esj.4.2.41

Rodríguez Aedo, N. C. (2016). *PRONÓSTICO DE DEMANDA DE AGUA POTABLE MEDIANTE REDES NEURONALES* [Tesis Pregrado, Universidad Técnica Federico Santa María]. https://repositorio.usm.cl/bitstream/handle/11673/13205/3560900232223UTFSM.pdf?sequence=1

Sahoo, G. B., Schladow, S. G., & Reuter, J. E. (2009). Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. *Journal of Hydrology*, *378*(3–4), 325–342. https://doi.org/10.1016/J.JHYDROL.2009.09.037

Sarmiento Maldonado, H. O., & Villa Acevedo, W. M. (2008). INTELIGENCIA ARTIFICIAL EN PRONOSTICO DE DEMANDA DE ENERGIA ELECTRICA: UNA APLICACION EN OPTIMIZACION DE RECURSOS ENERGETICOS. *Revista Colombiana de Tecnologías de Avanzada*, *2*(12), 94–100.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1). https://doi.org/10.1177/1536867X20909688

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms ofClassification in Data Mining. *International Journal of Science and Research*, *5*(4).

Silver, E., Pyke, D., & Thomas, D. (2017). *Inventory and Production Management in Supply Chains* (4th ed.). CRC Press, Taylor & Francis Group.

Sipper, D., & Bulfin, R. (1998). *Planeación y control de la produccción* (1st ed.). Mc Graw Hill.

Statnikov, A., Aliferis, C., Hardin, D., & Guyon, I. (2011). *A gentle introduction to support vector machines in biomedicine: Volume 1: Theory and methods*.

Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Sungwon, K., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S. I., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K., … Yaseen, Z. M. (2022). Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing*, *489*, 271–308. https://doi.org/10.1016/J.NEUCOM.2022.03.014

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Toro Ocampo, E. M., Mejía Giraldo, D. A., & Salazar Isaza, H. (2004). PRONÓSTICO DE VENTAS USANDO REDES NEURONALES. *Scientia Et Technica*, *X* (26), 25–30.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999.

Velásquez, J. D., Olaya, Y., & Franco, C. J. (2010). PREDICCIÓN DE SERIES TEMPORALES USANDO MÁQUINAS DE VECTORES DE SOPORTE TIME SERIES PREDICTION USING SUPPORT VECTOR MACHINES. *Revista Chilena de Ingeniería*, *18*(1), 64–75.

Villada, F., Muñoz, N., & García, E. (2012). Aplicación de las Redes Neuronales al Pronóstico de Precios en el Mercado de Valores. *Información Tecnológica*, *23*(4), 11–20.

Voß, S., & Woodruff, D. L. (2006). *Introduction to Computational Optimization Models for Production Planning in a Supply Chain* (2nd ed.). Springer.

Wettschereck, D., Aha, D., & Mohri, T. (1997). A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review, 11*(1–5), 273–314.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE, 87*(9), 1423–1447.

Yeh, I. C., & Lien, C. hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/J.ESWA.2007.12.020

Zavala Hepp, B. I. (2015). *Pronóstico de demanda desagregada para una empresa de productos de consumo masivo* [Tesis Pregrado, Universidad de Chile]. https://repositorio.uchile.cl/handle/2250/137650

Zhou, S. L., McMahon, T. A., Walton, A., & Lewis, J. (2002). Forecasting operational demand for an urban water supply zone. *Journal of Hydrology*, *259*(1–4), 189–202. https://doi.org/10.1016/S0022-1694(01)00582-0

ANEXOS

Anexo 1. Codificación de la serie temporal de la demanda de agua potable

```
Datos <- read.csv("Base Datos.csv")
Datos

#convertir la base de datos en serie temporal(ts)
Datosts=ts(Datos$Consumos,freq=12,start=c(2016,1))
Datosts
plot(Datosts)

#calculando la estacionalidad
ts_seasonal(Datosts,type="all")
#Estacionariedad prueba dickey-fuller
adf.test(Datosts)
```

Anexo 2. Codificación del Modelo de Redes Neuronales Artificial

```
#pronosticar la serie
y <- Datosts
y

#las variables de entrada son 12 meses
h <- 1*frequency(y)
frequency(y)

#Entrenamiento automÃ¡tico
fit1<-mlp(y,hd = c(22,26,12),sel.lag=FALSE, lag=1:12,difforder=c(3,10),reps = 20)
print(fit1)

Fit1<- mlp(y, reps = 200, lags =NULL,difforder = NULL,hd.max = NULL)
plot(Fit1)
forecast(Fit1)
print(Fit1)
plot(forecast(Fit1))


##Mejora de entrenamiento
Fit2<-mlp(y, model=Fit1,retrain=20)
print(Fit2)
plot(Fit2)
plot(forecast(Fit2,h=h))
summary(forecast(Fit2,h=h))
```

Anexo 3. Codificación del Modelo K Nearest Neighbor Regression

```
#pronosticar la serie
y <- Datosts
y

#las variables de entrada son 12 meses
h <- 1*frequency(y)
frequency(y)


# Time Series Forecasting Using Nearest Neighbors
pred <- knn_forecasting(y, h = 12, lags = NULL, k = 2)
autoplot(pred)
autoplot(pred, highlight = "neighbors")

pred <- knn_forecasting(y, h = 1, lags = NULL, k = 2)
knn_examples(pred)

pred <- knn_forecasting(y, h = 12, lags = NULL, k = 2)
pred$prediction # To see a time series with the forecasts
plot(pred) # To see a plot with the forecast

pred <- knn_forecasting(y, h = 4, lags = NULL, k = 2, msas = "MIMO")
nearest_neighbors(pred)

pred <- knn_forecasting(y, h = 4, k = 1, msas = "recursive")
new_pred <- predict(pred, h = 12)
print(new_pred$prediction)
plot(new_pred) # To see a plot with the forecast

pred <- knn_forecasting(y, h = 4, lags = NULL, k = 2)
ro <- rolling_origin(pred)
print(ro$global_accu)
```

Anexo 4. Codificación del Modelo Support Vector Machine

```
# train an svm model, consider further tuning parameters for lower MSE
training_data <- window(Datosts, end = c(2020, 12))
testing_data <- window(Datosts, start = c(2021, 1))

svmodel <- ARml(training_data, maxlag = 5, caret_method = "svmLinear",
     lambda = 0.9,cost=11000000)

data <- caretForecast::forecast(svmodel, h = 12, level = NULL)-> fc

accuracy(fc, testing_data)

date <-seq(as.Date("2016/1/1"), as.Date("2021/12/1"), "months")
data
df2021 <- fc$mean
df2016 <- fc[[1]]
c(unclass(df2016),unclass(df2021))
# visualize the forecasts
FINAL_df <- data.frame(date=date,value=Datosts)

#Grafico
FINAL_df$forcast <-c(unclass(df2016),unclass(df2021))

plot_fc <- FINAL_df %>%
  ggplot(aes(x = date)) +

  geom_line(aes(y = forcast),color="red") +
  geom_line(aes(y = value),color="blue") +
  theme_minimal() +
  labs(

    x = "Lína de tiempo",
    y = "Consumo de agua"
  )
plot_fc

get_var_imp(fc)
get_var_imp(fc, plot = F)

## predict
data_for_model<- window(Datosts, end = c(2021, 12))
svmodel <- ARml(data_for_model, maxlag = 5, caret_method = "svmLinear",
                lambda = 1)
data_predict <- caretForecast::forecast(svmodel, h = 12, level = NULL)-> fc
fc
```

Anexo 5. Codificación del Modelo Random Forest

```
# Obtengo las diferencias
tax_ts_org <- window(Datosts, end = c(2021, 12))
n_diffs <- nsdiffs(tax_ts_org)

# Se ordena en forma de matriz con x y y
tax_ts_trf <- tax_ts_org
tax_ts_mbd <- embed(tax_ts_trf,2)
tax_ts_mbd
# Se obtiene le test y train
y_train <- tax_ts_mbd[, 1] # the target
X_train <- tax_ts_mbd[, -1] # everything but the target
#La parte test
y_test <- window(Datosts, start = c(2020, 1),end=c(2021,12))
X_test <- tax_ts_mbd[nrow(tax_ts_mbd), c(1:2)] # the test set consisting
# of the six most recent values (we have six lags) of the training set. It's the
# same for all models.

# Creo mi dataframe para pronosticar mis datos
X_test <- data.frame(y_train=tax_ts_mbd[nrow(tax_ts_mbd),1 ],X_train=tax_ts_mbd[nrow(tax_ts_mbd),2 ])
horizon <- 12

#Hago mi cáldulo de modelo y forcast para cada mes
forecasts_rf <- numeric(horizon)
for (i in 1:horizon){
  # set seed
  set.seed(2019)
  # fit the model
  fit_rf <- randomForest(y_train~ X_train )
  # predict using the test set
  forecasts_rf[i] <- predict( fit_rf, newdata = X_test)
  # here is where we repeatedly reshape the training data to reflect the time distance
  # corresponding to the current forecast horizon.
  y_train <- y_train[-1]
  X_train <- X_train[-length(X_train) ]
}


#Obtengo mi predición
y_pred <- ts(
  forecasts_rf,
  start = c(2022, 1),
  frequency = 12
)
y_pred
```