ANÁLISIS COMPARATIVO DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO, APLICADAS A DATOS HIDROLÓGICOS DE LA CIUDAD DE CARTAGENA DESDE 1981 HASTA 2016

YESENIA CASTILLO ROMAN
LIZ DE AVILA BELLIDO

UNIVERSIDAD DEL SINÚ
ESCUELA DE INGENIERÍA DE SISTEMAS
CARTAGENA, COLOMBIA
AÑO 2018

ANÁLISIS COMPARATIVO DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO, APLICADAS A DATOS HIDROLÓGICOS DE LA CIUDAD DE CARTAGENA DESDE 1981 HASTA 2016

Proyecto de grado para optar el título de ingeniero de sistemas

YESENIA CASTILLO ROMAN
LIZ DE AVILA BELLIDO

Asesor Disciplinar
RAFAEL MONTERROZA
Asesor Metodológico
EUGENIA ARRIETA RODRIGUEZ

UNIVERSIDAD DEL SINÚ
ESCUELA DE INGENIERÍA DE SISTEMAS
CARTAGENA, COLOMBIA
AÑO 2018

Resumen

Con el fin de realizar la aplicación de modelos de inteligencia artificial, específicamente en técnicas de aprendizaje automático supervisado Maquinas de soporte vectorial, Redes neuronales y Regresión logística, se evalúan datos hidrológicos registradas por la estación Hidrológica 29037610 Kilómetro 107 de corriente canal del dique, cuya información fue suministrada por el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia (IDEAM) en las fechas del 10/04/1981 hasta el 31/08/2016; se concluye que la técnica que la técnica Maquinas de soporte vectorial, obtuvo un mejor desempeño a nivel de Exactitud con 0.547, Precisión de 0.567 y obteniendo en Recall o Sensibilidad 0.830 presentando una leve diferencia de puntuación de 0.012 con la técnica Regresión Logística la cual obtuvo mejores resultados.

Abstract

In order to carry out the application of artificial intelligence models, specifically in supervised automatic learning techniques Vector support machines, Neural Networks and Logistic Regression, hydrological data recorded by the hydrological station 29037610 Kilometer 107 of the dike channel current are evaluated. Information was provided by the Institute of Hydrology, Meteorology and Environmental Studies of Colombia (IDEAM) on the dates of 04/10/1981 until 08/31/2016; It is concluded that the technique that the Vector Support Machines technique, obtained a better performance at the level of Accuracy with 0.547, Precision of 0.567 and obtaining in Recall or Sensitivity 0.830 presenting a slight difference of score of 0.012 with the Logistic Regression technique which obtained best results.

Contenido

Lista de Figuras

1. Introducción

Este proyecto consiste en el análisis comparativo de técnicas de aprendizaje automático supervisado, en la cual se plantea una explotación y análisis de datos por medio de bases de datos otorgadas por el IDEAM (Instituto de hidrología, meteorología y estudios ambientales), esta información será útil para la identificación de patrones y predicción de comportamientos hidrológicos, junto con el análisis descriptivos entre las técnicas de aprendizaje automático supervisado en el comportamiento de la hidrología en la ciudad de Cartagena desde 1981 hasta 2016.

Colombia es un país con muchos recursos hídricos, los cuales están representados en aguas oceánicas, depositadas o estancadas, aguas de escurrimiento y aguas subterráneas; constituida por el mar Caribe y el océano pacifico que bañan el territorio continental por el norte y el occidente respectivamente. El abundante recurso hídrico con el que cuenta el país impone un valor agregado al momento de darle importancia a la hidrología; por lo anterior se tomó como muestra de estudio la ciudad de Cartagena la cual es rica en recursos hidrológicos y que en el trascurso de los años ha manifestado inconvenientes y problemáticas con respecto a inundaciones; considerando lo anterior se toma como muestra los datos correspondientes a la ciudad de Cartagena los cuales fueron otorgados por el IDEAM,desde el año 1981 hasta el año 2016.

Como proyecto investigativo se busca explorar y analizar datos hidrológicos en la ciudad en el periodo de tiempo comprendido entre 1981 hasta 2016 ya que hasta la fecha no se evidencian estudios o casos en los cuales se apliquen o estudien técnicas de inteligencia artificial en los datos suministrados por el ente control (IDEAM) en la región Caribe; se ha definido el estudio de los datos anteriormente indicados, mediante la utilización de la herramienta Microsoft Azure Machine Learning Studio la cual posee una gama amplia de técnicas de aprendizaje automático supervisado que permitirán la realización de un análisis comparativo del comportamiento de las técnicas aplicadas.

Actualmente se cuenta con diferentes técnicas para la predicción o identificación de patrones en diferentes campos, desde los modelos estadísticos hasta modelos avanzados con algoritmos computacionales que se basan en inteligencia artificial, como por ejemplo regresión logística redes neuronales, Kohonen y Grossberg entre otras.

Este proyecto no busca la construcción de un modelo de aprendizaje supervisado, si no la evaluación de algoritmos y técnicas para obtener un rendimiento o resultado para problemas relacionados con hidrología como son las precipitaciones en la ciudad.

2. Diseño metodológico

2.1. Planteamiento del problema

En esta sección se define el problema mediante la descripción, justificación y formulación,se aclara la problemática a tratar en el proyecto.

2.1.1. Descripción del Problema

El aprendizaje automático supervisado propone traer cada vez más inteligencia a todos los software de las máquinas y dispositivos en la actualidad, desde un teléfono inteligente a una máquina de café o un dispositivo para el hogar. Por tal razón dicho aprendizaje se manifiesta como una tecnología y sobre todo como una estrategia de modelado matemático el cual busca ayudar a comprender y/o abarcar el contenido de una base de datos.

Teniendo en cuenta la problemática de la ciudad de Cartagena en los últimos años, los cuales se evidencias en las noticias locales como el Diario el Universal el cual el día 20 de noviembre de 2016 que titula "Cartagena amanece bajo lluvia, bayunca y el pozón se inundaron - Las lluvias no cesan en Cartagena y los estragos causados por la falta de un eficiente sistema de drenaje fluvial mantiene padeciendo a los habitantes de diferentes sectores de la ciudad. Hoy amanecieron el corregimiento de Bayunca y el barrio El Pozón inundados tras el aguacero que baña a La Heroica desde la madrugada de este domingo-" al igual que el diario del día 7 de mayo de 2017 en el cual titula "fuertes vientos y lluvias a esta hora en Cartagena - A través de las redes sociales, usuarios comparten sus reacciones por la fuerte lluvia que cae en la mañana de este domingo en Cartagena. José Magallanes, Comandante del Cuerpo de Bomberos confirmó que hasta el momento no hay reporte de emergencias en los barrios de la ciudad, sin embargo, las unidades de monitoreo están atentas para atender cualquier situación-" y las noticias nacionales como el Diario el Heraldo del día 17 de noviembre de 2017 en el que se titula "inundaciones en Cartagena luego de tres horas de lluvias - Por más de tres horas un intenso aguacero se registró en Cartagena. El Cuerpo de Bomberos reportó que no se habían presentando emergencias y solo las inundaciones que son ya frecuentes cuando suele llover de manera torrencial-".

Por lo anterior en este proyecto se enfoca en el comportamiento de la hidrología de la ciudad de Cartagena desde 1981 hasta 2016.

2.1.2. Formulaciónón del Problema

¿Cuál es la técnica de inteligencia artificial que permite expresar claramente el comportamiento de los datos hidrológicos de la ciudad de Cartagena, tomando para el estudio los datos hidrológicos de la ciudad desde 1981 hasta 2016?

2.1.3. Justificación

Con el auge de los sistemas de información de la última década, cada día se cuenta con más y más información. Con este creciente volumen de información nace la necesidad de analizar estos datos para la toma de decisiones, es por esto que se hace inevitable, implementar técnicas especializadas como es el aprendizaje automático supervisado, que consiste en aprender una métrica determinada respondiendo específicamente a las características de los datos históricos. Es importante resaltar que estos algoritmos no requieren información de etiqueta de clases, y se han utilizado principalmente para mejorar los resultados de métodos de agrupamiento.
Un caso puntual es el IDEAM que se encarga de recopilar, procesar, interpretar y hacer públicos los datos hidrológicos, meteorológicos y geográficos sobre aspectos biofísicos, geomorfologicos , suelos y cobertura vegetal, para el manejo adecuado y aprovechamiento racional de los recursos biofísicos del país.
El monitoreo del agua en la integralidad del dominio del ciclo hidrológico, proporciona las herramientas conceptuales y metodológicas para evaluar el estado y la dinámica del agua, en cantidad y calidad. Refiere las variables determinadas por los procesos del ciclo hidrológico, como la precipitación, evapotranspiración, escorrentía (niveles y caudales), aguas subterráneas (niveles piezométricos, variables hidráulicas), sedimentos, así como también variables relacionadas con el estado y la dinámica de la calidad del agua, en sus manifestaciones fisicoquímicas e hidrobiológicas. El protocolo no incluye el seguimiento a variables asociadas con agua potable y saneamiento básico.
Este proyecto está enfocado al análisis comparativo de dos técnicas de aprendizaje automático supervisado, aplicando inteligencia artificial. El sistema deberá mostrar la comparación gráfica de las técnicas y a su vez permitirá hacer el análisis correspondiente.
El motivo del desarrollo de esta investigación es el gran impacto e importancia que representa el estudio de los algoritmos de aprendizaje automático supervisado aplicados al estudio de las precipitaciones, para así convertirlos en información útil, usando técnicas avanzadas; las cuales serán evaluadas para medir su grado de efectividad.

2.2. Objetivos

Realizar un análisis comparativo del rendimiento de las técnicas de aprendizaje automático supervisado para el comportamiento de los fenómenos hidrológicos de la ciudad de Cartagena de

los años 1981 al 2016.
Para lograr cumplir con este propósito se plantean los siguientes objetivos específicos:

- Seleccionar las variables del estudio, usando técnicas estadísticas que permitan identificar las de mayor relación con el comportamiento de las precipitaciones de la ciudad de Cartagena de los años 1981 al 2016.
- Construir el conjunto de datos a analizar, teniendo en cuenta las varíales de estudio seleccionadas, para con ello llevar a cabo el análisis de los mismos.
- Seleccionar técnicas de aprendizaje supervisado para identificar los comportamientos de los datos.
- Aplicación de las técnicas de aprendizaje automático supervisado; utilizando redes neuronales, maquinas de soporte vectorial y regresión logística.
- Analizar los resultados obtenidos de los modelos evaluados haciendo un comparativo de las técnicas utilizadas.

2.3. Alcance

Este proyecto tiene un enfoque investigativo en el cual se analizaran diversas técnicas de aprendizaje automático supervisado, con el fin de analizar los resultados obtenidos de la mismas y con ello realizar un análisis del comportamiento de las técnicas con el rendimiento o resultado de mayor relevancia.
Se desarrolla el análisis del modelo de aprendizaje supervisado, utilizando las técnicas de regresión logística, maquina de soporte vectorial y redes neuronales; con las cuales se realizan comparaciones de los resultados y/o rendimientos de las mismas. Con lo anterior se llega a la evaluación y medición de datos, al igual que a el análisis gráfico de los resultados de cada técnica evaluada y analizada.

Como soporte del proyecto de investigación realizado, se entregara el presente documento y un articulo investigativo, en el cual se evidenciaran los análisis, desarrollos, resultados y conclusión del presente proyecto.

3. Capítulo 2: Marcos de Referencia

3.1. Estado del arte

Al transcurrir de los años se ha evidenciado que los sistemas de información son parte fundamental en las instituciones tanto privadas como publicas a nivel mundial, esto ha generado un aumento a gran magnitud de la informaron ingresada, procesada y almacenada en los sistemas de información, generando la necesidad de analizar grandes cantidades de datos que de por sí solos no tendrían relevancias trascendentales.
En el trascurso de la búsqueda de estudios, tesis, artículos e investigaciones realizadas en el ámbito de inteligencia artificial y aprendizaje automático supervisado se destaca el trabajo realizado por Gustavo Ovando, Mónica Bocco y Silvina Sayago en el año 2005 titulado "Redes neuronales para modelar predicción de heladas" en el cual se desarrollaron modelos basados en redes neuronales del tipo "backpropagation", para predecir la ocurrencia de heladas, a partir de datos meteorológicos de temperatura, humedad relativa, nubosidad, dirección y velocidad del viento. El entrenamiento y la validación de las redes se realizaron utilizando 24 años de datos meteorológicos correspondientes a la estación de R o Cuarto, Córdoba, Argentina, separados en 10 años como conjunto de datos de entrenamiento y 14 como conjunto de datos de validación.

En la fase de entrenamiento se utilizaron 3650 datos, y a los efectos de comparación entre los distintos modelos planteados como de tiempos de computación dedicados a los mismos, se considera que el número de 20.000 iteraciones era su cliente para lograr un error significativo. Para los modelos analizados, se encontró que en general el porcentaje de datos con error de pronóstico se encuentra en aproximadamente el 2 % para 14 años de validación.
Estos errores se incrementan en porcentajes que oscilan, para el mismo periodo, entre un 10 y un 23 % cuando solo se consideran d as de heladas efectivas no pronosticadas.
Considerando la respuesta de las redes neuronales propuestas se puede asegurar que la dependencia de estas variables con la ocurrencia de heladas responde a una función no lineal, si bien las redes no proporcionan la expresión matemática de la misma; si se disminuye el número de neuronas de las capas ocultas en el planteamiento, o el numero de iteraciones de entrenamiento, los errores no varían en forma importante, por lo cual ambos parámetros pueden decidirse en función del tiempo de entrenamiento. [4]

Es evidente que los avances tecnológicos se extienden hacia todos los ámbitos en el desarro-

llo de las ciencias por lo que dentro de este se destaca el trabajo realizado por Solangel Rodríguez Vazque y Andy Vidal Martínez en 2015 el cual titulan como "clasificación de células cervicales con maquinas de soporte vectorial empleando rasgos del núcleo" en el cual se presenta el uso de las maquinas de soporte vectorial (SVM) como método computacional para la clasificación de las células cervicales en las condiciones normal y anómala, basándose solamente en las características extra das de la región ocupada por el núcleo, sin hacer uso de las características del citoplasma. La importancia de este enfoque viene dada porque los núcleos son las zonas que pueden ser segmentadas más fácilmente en imágenes complejas de frotis de Papanicolaou. Dichas imágenes presentan un alto grado de células superpuestas y es difícil lograr diferenciar las fronteras exactas de las regiones ocupadas por los citoplasmas; en esta técnica entre un 79 % y 86 % de predictibilidad negativa, de la misma manera que la predictibilidad positiva y el área bajo la curva ROC se mantienen entre rangos de valores que permiten validar la e ciencia del clasificador empleado para cada uno de los conjuntos de datos. Los valores obtenidos de acuerdo a las medidas F y H de igual forma se mantienen entre un 90-92 % y 85-91 % respectivamente, lo que muestra el nivel de efectividad del clasificador.

Los resultados obtenidos muestran que a través de la medida H es posible evaluar el comportamiento de la tasa de falsos negativos, mientras mayor sea el % de la media H, menor será la tasa de falsos negativos lo que brinda un buen desempeño en la realización de la prueba de Papanicolaou. [13]

Así mismo se destaca el articulo realizado en colaboración por Lara Vilar del Hoyo, María Pilar Martin Isabel y Javier Martínez Vega en el año 2008 titulado "Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional" en el cual Se aborda la realización de modelos de riesgo humano de incendio forestal mediante el empleo de técnicas de regresión logística, estimando la probabilidad de ocurrencia del fenómeno a partir de variables de tipo socioeconómico relacionadas con la ocurrencia de incendios forestales en las Comunidades Autónomas de Madrid y Valencia. Las variables independientes de riesgo se generan a partir de herramientas de Sistemas de Información Geográfica (SIG), a una resolución de 1 km2.

Los resultados obtenidos en la C. de Madrid tras llevar a cabo las correlaciones no paramétricas de Spearman señalan que no han de incluirse en el análisis las variables buffer de carreteras, pistas y maquina por su alta correlación con otras variables. A partir de tests no paramétricos de estadística comparativa se observa que las variables buffer líneas de ferrocarril, buffer líneas eléctricas, campos de tiro-canteras y montes consorciados no presentan diferencias significativas al 95 por ciento de confianza (p-valor mayor de 0,05) para dos muestras independientes del primer y cuarto cuartil (resultados del test de la U-Mann-Whitney) y que la variable buffer l neas eléctricas no es significativa en la comparación de las 4 muestras independientes, al 95 por ciento de confianza (resultados de la prueba de Kruskal-Wallis).

Los modelos obtenidos en las dos áreas de estudio han ofrecido resultados muy similares en cuanto al porcentaje de acierto, si bien en la C. Madrid se alcanza un porcentaje de acierto global

ligeramente superior (70,6 por ciento). En ambas zonas los modelos logran una mejor predicción de la baja incidencia de incendio, alcanzando mayor valor de acierto en la C. Madrid (76,4 por ciento) que en la C. Valenciana (57,4 por ciento).
A pesar de las limitaciones se ha demostrado que la metodología propuesta es aplicable a ámbitos geográficos muy diversos siempre que el conjunto de variables independientes representen adecuadamente los factores relacionados con la ocurrencia de incendios en la zona de interés. La consecución de modelos a una resolución espacial como la que se propone en este trabajo puede ser de gran interés para los gestores, permitiendo identificar zonas de alta ocurrencia de incendios y tipos de variables de riesgo humano influyentes en el mismo. Este análisis refleja la importancia de la distribución de usos en los diferentes territorios, y de como la acción del hombre está influyendo en el fenómeno de los incendios forestales. Indica la importancia de estos factores socioeconómicos y del interés en incluirlos en los sistemas generales de riesgo de incendio forestal. [14]

3.2. Marco Teórico

3.2.1. El nacimiento de la Inteligencia Artificial (1956)

El primer trabajo que es reconocido generalmente como perteneciente a la inteligencia artificial fue realizado por Warren McCulloch y Walter Pitts. Ellos propusieron un modelo de neurona artificial, en el cual, cada neurona se caracterizaba por un estado de "on"-"off"; el cambio a "on" ocurría en respuesta a la estimulación hecha por un número suficiente de neuronas vecinas. Ellos mostraron que cualquier función computable puede ser programada por una red de neuronas conectadas y que todos las conexiones lógicas (and, or, not, ...) pueden ser implementadas por estructuras de red simples. McCulloch y Pitts también sugirieron que las Redes de Neuronas Artificiales podrían aprender.

Donald Hebb desarrolló una regla simple para modificar el peso de las conexiones entre las neuronas. Su regla (Hebbian Learnig) sigue siendo un modelo útil a día de hoy. En 1950, Marvin Minsky y Dean Edmons construyeron el primer ordenador neuronal: el SNARC, que simulaba una red de 40 neuronas. Minsky siguió estudiando la computación universal usando redes de neuronas, siendo bastante escéptico en cuando a las posibilidades reales de las Redes de Neuronas Artificiales [MP69]. Fue el autor de influyentes teoremas que demostraban las limitaciones de las redes de neuronas artificiales.

Años después, el "nacimiento oficial" de la inteligencia artificial tomo lugar en el verano de 1956 en el Dartmouth College de Stanford. El padre fue John McCarthy que convenció a Minsky, Claude Shannon, y Nathaniel Rochester para juntar a los investigadores más ilustres en los campos de la teoría de autómatas, redes neuronales y del estudio de la inteligencia, con el fin de organizar unas jornadas de trabajo durante los dos meses del verano de 1956. Las jornadas de

Dartmouth [MMRS55] no incluyeron ninguna línea rompedora, pero el nuevo campo de la Inteligencia Artificial estuvo sometido por los participantes y sus alumnos durante las siguientes dos décadas. En Dartmouth, se definió por qué es necesaria una nueva disciplina en vez de agrupar los estudios en Inteligencia Artificial dentro de alguna de las ya existentes: teoría del control, de la toma de decisiones, operaciones, matemáticas. La primera razón es porque la inteligencia artificial trata de duplicar facultades humanas como la creatividad, el auto aprendizaje o el uso del lenguaje. Otra razón, es porque la metodología usada parte de la ciencia de la computación y la Inteligencia Artificial es la única especialidad que trata de hacer máquinas las cuales puedan funcionar autónomamente en entornos complejos y dinámicos.[12]

3.2.2. La Inteligencia Artificial adopta el método científico (1987 - presente)

Desde los últimos años de los 80 y hasta el presente, se ha creado una revolución tanto en el contenido como en la metodología de trabajo de la inteligencia artificial. Últimamente, es más común crear a partir de teorías ya existentes que desarrollar nuevas, proporcionando a estas teorías de rigor matemático y mostrando su eficacia en problemas reales más que en simulaciones o ejemplos simples de laboratorio. En términos metodológicos, la inteligencia artificial ha adoptado firmemente el método científico. Para que una hipótesis sea aceptada, debe estar sometido a experimentos empíricos implacables, y los resultados deben ser analizados estadísticamente para medir su importancia. Ahora es posible replicar los experimentos usando repositorios de datos compartidos, así como datos y código de testeo. Un ejemplo para ilustrar lo anterior sería el reconocimiento del habla.

En la década de los 70, una gran variedad de arquitecturas y aproximaciones fueron probadas; muchas de ellas fueron hechas "ad-hoc" y con un planteamiento teórico muy débil. Siendo probadas en sólo unos pocos experimentos muy restringidos. En los últimos años, aproximaciones basadas en los modelos ocultos de Markov (Hidden Markov Models) parecen controlar este área. Dos características son importantes en los modelos de Markov: están basados en una teoría matemática rigurosa y son generados a partir de un gran conjunto de datos reales del habla.

Las redes neuronales artificiales han tenido un proceso similar: en un principio el enfoque de muchos desarrollos era mostrar cómo las redes de neuronas diferían de las técnicas "tradicionales". Con el desarrollo de la metodología y de unos marcos teóricos robustos, se consigue comparar las redes neuronales a las técnicas estadísticas, reconocimiento de patrones y, en general a las técnicas más relevantes de cada aplicación. A raíz de estos desarrollos, la minería de datos se ha convertido en una nueva y vigorosa industria. Por último, es importante resaltar el papel del razonamiento probabilístico en muchos campos de la Inteligencia Artificial, basado fundamentalmente en las redes bayesianas. Las redes bayesianas fueron inventadas para permitir una representación eficiente y un razonamiento riguroso del conocimiento incierto (uncertain knowledge).

Quizás por el avance de la Inteligencia Artificial en la solución de problemas muy específicos, se ha vuelto a plantear la cuestión de la solución de problemas generales o desde un punto de

vista holístico. Esto da lugar a los sistemas de agentes inteligentes, donde agentes autónomos y especializados en ciertas tareas colaboran entre sí, generando un conocimiento mucho más global. Uno de los ejemplos más conocidos de arquitectura basadas en agentes es el sistema SOAR. Y uno de los entornos más importantes para los agentes inteligentes es Internet: motores de búsquedas, sistemas de recomendación o sistemas de agregación de sitios web. A pesar de todo, algunos autores de los más influyentes en el campo de la Inteligencia Artificial (John McCarthy, Marvin Minsky, Nils Nilsson y Patrick Winston) han expresado su descontento con los progresos de la Inteligencia Artificial, ellos piensan que más que seguir mejorando el rendimiento en ciertas áreas o ejemplos concretos; la inteligencia artificial debe retornar al principio expresado por Simon: "máquinas que piensan, que aprenden y que crean". De esta corriente han surgido nuevas líneas de trabajo: Inteligencia Artificial Humana,Inteligencia Artificial General e Inteligencia Artificial amigable. [12]

3.2.3. Aprendizaje Automático Supervisado

El objetivo principal del aprendizaje automático, es crear algoritmos capaces de extender comportamientos y reconocer patrones a partir de una información proporcionada en forma de ejemplos. Por lo tanto, es un proceso de incitación del conocimiento, es decir, un método que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares.

Cuando se han observado todos los casos particulares la inducción se considera completa, por lo que la generalización a la que da lugar se considera válida. No obstante, en la mayoría de los casos es imposible alcanzar una inducción completa, por lo que el enunciado a que da lugar queda sujeto a un cierto grado de inseguridad, y en consecuencia no se puede considerar como un esquema de inferencia formalmente válido ni se puede justificar empíricamente. En muchas ocasiones el campo de actuación del aprendizaje automático se oculta con el de Data Mining, ya que las dos disciplinas están enfocadas en el análisis de datos, sin embargo el aprendizaje automático se centra más en el estudio de la complejidad computacional de los problemas con la intención de hacerlos posibles desde el punto de vista práctico, no únicamente teórico.

A un nivel muy básico, se puede decir que una de las tareas del Aprendizaje Automático es intentar extraer conocimiento sobre algunas propiedades no observadas de un objeto basándose en las propiedades que sí han sido observadas de ese mismo objeto (incluso de propiedades observadas en otros objetos similares)o, en palabras más llanas, predecir comportamiento futuro a partir de lo que ha ocurrido en el pasado. Un ejemplo de mucha actualidad sería, por ejemplo, el de predecir si un determinado producto le va a gustar a un cliente basándose en las valoraciones que ese mismo cliente ha hecho de otros productos que sí ha probado.

En cualquier caso, como el tema del que se esta hablando está relacionado con el aprendizaje, lo primero que se debe cuestionar es ¿Qué se entiende por aprender? y, ya que se quiere dar metodologías generales para producir un aprendizaje de forma automática, una vez que se fije este concepto habremos de dar métodos para medir el grado de éxito/fracaso de un aprendizaje.

En cualquier caso, se traslada un concepto intuitivo y que se usa normalmente en la vida diaria a un contexto computacional, ha de tenerse en cuenta que todas las definiciones que se utilicen de aprendizaje desde un punto de vista computacional, así como las diversas formas de medirlo, estarán íntimamente relacionadas con contextos muy concretos y posiblemente lejos de lo que intuitivamente, y de forma general, se entiende por aprendizaje.[3]

3.2.4. Métodos de Aprendizaje Automático Supervisado

Dentro del aprendizaje automático supervisado existen tres métodos de aplicación diferenciados, que son los siguientes:

Método de Regresión

Este método es una técnica estadística empleada para estudiar la relación entre dos o mas variables. En el entorno de la investigación es utilizada para predecir un gran rango de fenómenos.

Método de Clasificación

Este método se utiliza para predecir los resultados de un atributo con valor reservado (a, b, c, . . .) dadas unas características (Xo, X1, X2, X3,.... Xn). El método simple de clasificación es el binario, donde se clasifica un registro de variables de entrada en 1 o 0. La clasificación múltiple es una extensión de la clasificación binaria.

3.2.5. Algoritmos de Clasificación Supervisado

Los algoritmos utilizados para el problema de la clasificación supervisada actúan usualmente sobre la información proporcionada por un conjunto de muestras, patrones, ejemplos o prototipos de entrenamiento que son aceptados como representantes de las clases, y los mismos poseen una etiqueta de clase correcta. A este conjunto de prototipos correctamente etiquetados se le llama conjunto de entrenamiento, y es el conocimiento empleado para la clasificación de nuevas muestras.
Por otra parte también se puede decir que estos inspeccionan todo el conocimiento almacenado en el conjunto de entrenamiento para así determinar cuál será la clase a la que corresponde una nueva muestra, pero únicamente tiene en cuenta el vecino más próximo a ella, por lo que es lógico pensar que es posible que no se esté aprovechando de forma eficaz toda la información que se podría extraer del conjunto de entrenamiento.
Estos algoritmos tienen como objetivo principal, decidir cuál es la clase, de las que ya se tiene conocimiento, a la que debe pertenecer una nueva muestra, teniendo en cuenta la información que se puede extraer del conjunto de entrenamiento.

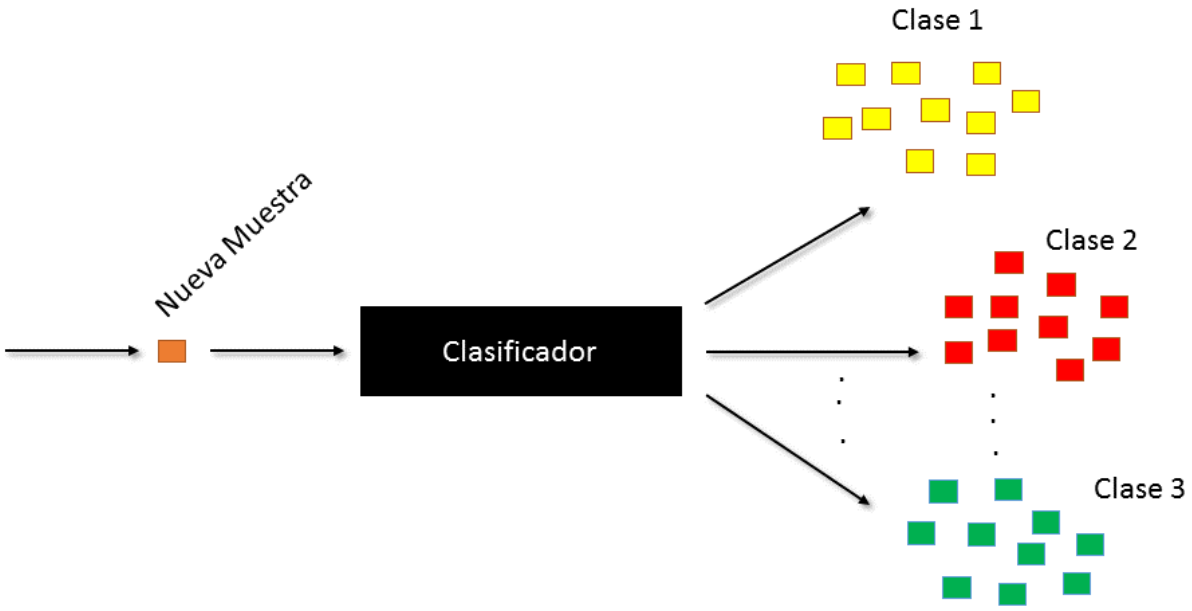


Figura 3-1.: Clasificación Supervisada

El aprendizaje Automático supervisado puede ser aplicada en tareas de:

- **Clasificación:** Donde el enfoque principal del estudio de la técnica es identificar a que clase pertenece una nueva entrada, ejemplos específicos son la clasificación de documentos, imágenes, diagnóstico médico, etc.
- **Regresión:** En el cual se predice un valor real para cada ítem, ejemplos específicos son la predicción de la demanda, stocks de inventarios, variables económicas, tasas, etc.
- **Ranking:** Enfocado y utilizado para la organización de ítems basado en cierto criterio, como por ejemplo las búsquedas en la web.
- **Clustering:** Son utilizadas en mayor escala en procesos comerciales, donde se segmentan clientes y productos, con el fin de facilitar los procesos de decisiones concernientes a que vender y a quienes.
- **Reducción de Dimensionalidad:** Convierte la representación de los ítems originarios en una representación de baja dimensionalidad, perseverando las propiedades de la inicial representación. Un ejemplo específicos esto se evidencia en el reprocesamiento de imágenes digitales.

También podrían encontrar la aplicación del aprendizaje automático supervisado en el procesamiento de lenguaje natural, reconocimiento de sonido, reconocimiento de caracteres ópticos (OCR). Reconocimiento de rostro, etc.
Entre las técnicas de aprendizaje automático supervisado usadas en tareas de regresión se encuentran:

- Multilayer Perceptron Neural Network (MLP)
- Radial Basis Function Neural Network (RBF)
- Generalized Regression Neural Network (GRNN)
- K Nearest Neighbor Regression (KNN)
- Classification and regression Trees (CART)
- Support Vector Regression (SVR)
- Gaussian Processes (GP)
- Ridge Regression
- Kernel Ridge Regression

3.3. Modelos de Aplicación

En esta sección se describirán los modelos (técnicas) seleccionadas para el análisis y desarrollo del presente proyecto investigativo.

3.3.1. Regresión logística

El concepto de regresión es uno de los pilares de la estadística, y data al menos de principios de 1800 con los trabajos de Legendre, Gauss y Laplace. El término regresión fue introducido en 1889 por Francis Galton, en su libro Natural inheritance, quien acuñó el término regresión hacia la media". Esta denominación proviene de que los valores pronosticados en la variable dependiente (VD) a partir de los valores de la variable independiente (VI), tienen varianza menor que la de la variable dependiente empírica (Var(Y') $<$ Var(Y)).
La regresión logística es una de las técnicas estadístico-inferenciales más empleadas en la producción científica contemporánea. Surge en la década de los 60, su generalización dependía de la solución que se diera al problema de la estimación de los coeficientes. El algoritmo de Walker-Duncan para la obtención de los estimadores de máxima verosimilitud vino a solucionar en parte este problema, pero era de naturaleza tal que el uso de ordenadores era imprescindible.

Esta técnica va a contestar a preguntas tales como: ¿Se puede predecir con antelación si un cliente que solicita un préstamo a un banco va a ser un cliente moroso?. ¿Se puede predecir si una empresa va a entrar en bancarrota?. ¿Se puede predecir de antemano que un paciente corra riesgo de un infarto?, por lo tanto se utiliza cuando se desea pronosticar la probabilidad de que ocurra o no un suceso determinado.

Se dice que un proceso es binomial cuando sólo tiene dos posibles resultados: "éxito" y "fracaso", siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones.
Un proceso binomial está caracterizado por la probabilidad de éxito, representada por p, la probabilidad de fracaso se representa por q y, ambas probabilidades están relacionadas por p+q=1. En ocasiones, se usa el cociente p/q, denominado "odds", y que indica cuánto más probable es el éxito que el fracaso, como parámetro característico de la distribución binomial.[8]
Los modelos de regresión logística son modelos de regresión que permiten estudiar si una variable binomial depende, o no, de otra u otras variables (no necesariamente binomiales): Si una variable binomial de parámetro p es independiente de otra variable X, se cumple p=p $\rightarrow$ X, por consiguiente, un modelo de regresión es una función de p en X que a través del coeficiente de X permite investigar la relación anterior.
Existen varias implementaciones de regresión logística en la investigación estadística, que utilizan diferentes técnicas de aprendizaje. El algoritmo de Regresión logística se ha implementado utilizando una variación del algoritmo de Red neuronal. Este algoritmo comparte muchas de las cualidades de las redes neurales pero es más fácil de entrenar.
Una de las ventajas de la regresión logística es que el algoritmo es muy flexible, puede tomar cualquier tipo de entrada y admite varias tareas analíticas diferentes:[9]

- Usar datos demográficos para realizar predicciones sobre los resultados, como el riesgo de contraer una determinada enfermedad.
- Explorar y ponderar los factores que contribuyen a un resultado. Por ejemplo, buscar los factores que influyen en los clientes para volver a visitar un establecimiento.
- Clasificar los documentos, el correo electrónico u otros objetos que tengan muchos atributos.

La dificultad para poder discriminar entre los efectos mencionados, y para la verificación empírica de los supuestos del modelo, conllevan que este primer concepto de regresión haya evolucionado. Actualmente la regresión, en un sentido amplio, designa al conjunto de procedimientos empleados para construir funciones matemáticas (con su correspondiente término de error en el caso de los modelos lineales), y sus transformaciones "logit", que permiten estimar o predecir el comportamiento de una o más variables a partir de otras variables, con las que se encuentran fuertemente correlacionadas.
En conclusión, si se desea que el modelo proporcione directamente la probabilidad de pertenecer a cada uno de los grupos, se debe transformar la variable respuesta de algún modo para garantizar que la respuesta prevista esté entre cero y uno. Si se tomaran,

$$p_i = F(\beta_0 + \beta_1^{'} x_i) \tag{3-1}$$

Se garantiza que pi esté entre cero y uno si se exige que F tenga esa propiedad. La clase de funciones no decrecientes, acotadas entre cero y uno, es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como F cualquier función de distribución.
Habitualmente se toma como F la función de distribución logística, dada por:

$$pi = \frac{1}{(1 + e^{-}(\beta_0 + \beta_1^{'} x_i)} \tag{3-2}$$

Esta función tiene la ventaja de ser continua. Además, como,

$$1 - pi = \frac{e^{-}(\beta_0 + \beta_1^{'} x_i)}{(1 + e^{-}(\beta_0 + \beta_1^{'} x_i)} = \frac{1}{(1 + e^{-}(\beta_0 + \beta_1^{'} x_i)} \tag{3-3}$$

resulta que

$$g_i = \log \frac{p_i}{1 - p_i} = \log \left(\frac{\frac{1}{1+e^{-(\beta_0+\beta_1^{'} x_i)}}}{\frac{e^{-(\beta_0+\beta_1^{'} x_i)}}{1+e^{-(\beta_0+\beta_1^{'} x_i)}}} \right) = \log \left(\frac{1}{e^{-(\beta_0+\beta_1^{'} x_i)}} \right) = \beta_0 + \beta_1^{'} x_i \tag{3-4}$$

de modo que, al hacer la transformación, se tiene un modelo lineal que se denomina logit.
La variable g representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones y, al ser una función lineal de las variables explicativas, facilita la estimación y la interpretación del modelo.

3.4. Máquinas de Soporte Vectorial

Las maquinas de vectores de soporte (SVM por sus siglas en ingles, "Support Vector Machine"), fueron desarrolladas por Vapnik (1995), para el problema de clasificación pero la forma actual de SVM para regresión fue desarrollada en los laboratorios de ATT por Vaptnik. SVM esta ganando gran popularidad como herramienta para la identificación de sistemas no lineales, esto debido principalmente a que SVM esta basado en el principio de minimización del riesgo estructural (SRM por sus siglas en ingles, "Structural Risk Minimization"), principio originado de la teoría de aprendizaje estadístico desarrollada por Vapnik en el cual ha demostrado ser superior al principio de minimización del riesgo empírico (ERM por sus siglas en ingles .Em pirical Risk Minimization"), utilizado por las redes neuronales convencionales. Algunas de las razones por las que este método ha tenido éxito es que no padece de mínimos locales y el modelo solo depende de los datos con mas información llamados vectores de soporte(SV por sus siglas en ingles, "Support Vectors"). Las grandes ventajas que tiene SVM son:

- Una excelente capacidad de generalización, debido a la minimización del riesgo estructurado.
- Existen pocos parámetros a ajustar; el modelo solo depende de los datos con mayor información.
- La estimación de los parámetros se realiza a través de la optimización de una función de costo convexa, lo cual evita la existencia de un mínimo local.
- La solución de SVM es separse, esto es que la mayoría de las variables son cero en la solución de SVM, esto quiere decir que el modelo final puede ser escrito como una combinación de un numero muy pequeño de vectores de entrada, llamados vectores de soporte.

Las Máquinas de Soporte Vectorial actualmente es una nueva técnica de clasificación y ha tomado mucha atención en años recientes. La teoría de la SVM está basada en la idea de minimización de riego estructural (SRM). En muchas aplicaciones, las SVM han mostrado tener gran desempeño, más que las máquinas de aprendizaje tradicional como las redes neuronales y han sido introducidas como herramientas poderosas para resolver problemas de clasificación. Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor (i.e.: si los puntos de entrada están en 2 entonces son mapeados por la SVM a 3) y encuentra un hyperplano que los separe y maximice el margen m entre las clases en este espacio como se aprecia en la Figura 3-2. [1]

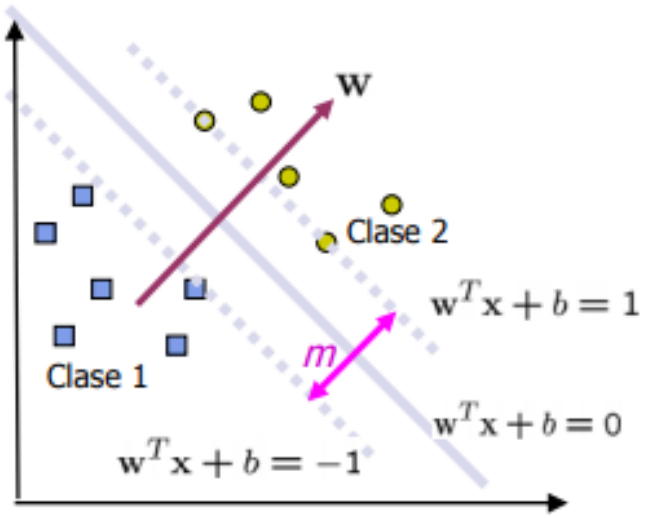


Figura 3-2.: La frontera de desición debe estar tan lejos de los datos de ambas clases como sea.[1]

Maximizar el margen m es un problema de programación cuadrática (QP) y puede ser resuelto por su problema dual introduciendo multiplicadores de Lagrange. Sin ningún conocimiento del

mapeo, la SVM encuentra el hyperplano óptimo utilizando el producto punto con funciones en el espacio de características que son llamadas kernels. La solución del hyperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte.[1]

Espacios inducidos por la función Kernel

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal estas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio del Kernel ofrece una solución alternativa a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de la máquinas de aprendizaje lineal. La forma más común en que las máquinas de aprendizaje lineales aprenden una función objetivo es cambiando la representación de la función, esto es similar a mapear el espacio de entradas X a un nuevo espacio de características.[7]

$$F = \{\phi(x)|x \in\} \tag{3-5}$$

Esto es:

$$x = \{x_1, x_2, ...x_n\} \rightarrow \phi(x) = \{\ \phi(x)_1, \phi(x)_2, ...\phi(x)_n\} \tag{3-6}$$

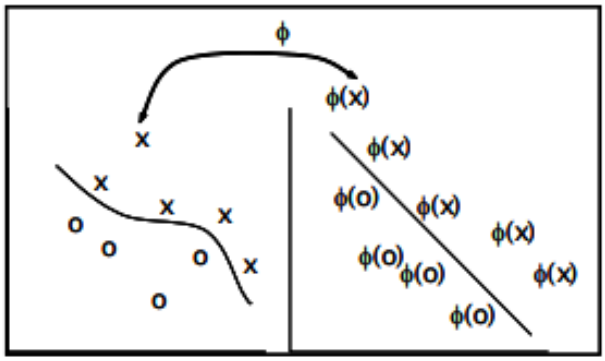


Figura 3-3.: Mapeo del espacio de entradas a un espacio de caracteristicas de mayor dimensión.[7]

En la Figura 3.3 se muestra un mapeo de un espacio de entradas de dos dimensiones a un espacio de características de dos dimensiones, donde la información no puede ser separada por una máquina lineal en el espacio de entradas mientras que en el espacio de características esto resulta muy sencillo.
Las máquinas de aprendizaje lineales son funciones reales

$$f : X \in R^n \rightarrow Y \in R \tag{3-7}$$

La función f se considera como una función lineal de x ∈ X, tal que se pude escribir como

$$f(x) = (w.r) + b \tag{3-8}$$

$$= wx^T + b \tag{3-9}$$

$$= \sum_{i=1}^{n} w_i x_i + b \tag{3-10}$$

donde w es el vector de pesos y b es el bias, términos tomados de la literatura de redes neuronales. Este tipo de máquinas admiten una representación dual, esto es si se define a

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{3-11}$$

se tiene que la función lineal se puede escribir en su forma dual esto es:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i (x_i.x_i) + b \tag{3-12}$$

donde (.) es el producto interno. Una propiedad importante de la representación dual es que la información de entrenamiento entra a la función a través de las entradas de la matriz de Gram = (xi.xi).

3.5. Redes Neuronales

El hombre se ha caracterizado siempre por su búsqueda constante de nuevas vías para mejorar sus condiciones de vida. Estos esfuerzos le han servido para reducir el trabajo en aquellas operaciones en las que la fuerza juega un papel primordial. Los progresos obtenidos han permitido dirigir estos esfuerzos a otros campos, como por ejemplo, a la construcción de máquinas calculadoras que ayuden a resolver de forma automática y rápida determinadas operaciones que resultan tediosas cuando se realizan a mano. Uno de los primeros en acometer esta empresa fue Charles Babbage, quien trató infructuosamente de construir una máquina capaz de resolver problemas matemáticos. Posteriormente otros tantos intentaron construir máquinas similares, pero no fue hasta la Segunda Guerra Mundial, cuando ya se disponía de instrumentos electrónicos, que se empezaron a recoger los primeros frutos. En 1946 se construyó la primera computadora electrónica, ENIAC. Desde entonces los desarrollos en este campo han tenido un auge espectacular. Estas

máquinas permiten implementar fácilmente algoritmos para resolver multitud de problemas que antes resultaban engorrosos de resolver. Sin embargo, se observa una limitación importante: ¿qué ocurre cuando el problema que se quiere resolver no admite un tratamiento algorítmico, como es el caso, por ejemplo, de la clasificación de objetos por rasgos comunes? Este ejemplo demuestra que la construcción de nuevas máquinas más versátiles requiere un enfoque del problema desde otro punto de vista. Los desarrollos actuales de los científicos se dirigen al estudio de las capacidades humanas como una fuente de nuevas ideas para el diseño de las nuevas máquinas. Así, la inteligencia artificial es un intento por descubrir y describir aspectos de la inteligencia humana que pueden ser simulados mediante máquinas. Esta disciplina se ha desarrollado fuertemente en los últimos años teniendo aplicación en algunos campos como visión artificial, demostración de teoremas, procesamiento de información expresada mediante lenguajes humanos... etc.
Las redes neuronales son más que otra forma de emular ciertas características propias de los humanos, como la capacidad de memorizar y de asociar hechos. Si se examinan con atención aquellos problemas que no pueden expresarse a través de un algoritmo, se observará que todos ellos tienen una característica en común: la experiencia. El hombre es capaz de resolver estas situaciones acudiendo a la experiencia acumulada. Así, parece claro que una forma de aproximarse al problema consista en la construcción de sistemas que sean capaces de reproducir esta característica humana. En definitiva, las redes neuronales no son más que un modelo artificial y simplificado del cerebro humano, que es el ejemplo más perfecto del que se dispone para un sistema que es capaz de adquirir conocimiento a través de la experiencia. Una red neuronal es "un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona".[2]

Ventajas que ofrecen las redes neuronales

Debido a su constitución y a sus fundamentos, las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro. Por ejemplo, son capaces de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas. Entre las ventajas se incluyen:

- Aprendizaje Adaptativo. Capacidad de aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.
- Auto-organización. Una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.
- Tolerancia a fallos. La destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

- Operación en tiempo real. Los cómputos neuronales pueden ser realizados en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.
- Fácil inserción dentro de la tecnología existente. Se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilitará la integración modular en los sistemas existentes.

La primera regla para actualizar los pesos de una red neuronal se conoce como la regla del Perceptrón, o también el procedimiento de convergencia del perceptrón. Esta primera regla modifica los elementos $W_(k)$ de acuerdo al algoritmo básico de la regla del perceptrón:

$$w_k + 1 = w_k + \alpha \left(\frac{\epsilon_k}{2}\right) X_k \tag{3-13}$$

Desarrollada por Rosenblatt, esta regla actualiza $W_(k)$ sólo si el error $\varepsilon\ \kappa$ es diferente de cero. El vector de pesos es $W_(k+1)$ y α es la tasa de aprendizaje del sistema (un valor constante muy pequeño que no cambia en el tiempo).
Un poco después, Mays desarrollo su famosa serie de algoritmos adaptativos para redes neuronales, los cuales son una versión mas general de la regla del Perceptrón. Esta nueva regla tambien es llamada de incremento adaptativo o algoritmo de Mays.

$$w_k + 1 = \begin{cases} W_k + (\alpha\epsilon_k)\left(\frac{X_k}{2|X_k|^2}\right) si|s_k| \geq y \\ W_k + (\alpha d_k)\left(\frac{X_k}{|X_k|^2}\right) si|s_k| \geq y \end{cases} \tag{3-14}$$

Donde $\varepsilon\ \kappa$ es el error κ-y κ, siendo "d"la respuesta deseada y ε"la salida por la red. Por lo general , la "zona muerta"delimitada por γ es un valor pequeño, y si vale 0 el algoritmo de Mays se convierte en la regla del perceptrón.

3.6. Microsoft Azure Machine Learning Studio

Microsoft Azure Machine Learning Studio es una herramienta de arrastrar y colocar que le permite crear, probar e implementar soluciones de análisis predictivo en sus datos. Machine Learning Studio publica modelos como servicios web que pueden utilizarse fácilmente en aplicaciones personalizadas o herramientas de BI como Excel. Machine Learning Studio es el lugar en el que confluyen la ciencia de datos, el análisis predictivo, los recursos en la nube y sus datos.
Para desarrollar un modelo de análisis predictivo, normalmente se utilizan datos de una o varias fuentes, se transforman y analizan los datos a través de diversas funciones estadísticas y de manipulación de datos y se genera un conjunto de resultados. Desarrollar un modelo como este es un proceso iterativo: a medida que se modifican las diversas funciones y sus parámetros, sus resultados convergen hasta que esté satisfecho con un modelo entrenado y efectivo. Azure

Machine Learning Studio le proporciona un área de trabajo visual e interactiva para generar, probar e iterar con toda facilidad sobre un modelo de análisis predictivo. Se arrastran y colocan conjuntos de datos y módulos de análisis en un lienzo interactivo, conectándolos todos para formar un experimento que se ejecuta en Machine Learning Studio. Para iterar su diseño de modelo, se puede editar el experimento, guardar una copia si así se desea y ejecutarlo de nuevo. Cuando esté listo, puede convertir el experimento de entrenamiento en un experimento predictivo, y luego publicarlo como servicio web para que otros usuarios puedan acceder al modelo. No se requiere ningún tipo de programación, basta con conectar visualmente conjuntos de datos y módulos para construir el modelo de análisis predictivo.[10]

3.6.1. Entorno Microsoft Azure Machine Learning Studio

En el entorno al iniciar sesion en Microsoft Azure Machine Learning Studio, se cuenta con las siguientes pestañas:

- **Proyectos :**Colecciones de experimentos, conjuntos de datos, cuadernos y otros recursos que representan un proyecto individual
- **Experimentos:**Experimentos que ha creado y ejecutado, o que ha guardado como borrador.
- **Servicio Web :**Servicios web que implementó a partir de los experimentos.
- **Cuadernos :**cuadernos de Jupyter que creó.
- **Conjuntos De Datos :** Conjuntos de datos que cargó en Estudio.
- **Modelos Entrenados :**Modelos que entrenó en experimentos y guardó en Estudio.
- **Configuración :**Una colección de ajustes que puede utilizar para configurar la cuenta y los recursos.

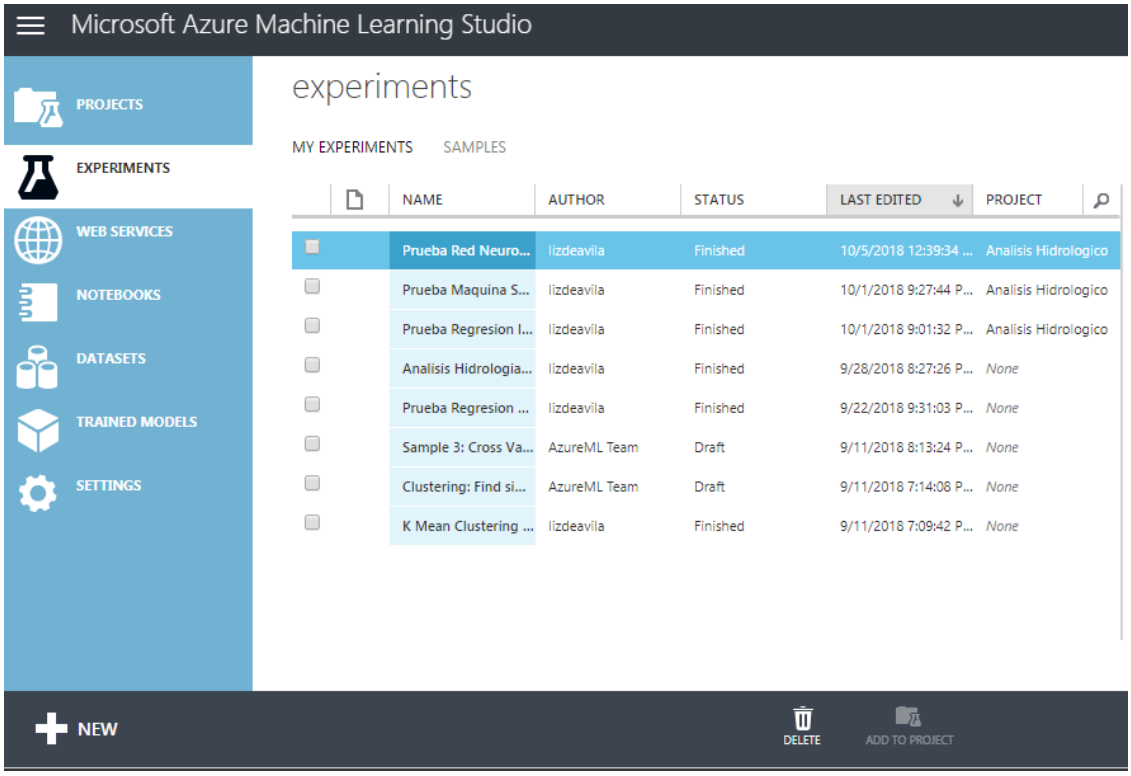


Figura 3-4.: Entorno inicial de Microsoft Azure Machine Learning Studio.

3.6.2. Componentes de un experimento en Microsoft Azure Machine Learning Studio

Un experimento consta de conjuntos de datos que proporcionan datos a módulos analíticos, que se conectan en conjunto para construir un modelo de análisis predictivo. En concreto, un experimento válido tiene estas características:

- Tiene al menos un conjunto de datos y un módulo.
- Los conjuntos de datos pueden estar solo conectados a módulos.
- Los módulos pueden conectarse a conjuntos de datos o a otros módulos.
- Todos los puertos de entrada de los módulos deben tener alguna conexión al flujo de datos.
- Deben establecerse todos los parámetros necesarios para cada módulo.

Puede crear un experimento desde cero, o puede usar un experimento de ejemplo existente como plantilla.

Conjuntos de datos

Un conjunto de datos son datos que se han cargado en Machine Learning Studio para utilizarse en el proceso de modelado.

Módulos

Un módulo es un algoritmo que puede aplicar sobre sus datos. Machine Learning Studio cuenta con diversos módulos que van desde las funciones de incorporación de datos hasta procesos de entrenamiento, puntuación y validación. Algunos de los ejemplos de los módulos incluidos:

- **Convertir a ARFF:** Convierte un conjunto de datos serializados de .NET a formato ARFF.
- **Estadísticas elementales de procesos:** Calcula estadísticas elementales como la media, la desviación estándar, etc.
- **Regresión lineal:** Crea un modelo de regresión lineal basado en un descenso de gradiente en línea.
- **Puntuar modelo:** Puntúa un modelo entrenado de clasificación o regresión.

Un módulo puede tener un conjunto de parámetros que puede utilizar para configurar los algoritmos internos del módulo. Al seleccionar un módulo en el lienzo, los parámetros del módulo se muestran en el panel Propiedades a la derecha del lienzo. Puede modificar los parámetros en ese panel para ajustar su modelo.

3.7. Marco Legal

Con el fin de tratar de controlar la utilización de productos o software por personas externas sin el consentimiento de las personas que son autores de los mismos; la Dirección Nacional del Derecho de Autor se ha hecho presente en Intersoftware 95, esto con el objetivo de promover entre los programadores la utilización del Registro Nacional de Derecho de Autor, el cual es un espacio en el que se inscriben las obras literarias y artísticas que se crean en el territorio nacional.

La normalización en relación a los derechos que obtiene el autor de soporte lógico o de software, y las consecuencias jurídicas que sobrevienen a su licenciamiento, transferencia, distribución, reproducción o, en general, cualquier utilización que se haga de ellos, están contempladas en la Ley 23 de 1982, la Decisión 351 del Acuerdo de Cartagena y el Decreto 1360 de junio 23 de 1989.
Según el Acuerdo de Cartagena, se estipula que los programas operativos como a los aplicativos, sea este en forma de código fuente o código objeto, estos están resguardados por la Ley, en la que se establece, que “El propietario de un ejemplar de programa de computador de circulación lícita,

puede realizar copias o adaptaciones del mismo, siempre y cuando sean indispensables para su utilización o se realicen con fines de archivo o sustitución del original en caso de daño o pérdida".

No se considera lícito la adaptación de un programa de carácter personal por parte de uno o diversos usuarios ya sea por medio de la instalación de redes, estaciones de trabajo o cualquier procedimiento análogo, sin el consentimiento del titular de los derechos; Sin embargo, los autores o titulares pueden autorizar las modificaciones o adaptaciones necesarias para la correcta utilización de los programas.

La Ley 44 de 1993 especifica penas entre dos y cinco años de cárcel, así como el pago de indemnizaciones por daños y perjuicios a quienes comentan el delito de piratería de software. Se considera delito el uso o reproducción de un programa de computador de manera diferente a como esté estipulado en la licencia. Los programas que no tengan licencia son ilegales y es necesaria una licencia por cada copia instalada en los computadores; a partir del mes de julio de 2001, y gracias a la reforma hecha al Código de procedimiento penal, quien sea encontrado usando, distribuyendo o copiando software sin licencia tendrá que pagar con cárcel hasta por un periodo de 5 años.

3.7.1. Licencias

Licencias GPL

Es una de las más utilizadas y se suele denominar como GNU GPL. Con esta licencia el desarrollador conserva los derechos de autor, pero permite su libre distribución, modificación y uso siempre y cuando, en el caso de que el software se modifique, el nuevo software que se desarrolle como resultado quede obligatoriamente con la misma licencia. Incluso si el software con licencia GPL solo fuera una parte de otro programa, este programa también tendría que mantener la licencia. Está considerada la primera licencia copyleft y, bajo esta filosofía, cualquier código fuente licenciado bajo GPL, debe estar disponible y accesible, para copias ilimitadas y a cualquier persona que lo solicite. De cara al usuario final, el software licenciado bajo GPL es totalmente gratuito, pudiendo pagar únicamente por gastos de copiado y distribución. [5]

Licencia AGPL

Se engloba dentro de las licencias destinadas a modificar el derecho de autor derivadas de GNU. La novedad de AGPL es que, aparte de las cláusulas propias de una GNU GPL, ésta obliga a que se distribuya el software que se destine a dar servicios a través de una red de ordenadores, es decir, si se quiere usar como parte del desarrollo de un nuevo software, éste quedaría obligado a su libre distribución.

Licencia BSD

Es un buen ejemplo de una licencia permisiva que casi no impone condiciones sobre lo que un usuario puede hacer con el software. El software bajo esta licencia es la menos restrictiva para los desarrolladores, ya que, por ejemplo, el software puede ser vendido y no hay obligaciones de incluir el código fuente. Además, una aplicación licenciada con BSD permite que otras versiones puedan tener otros tipos de licencias, tanto libres como propietarias; un buen ejemplo de ello es el conocido sistema operativo Mac OS X, desarrollado bajo esta licencia. También, BSD permite el cobro por la distribución de objetos binarios.

Licencia Apache

El software bajo este tipo de licencia permite al usuario distribuirlo, modificarlo, y distribuir versiones modificadas de ese software pero debe conservar el copyright y el disclaimer. La licencia Apache no exige que las obras derivadas (las versiones modificadas) se distribuyan usando la misma licencia, ni siquiera que se tengan que distribuir como software libre, solo exige que se informe a los receptores que en la distribución se ha usado código con la licencia Apache. En este sentido, al crear nuevas piezas de software, los desarrolladores deben incluir dos archivos en el directorio principal de los paquetes de software redistribuidos: una copia de la licencia y un documento de texto que incluya los avisos obligatorios del software presente en la distribución.

4. Diseño Metodológico

A continuación se detallan los aspectos metodológicos utilizados para el cumplimiento de los objetivos planteados en este proyecto.

4.1. Línea de investigación

La institución cuenta con grupos de investigación tecnológica, la escuela de ingeniería de sistemas de la universidad del Sinú, las cuales constan de la línea de Desarrollo de Software, Inteligencia Artificial y Redes de Computo; este proyecto está enfocado en la línea de conocimiento de Inteligencia Artificial, permitiendo con ello fortalecer la misma, ya que se utilizan técnicas como Machine Learning, la cual consiste en una serie de algoritmos que permiten que un dispositivo o aplicación sean artificialmente inteligentes.

4.2. Tipo de investigación

La investigación aplicada busca la generación de conocimiento con aplicación directa a los problemas de la sociedad o el sector productivo. Esta se basa fundamentalmente en los hallazgos tecnológicos de la investigación básica, ocupándose del proceso de enlace entre la teoría y el producto. El proyecto presenta una visión sobre los pasos a seguir en el desarrollo de investigación aplicada, la importancia de la colaboración entre la universidad y la industria en el proceso de transferencia de tecnología, así como los aspectos relacionados a la protección de la propiedad intelectual durante este proceso. [6]

Por otra parte, Padrón 2006, dice que la investigación aplicada recibió el nombre de "investigación práctica o empírica", y se caracterizaba porque buscaba la aplicación o utilización de los conocimientos adquiridos, y a la vez que se pueden adquirir otros, después de implementar y sistematizar la práctica basada en investigación. [15]

Finalmente, el proceso investigativo de maduración y transferencia de la tecnología permite la creación de prototipos que materializan el concepto y que se pueden transferir a la industria para que se transformen en productos.

4.3. Planificación

En esta sección se describe la metodología utilizada.

4.3.1. Definición de la metodología

Para la formulación y desarrollo de este proyecto se determina trabajar con la metodología de investigación entregado por Vaishnavi y Kuechler en el libro (Métodos y patrones de investigación en ciencias del diseño: Innovadora tecnología de la información y la comunicación, 2ª edición). En este Señala que las escuelas profesionales anhelaban de las científicas, la «respetabilidad académica» y dice que "la respetabilidad académica requiere que un tema que sea intelectualmente robusto, analítico, formalizable y enseñable". Sin embargo, reconoce que "En el pasado, mucho, si no la mayoría, de lo que se sabia sobre el diseño y sobre las ciencias artificiales era intelectualmente débil, intuitivo, informal y poco reflexivo".

Con base en este razonamiento, propone el desarrollo de una enseñanza profesional que pudiera alcanzar simultáneamente dos objetivos: la enseñanza en las ciencias de lo artificial como en las ciencias de lo natural a un nivel intelectual alto. "Las escuelas profesionales, pueden volver a asumir sus responsabilidades profesionales sólo en la medida en que descubran y enseñen una ciencia del diseño, un cuerpo de de pensamiento intelectualmente robusto, analítico, parcialmente formalizable, en parte empírico, una doctrina enseñable sobre el proceso de diseño".[11]

En una forma mas resumida, la metodología consiste en una serie de pasos que permitirá llevar a cabo el proyecto de investigación que se propone. Se inicia con la identificación del problema, la propuesta de investigación, el desarrollo y la implementación del diseño propuesto, la evaluación de los artefactos (medidas de desempeño), la presentación de resultados y las conclusiones.

4.3.2. Desarrollo de la metodología

Se identificara y se seleccionara la variables con las cuales se realizara el estudio de las técnicas de aprendizaje automático supervisado; con el conjunto de datos seleccionados se analizará y se implementará la metodología a utilizar y a su vez se cargara la base de datos con la información a estudiar. Se procederá a la construcción de algoritmos que permitirá la generación de datos o resultados a analizar, utilizando como herramienta la plataforma, microsoft azune. Posteriormente se realizaran pruebas en las que se verificaran los resultados que se obtuvieron al momento de operar los algoritmos. Se procederá analizar los datos obtenidos y se generara comparativo de los mismos, teniendo en cuenta las técnicas implementadas.

5. Desarrollo de la solución

En esta sección se definen los métodos de recolección de la información, los datos, a aplicación de técnicas de aprendizaje automático supervisado, los resultados y comparación de los mismos, los cuales se desarrollaran en el proyecto.

5.1. Recolección de la información

Se realizaron llamadas telefónicas al IDEAM, en las cuales se solicito información con respecto a las estaciones ubicadas en Cartagena y el tiempo en que estas están en funcionamiento; a lo cual indicaron que mediante la pagina Web se debía realizar un trámite de solicitud de información, en la cual se identifica la estación hidrológica, los datos y el periodo de tiempo. Se consulto en la página del IDEAM las estaciones hidrológicas en Cartagena que se encuentran en estado Activo.

AREA OPER.	CODIGO CAT.	NOMBRE	CLASE	CORRIENTE	LATITUD	LONGITUD	ALTITUD	FECHA INST.
AREA OPERATIVA 02	29037610	KILOMETRO 107	HID	CANAL DEL DIQUE	10.22805556	-75.52305556	2	15/04/1981

Figura 5-1.: Estaciones en estado activo

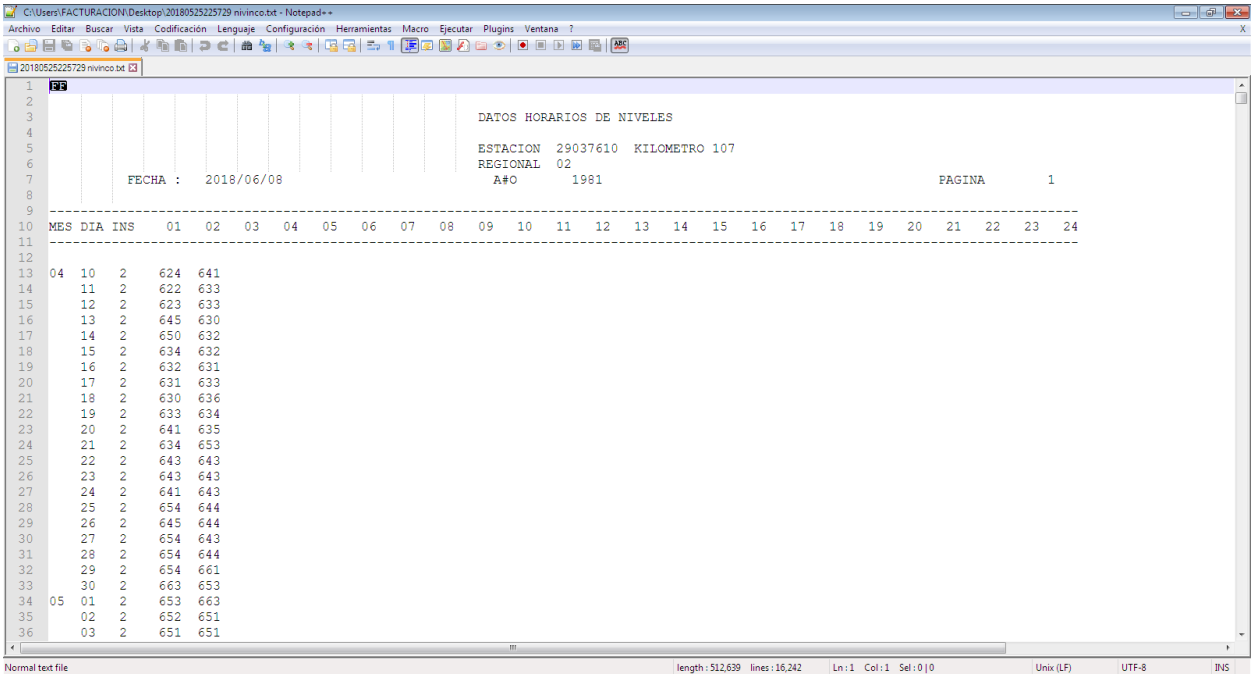
Por medio de la página del IDEAM se realizo solicitud de la información de la estación anteriormente mencionada.

Código	Nombre	Corriente	Elevación	Municipio	Tipo
4743	KILOMETRO 107 [29037610]	CANAL DEL DIQUE	2.0		Limnimetrica

	Parámetro	Periodicidad	Desde		Hasta	
☑	Caudales máximos (m3/seg)	Diario	Enero	2012	Abril	2018
☑	Caudales Medios (m3/seg)	Diario	Enero	2012	Abril	2018
☑	Caudales mínimos (m3/seg)	Diario	Enero	2012	Abril	2018
☑	Concentración sedimientos máximos	Diario	Enero	2012	Abril	2018
☑	Concentración sedimientos medios	Diario	Enero	2012	Abril	2018
☑	Concentración sedimientos mínimos	Diario	Enero	2012	Abril	2018
☑	Granulometría	Diario	Enero	2012	Abril	2018
☑	Nivel del mar - alturas horarias (anual) 12 datos	Diario	Enero	2012	Abril	2018
☑	Nivel del mar - alturas horarias (mensual) entre 672 a 744 datos	Diario	Enero	2012	Abril	2018
☑	Nivel del mar - horas y alturas (mensual) 4 datos/ día por número días mes	Diario	Enero	2012	Abril	2018
☑	Niveles horarios de nivel (nivinco lecturas de mira)	Diario	Enero	2012	Abril	2018
☑	Niveles máximos (cm)	Diario	Enero	2012	Abril	2018
☑	Niveles medios (cm)	Diario	Enero	2012	Abril	2018
☑	Niveles mínimos (cm)	Diario	Enero	2012	Abril	2018
☑	Perfil transversal	Diario	Enero	2012	Abril	2018
☑	Resumen aforos líquidos	Diario	Enero	2012	Abril	2018
☑	Resumen aforos sólidos	Diario	Enero	2012	Abril	2018
☑	Tabla conversión nivel caudal (curva de gastos)	Diario	Enero	2012	Abril	2018
☑	Transp. sedimentos máximos	Diario	Enero	2012	Abril	2018
☑	Transp. sedimentos medios	Diario	Enero	2012	Abril	2018
☑	Transp. sedimentos totales	Diario	Enero	2012	Abril	2018

Figura 5-2.: Variables solicitadas por cada estación

El instituto de hidrología, meteorología y estudios ambientales (IDEAM) suministra la información de 2 valores o tomas de los niveles indicados mediante la mira hidrométrica o limnímetro desde el día 10/04/1981 hasta el día 31/08/2016.



```
C:\Users\FACTURACION\Desktop\20180525225729 nivinco.txt - Notepad++
Archivo Editar Buscar Vista Codificación Lenguaje Configuración Herramientas Macro Ejecutar Plugins Ventana ?
20180525225729 nivinco.txt

                                        DATOS HORARIOS DE NIVELES

                                        ESTACION  29037610  KILOMETRO 107
                                        REGIONAL  02
          FECHA :   2018/06/08            A#O       1981                                        PAGINA       1

-------------------------------------------------------------------------------------------------------------------------
MES DIA INS    01   02   03   04   05   06   07   08   09   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24
-------------------------------------------------------------------------------------------------------------------------

04  10   2    624  641
    11   2    622  633
    12   2    623  633
    13   2    645  630
    14   2    650  632
    15   2    634  632
    16   2    632  631
    17   2    631  633
    18   2    630  636
    19   2    633  634
    20   2    641  635
    21   2    634  653
    22   2    643  643
    23   2    643  643
    24   2    641  643
    25   2    654  644
    26   2    645  644
    27   2    654  643
    28   2    654  644
    29   2    654  661
    30   2    663  653
05  01   2    653  663
    02   2    652  651
    03   2    651  651
Normal text file  length : 512,639  lines : 16,242  Ln : 1  Col : 1  Sel : 0 | 0  Unix (LF)  UTF-8  INS
```

Figura 5-3.: niveles indicados mediante la mira hidrométrica o limnímetro

Niveles: Se denomina nivel del agua en una corriente (río, quebrada, arroyo, caño) o en un cuerpo de agua (ciénaga, lago, laguna, embalse), a la elevación o altura de la superficie del agua en un punto determinado, el cual está ligado topográficamente a un origen de referencia identificado con una cota arbitraria o al nivel medio del mar.

Toma de Muestra: La información de niveles se puede obtener por observaciones directas en forma sistemática de una manera relativamente fácil en corrientes (ríos, quebradas, arroyos) y cuerpos de agua (embalses, lagunas). En Colombia y por recomendación y estandarización mundial de la organización Meteorológica Mundial (OMM), en las estaciones hidrométricas se toma información diaria, realizando dos lecturas, a las 6 am y 6 pm; esto se realiza mediante la mira hidrométrica o limnímetro es una regla graduada dispuesta en tramos de (1) metro, que se utiliza para medir las fluctuaciones de los niveles del agua en un punto determinado de una corriente o de un cuerpo de agua. En cuanto a su funcionamiento, las miras hidrométricas directas se instalan sobre la orilla próxima al sector más profundo del cauce, cuidando que la cota cero quede 0.5 metros por debajo del fondo del cauce para ríos pequeños y 0.5 metros por debajo del nivel de aguas mínimas, en ríos grandes. El extremo superior del limnímetro debe sobrepasar por lo menos en un metro el nivel máximo de la creciente posible o la registrada históricamente según huellas y/o información de los habitantes de la región.

Adicional a los Niveles, se solicito información con respecto a los períodos en los cuales de presento el fenómeno del niño, de la niña o periodos neutros en el periodo de 10/04/1981 hasta el 31/08/2016; esto se realizo mediante una PQR en la página web del IDEAM con el radicado

No 20184000009661 y teniendo en cuenta los Boletines informativos en la misma.

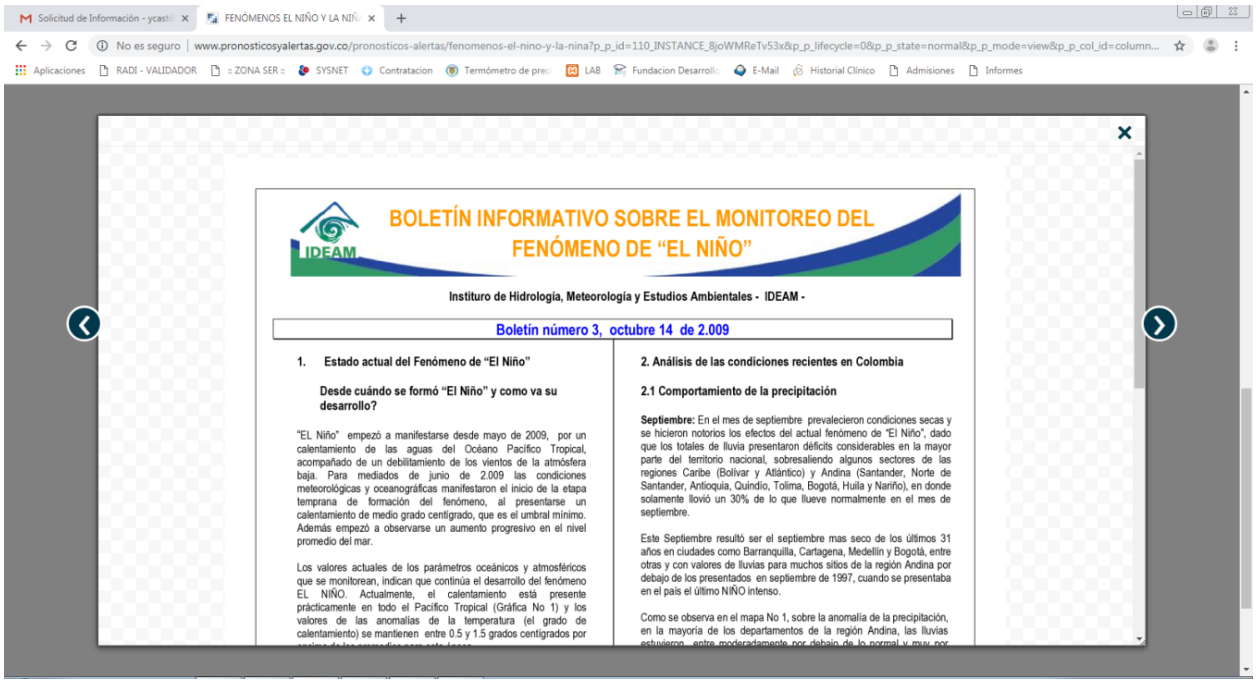


Figura 5-4.: períodos de presencia del fenómeno de la niña, niño o periodos neutros

Como respuesta se obtienen las fechas por días y periodos en que se presentaron los fenómenos, desde 10/04/1981 hasta el 31/08/2016. Esta información fue suministrada a través de boletines informativos mensuales, en los cuales se detalla el comportamiento océano-atmosférico (ver figura 5-4).

5.2. Construcción del conjunto de datos

Teniendo en cuenta la información recopilada se organiza la información de la siguiente forma:

Año	Mes	Dia	Nivel1	Nivel2	Fenomeno
1981	04	10	624	641	0
1981	04	11	622	633	0
1981	04	12	623	633	0
1981	04	13	645	630	0
1981	04	14	650	632	0
1981	04	15	634	632	0
1981	04	16	632	631	0
1981	04	17	631	633	0
1981	04	18	630	636	0
1981	04	19	633	634	0
1981	04	20	641	635	0
1981	04	21	634	653	0
1981	04	22	643	643	0
1981	04	23	643	643	0

Figura 5-5.: Conjunto de datos utilizados

- Año: Año en el cual se toma la muestra.
- Mes: Mes en el cual se toma la muestra..
- Día: Día en el cual se toma la muestra.
- Nivel1: Valor de la muestra o nivel (6am).
- Nivel2: Valor de la muestra o nivel (6pm).
- Fenómeno: Indica la presencia (1) o ausencia (0) de los fenómenos del niño o niña.

Se toma el valor cero (0) para todos los datos en conjunto Niño y Niña para los casos en que no se tenga dato o información de los mismos.

5.3. Aplicación de las técnicas de aprendizaje supervisado

En este capítulo se describirán las 3 técnicas utilizadas en la investigación y estudio de los datos hidrológicos en Cartagena desde 1981 hasta 2016; además de la comparación de las técnicas, para lo cual se determino que se evaluaran con base en la misma estructura y componentes, esto con el fin de que las técnicas no tengas variables diversas a la lógica y estructura de las mismas.

5.3.1. Regresión logística

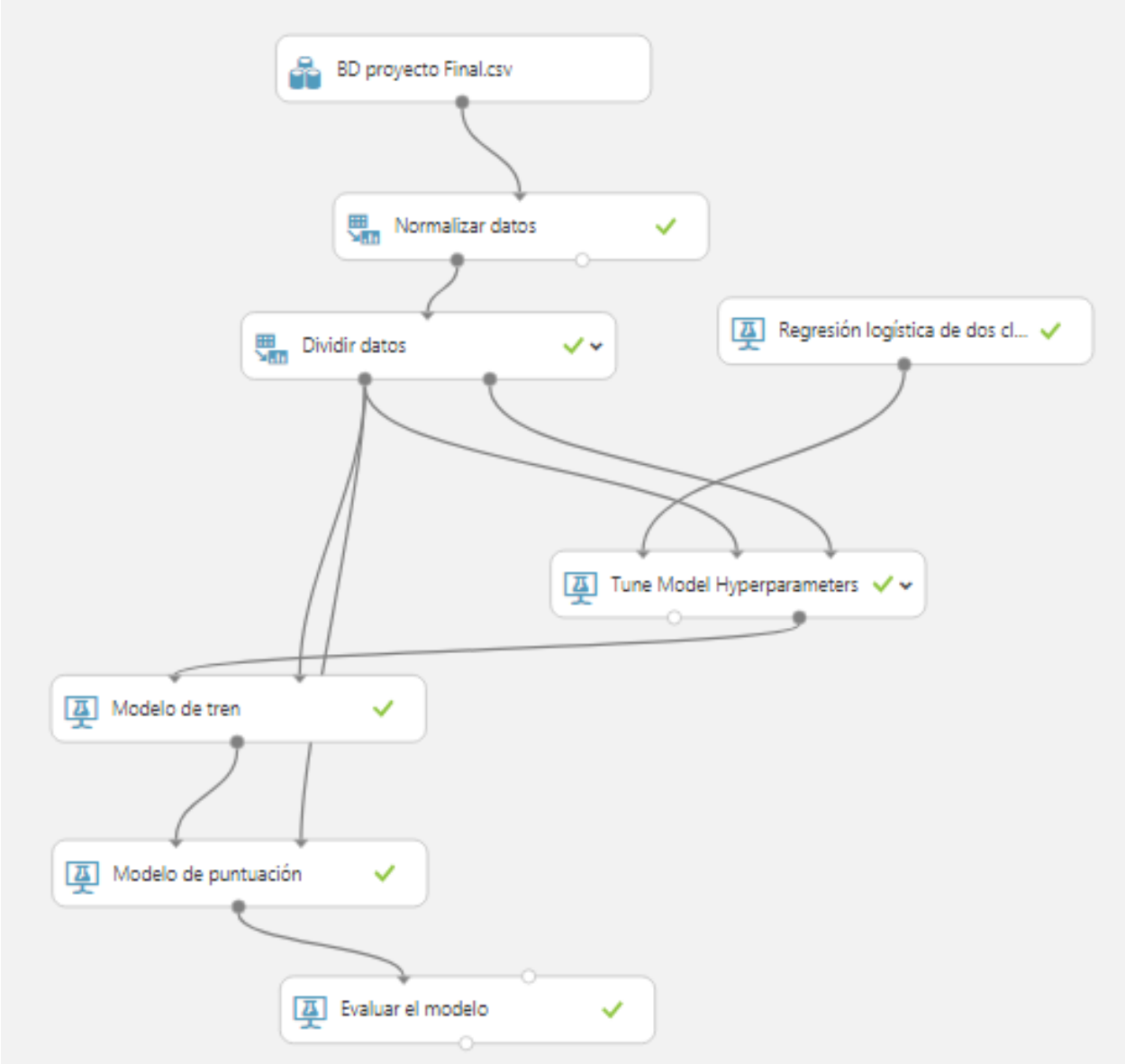


Figura 5-6.: Técnica de regresión logística

Como bien se sabe, la regresión logística es una técnica que se utiliza para el modelamiento de diversos problemas, en el caso puntual de proyecto se utilizan datos hidrológicos; este algoritmo es un método de aprendizaje supervisado. Se arma el modelo de la siguiente forma:

1. Se proporcionan un conjunto de datos que contienen los resultados y con ello capacitar al modelo; para la ejecución del estudio e investigación del mismo, lo primero que se realiza es la organización de los datos según la estructura que es solicitada por la herramienta, un archivo CSV separado por comas (,).

2. Se normalizan los datos, lo que consiste en una técnica que es aplicada como parte de la separación de los datos para el aprendizaje automático; teniendo como objetivo cambiar los valores de las columnas numéricas en el conjunto de datos para usar una escala común, sin que se distorsionen las diferencias en los rangos de valores o perder información. El método de transformación que se utiliza es Zscore el cual convierte todos los valores en un z-score. La media y la desviación estándar se calculan para cada columna por separado.

3. Se procede a dividir los datos, lo cual permite la separación de los datos en conjuntos de entrenamiento y prueba; el modo de división utilizado es dividir filas, el cual se utiliza para dividir los datos en 2 partes en 80-20.

4. El modelo o método de entrenamiento de los datos es configurado como parámetro único, se ingresa como tolerancia de optimización 1E-07, peso de regularización L1 y L2 con valores 1 y Tamaño de memoria con valor en 20.

5. El siguiente paso es la utilización de un Tune Model Hyperparameters el cual crea y prueba varios modelos, utilizando diversas combinaciones de configuraciones y compara las métricas de todos los modelos para obteniendo la combinación de configuraciones; el método utilizado para el barrido es cuadricula aleatorio, el cual desde el punto de vista computacional es considerado como más eficiente. Se especifica el número 5 como el máximo de ejecuciones en el barrido aleatorio, el cual indica cuántas veces debe entrenarse el modelo, utilizando una combinación aleatoria de valores de parámetros, se indica en la columna de etiqueta el dato Fenómeno como el valor a evaluar, se utiliza la métrica AUC para la clasificación que representa el área bajo la curva cuando los falsos positivos se trazan en el eje x y los verdaderos positivos se trazan en el eje y. La métrica utilizada para la regresión es el error absoluto medio en el cual se Promedia todo el error en el modelo, donde error significa la distancia del valor predicho del valor verdadero.

6. Se procede con el modelo de tren el cual permite el entrenamiento del modelo de manera supervisad, utilizando el conjunto de datos que contienen datos históricos para aprender patrones (Los datos contienen tanto el resultado (etiqueta) que está tratando de predecir, como los factores relacionados (variables). Ya que el modelo de aprendizaje automático precisa los resultados para determinar las características que mejor predicen los resultados.); se indica como etiqueta el dato fenómeno.

7. Se implementa el modelo de puntuación el cual generar predicciones utilizando una clasificación capacitada ingresando el modelo entrenado y el conjunto de datos obtenidos de Tune Model Hyperparameters.

8. Como paso final se procede a evaluar el modelo el cual permite medir la precisión del modelo capacitado mediante el cálculo de conjunto de métricas de evaluación con métricas estándar.

5.3.2. Resultados de la técnica de regresión logística

A continuación se indica la matriz de resultados arrojada por la técnica aplicada:

Año	Mes	día	nivel1	nivel2	fenómeno	Etiquetas puntuadas	Probabilidades anotadas
-1.524438	-0.442454	1.396148	-2.155806	-1.719531	0.878788	-1.13793	0.47289
-0.151368	0.722207	-0.877539	-1.400096	-0.192251	0.878788	0.878788	0.583437
1.319779	1.595702	-0.309117	0.489179	0.741088	-1.13793	0.878788	0.679644
-0.837903	1.595702	1.055095	0.25302	-0.998315	-1.13793	0.878788	0.58489
0.142862	1.595702	1.737201	0.819802	1.589577	0.878788	0.878788	0.628894
-1.524438	0.431042	1.737201	-1.49456	-1.125589	-1.13793	0.878788	0.507122
-0.151368	-0.442454	0.941411	-0.975009	-0.65892	0.878788	0.878788	0.537932
1.712085	-1.607115	-1.104908	0.25302	-0.828618	0.878788	0.878788	0.580435
1.221702	0.722207	-1.67333	1.103194	1.33503	0.878788	0.878788	0.645379
-1.524438	-1.024784	1.623517	-2.297502	-2.228625	0.878788	-1.13793	0.450089
-1.328285	1.013372	0.486673	-0.927777	0.104721	-1.13793	0.878788	0.539647
0.73132	-1.31595	1.16878	0.01686	-0.828618	0.878788	0.878788	0.545358
0.829397	-0.442454	1.396148	1.197657	0.995634	-1.13793	0.878788	0.583699
0.829397	-1.024784	0.14562	-0.786082	-1.083164	-1.13793	0.878788	0.561429
0.927473	1.304537	-1.218592	2.142295	1.759275	-1.13793	0.878788	0.653554
1.221702	-1.024784	1.737201	0.583643	0.019872	0.878788	0.878788	0.579276
1.614008	1.013372	1.16878	0.394715	0.444116	0.878788	0.878788	0.671346
-1.426361	0.139876	-0.309117	-0.408227	-0.616495	-1.13793	0.878788	0.501189
0.142862	1.595702	0.714042	1.292121	1.589577	0.878788	0.878788	0.629284
0.927473	-0.442454	0.941411	0.347484	0.95321	0.878788	0.878788	0.588201
1.319779	1.595702	-0.195433	0.441947	0.486541	-1.13793	0.878788	0.679605

Figura 5-7.: Tabla de resultados, técnica de regresión logística

Se detalla la tabla estadística,

Estadística	
Mean	0.5617
Median	0.5626
Min	0.4233
Max	0.6921
Standard Deviation	0.059
Unique Values	10266
Missing Values	0
Feature Type	Numeric Score

Figura 5-8.: Tabla estadísticas detallada, técnica de regresión logística

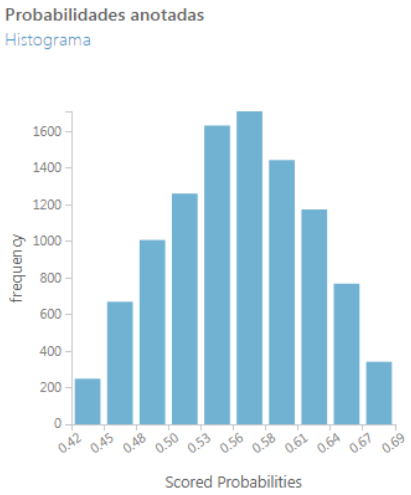


Figura 5-9.: Histograma de Probabilidades anotadas, técnica de regresión logística

5.3.3. Evaluación del modelo de regresión logística

En la siguiente gráfica se tiene como referencia la precisión de la técnica de regresión logística.

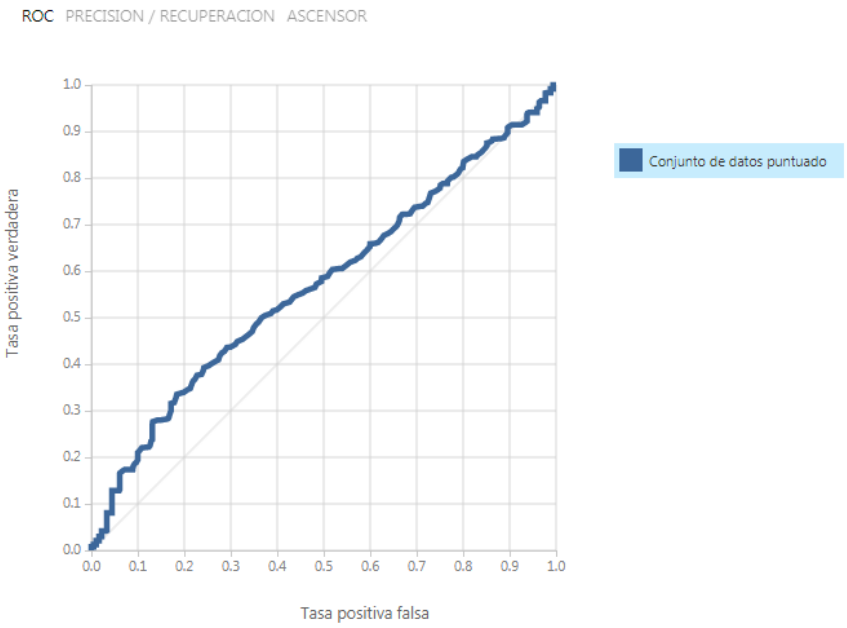


Figura 5-10.: Precisión de la técnica de regresión logística

Tabla de puntuación	Ejemplos positivos	Ejemplos negativos	Fracción por encima del umbral	Exactitud	Puntuación F1	Precisión	Recordar	Precisión negativa	Recuerdo negativo	AUC acumulativa
(0.900,1.000]	0	0	0	0.436	0	1	0	0.436	1	0
(0.800,0.900]	0	0	0	0.436	0	1	0	0.436	1	0
(0.700,0.800]	0	0	0	0.436	0	1	0	0.436	1	0
(0.600,0.700]	1965	892	0.278	0.541	0.455	0.688	0.34	0.484	0.801	0.038
(0.500,0.600]	2907	2746	0.829	0.557	0.682	0.573	0.842	0.479	0.188	0.399
(0.400,0.500]	914	842	1	0.564	0.721	0.564	1	1	0	0.571
(0.300,0.400]	0	0	1	0.564	0.721	0.564	1	1	0	0.571
(0.200,0.300]	0	0	1	0.564	0.721	0.564	1	1	0	0.571
(0.100,0.200]	0	0	1	0.564	0.721	0.564	1	1	0	0.571
(0.000,0.100]	0	0	1	0.564	0.721	0.564	1	1	0	0.571

Figura 5-11.: Tabla de variables generadas en la evaluación de la técnica de regresión logística

En la siguiente tabla se define cada variable de medición, donde el verdadero positivo es la proporción de casos positivos que están bien detectadas por la prueba.
El falso positivo, es la proporción de casos negativos que una prueba detecta como positivos.
El falso negativo, es la proporción de casos positivos que una prueba detecta como negativo.
El verdadero negativo, es la proporción de casos negativos que son bien detectados en una prueba.
Exactitud, es el sesgo de la estimación de casos negativos que sin bien detectados en una prueba.
El recall (sensibilidad), es la fracción de instancias relevantes que han sido recuperadas.
La precisión, siendo esta la dispersión del conjunto de valores obtenidos de mediciones repetidas en una magnitud.

Técnica	Verdadero Positivo	Falso Positivo	Falso Negativo	Verdadero Negativo	Exactitud	Recall (Sensibilidad)	Precisión
Regresión logística	4872	3638	914	842	0.557	0.842	0.573

Figura 5-12.: Tabla matriz de confusión, técnica de regresión logística

Teniendo en cuenta los resultados obtenidos se evidencia que la proporción de casos positivos que están bien detectadas por la técnica es de 4872, los negativos que la técnica detecta como positivos 3638, positivos que la técnica detecta como negativo 914, los negativos que son bien detectados en la técnica de 842 y teniendo en cuenta las mediciones del sesgo de la estimación o cuan cerca el valor real se encuentra el valor medio tiene exactitud de 0.557, al igual que la sensibilidad o fracción de instancias relevantes que han sido recuperadas de 0.842 y el 0.573 de dispersión del conjunto de valores obtenidos de mediciones repetidas en la magnitud o precisión; teniendo en cuenta lo anterior se evidencia que los resultados son óptimos al aplicar la técnica a datos hidrológicos que evalúan los fenómenos presentados.

5.3.4. Redes Neuronales

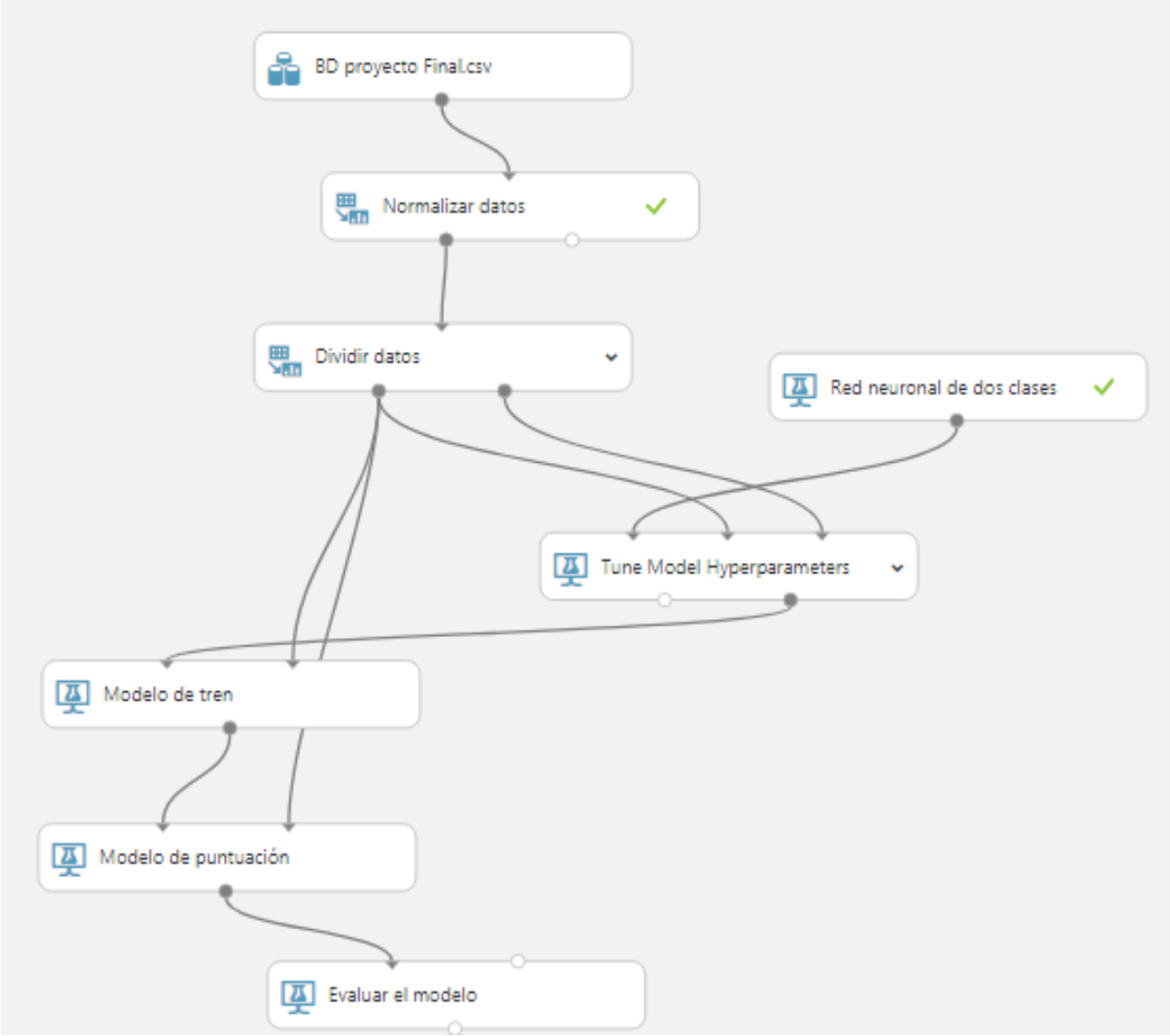


Figura 5-13.: Técnica de redes neuronales

Las redes neuronales son un método de aprendizaje supervisado que requiere un conjunto de datos etiquetados, que incluye una columna de etiqueta; entre las capas de entrada y salida puede insertar múltiples capas ocultas. La mayoría de las tareas predictivas se pueden realizar fácilmente con solo una o algunas capas ocultas; Se parametriza el modelo de la siguiente forma:

1. Se proporcionan un conjunto de datos que contienen los resultados y con ello capacitar al modelo; para la ejecución del estudio e investigación del mismo, lo primero que se realiza es la organización de los datos según la estructura que es solicitada por la herramienta, un archivo CSV separado por comas (,).

2. Se normalizan los datos, lo que consiste en una técnica que es aplicada como parte de la separación de los datos para el aprendizaje automático; teniendo como objetivo cambiar los valores de las columnas numéricas en el conjunto de datos para usar una escala común, sin que se distorsionen las diferencias en los rangos de valores o perder información. El método de transformación que se utiliza es Zscore el cual convierte todos los valores en un z-score. La media y la desviación estándar se calculan para cada columna por separado.

3. Se procede a dividir los datos, lo cual permite la separación de los datos en conjuntos de entrenamiento y prueba; el modo de división utilizado es dividir filas, el cual se utiliza para dividir los datos en 2 partes en 80-20.

4. El modelo o método de entrenamiento de los datos es configurado como parámetro único, se ingresa como tolerancia de optimización 1E-07, peso de regularización L1 y L2 con valores 1 y Tamaño de memoria con valor en 20.

5. El siguiente paso es la utilización de un Tune Model Hyperparameters el cual crea y prueba varios modelos, utilizando diversas combinaciones de configuraciones y compara las métricas de todos los modelos para obteniendo la combinación de configuraciones; el método utilizado para el barrido es cuadricula aleatorio, el cual desde el punto de vista computacional es considerado como más eficiente. Se especifica el número 5 como el máximo de ejecuciones en el barrido aleatorio, el cual indica cuántas veces debe entrenarse el modelo, utilizando una combinación aleatoria de valores de parámetros, se indica en la columna de etiqueta el dato Fenómeno como el valor a evaluar, se utiliza la métrica AUC para la clasificación que representa el área bajo la curva cuando los falsos positivos se trazan en el eje x y los verdaderos positivos se trazan en el eje y. La métrica utilizada para la regresión es el error absoluto medio en el cual se Promedia todo el error en el modelo, donde error significa la distancia del valor predicho del valor verdadero.

6. Se procede con el modelo de tren el cual permite el entrenamiento del modelo de manera supervisad, utilizando el conjunto de datos que contienen datos históricos para aprender patrones (Los datos contienen tanto el resultado (etiqueta) que está tratando de predecir, como los factores relacionados (variables). Ya que el modelo de aprendizaje automático precisa los resultados para determinar las características que mejor predicen los resultados.); se indica como etiqueta el dato fenómeno.

7. Se implementa el modelo de puntuación el cual generar predicciones utilizando una clasificación capacitada ingresando el modelo entrenado y el conjunto de datos obtenidos de Tune Model Hyperparameters.

8. Como paso final se procede a evaluar el modelo el cual permite medir la precisión del modelo capacitado mediante el cálculo de conjunto de métricas de evaluación con métricas estándar.

5.3.5. Resultados de la técnica de redes neuronales

A continuación se indica la matriz de resultados arrojada por la técnica aplicada:

Año	Mes	día	nivel1	nivel2	fenómeno	Etiquetas puntuadas	Probabilidades anotadas
-1.52444	-0.44245	1.396148	-2.15581	-1.71953	0.878788	-1.13793	0.43713
-0.15137	0.722207	-0.87754	-1.4001	-0.19225	0.878788	0.878788	0.579203
1.319779	1.595702	-0.30912	0.489179	0.741088	-1.13793	0.878788	0.614622
-0.8379	1.595702	1.055095	0.25302	-0.99832	-1.13793	0.878788	0.601222
0.142862	1.595702	1.737201	0.819802	1.589577	0.878788	0.878788	0.593708
-1.52444	0.431042	1.737201	-1.49456	-1.12559	-1.13793	-1.13793	0.375988
-0.15137	-0.44245	0.941411	-0.97501	-0.65892	0.878788	-1.13793	0.321566
1.712085	-1.60712	-1.10491	0.25302	-0.82862	0.878788	0.878788	0.739932
1.221702	0.722207	-1.67333	1.103194	1.33503	0.878788	0.878788	0.718196
-1.52444	-1.02478	1.623517	-2.2975	-2.22863	0.878788	0.878788	0.542988
-1.32829	1.013372	0.486673	-0.92778	0.104721	-1.13793	0.878788	0.595876
0.73132	-1.31595	1.16878	0.01686	-0.82862	0.878788	0.878788	0.551454
0.829397	-0.44245	1.396148	1.197657	0.995634	-1.13793	0.878788	0.503706
0.829397	-1.02478	0.14562	-0.78608	-1.08316	-1.13793	-1.13793	0.433045
0.927473	1.304537	-1.21859	2.142295	1.759275	-1.13793	0.878788	0.680265
1.221702	-1.02478	1.737201	0.583643	0.019872	0.878788	0.878788	0.633369
1.614008	1.013372	1.16878	0.394715	0.444116	0.878788	0.878788	0.691322
-1.42636	0.139876	-0.30912	-0.40823	-0.6165	-1.13793	-1.13793	0.356877
0.142862	1.595702	0.714042	1.292121	1.589577	0.878788	0.878788	0.601462
0.927473	-0.44245	0.941411	0.347484	0.95321	0.878788	-1.13793	0.456357
1.319779	1.595702	-0.19543	0.441947	0.486541	-1.13793	0.878788	0.611008
-1.32829	0.431042	-0.08175	-1.63626	-1.38014	0.878788	-1.13793	0.388541
0.829397	-0.15129	0.714042	1.433817	1.292606	-1.13793	-1.13793	0.487658
1.712085	-1.60712	0.259305	0.394715	-0.48922	0.878788	0.878788	0.747189
-0.15137	-0.44245	0.031936	-0.69162	-0.53165	0.878788	-1.13793	0.325231
0.829397	0.722207	-0.65017	1.622744	1.462303	0.878788	0.878788	0.68191
1.712085	-0.44245	-0.19543	1.055962	0.231994	-1.13793	0.878788	0.658083
-1.52444	-0.73362	1.623517	-2.01411	-1.76196	0.878788	-1.13793	0.481325
0.535167	-0.73362	1.282464	0.01686	-0.19225	-1.13793	-1.13793	0.401115

Figura 5-14.: Tabla resultados, técnica de redes neuronales

Estadística	
Media	0.5248
Mediana	0.5401
Min.	0.3087
Max	0.8056
Desviación estándar	0.1248
Valores únicos	10256
Valores faltantes	0
Tipo de función	Puntuación Numérica

Figura 5-15.: Tabla estadística detallada, técnica de redes neuronales

Histograma de Probabilidades anotadas,

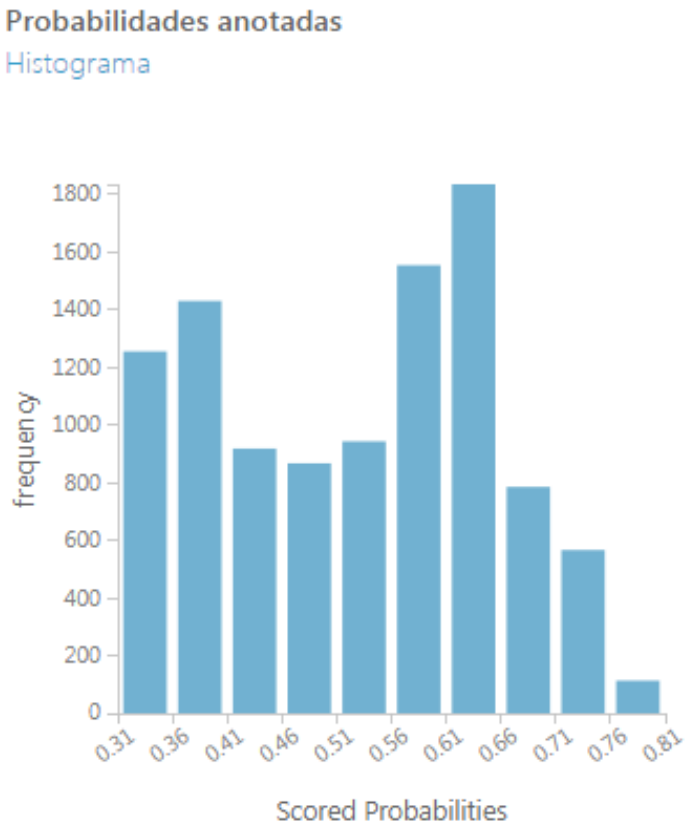


Figura 5-16.: Histograma de Probabilidades anotadas, técnica de redes neuronales

5.3.6. Evaluación del modelo de redes neuronales

Gráfica teniendo como referencia la precisión de la técnica.

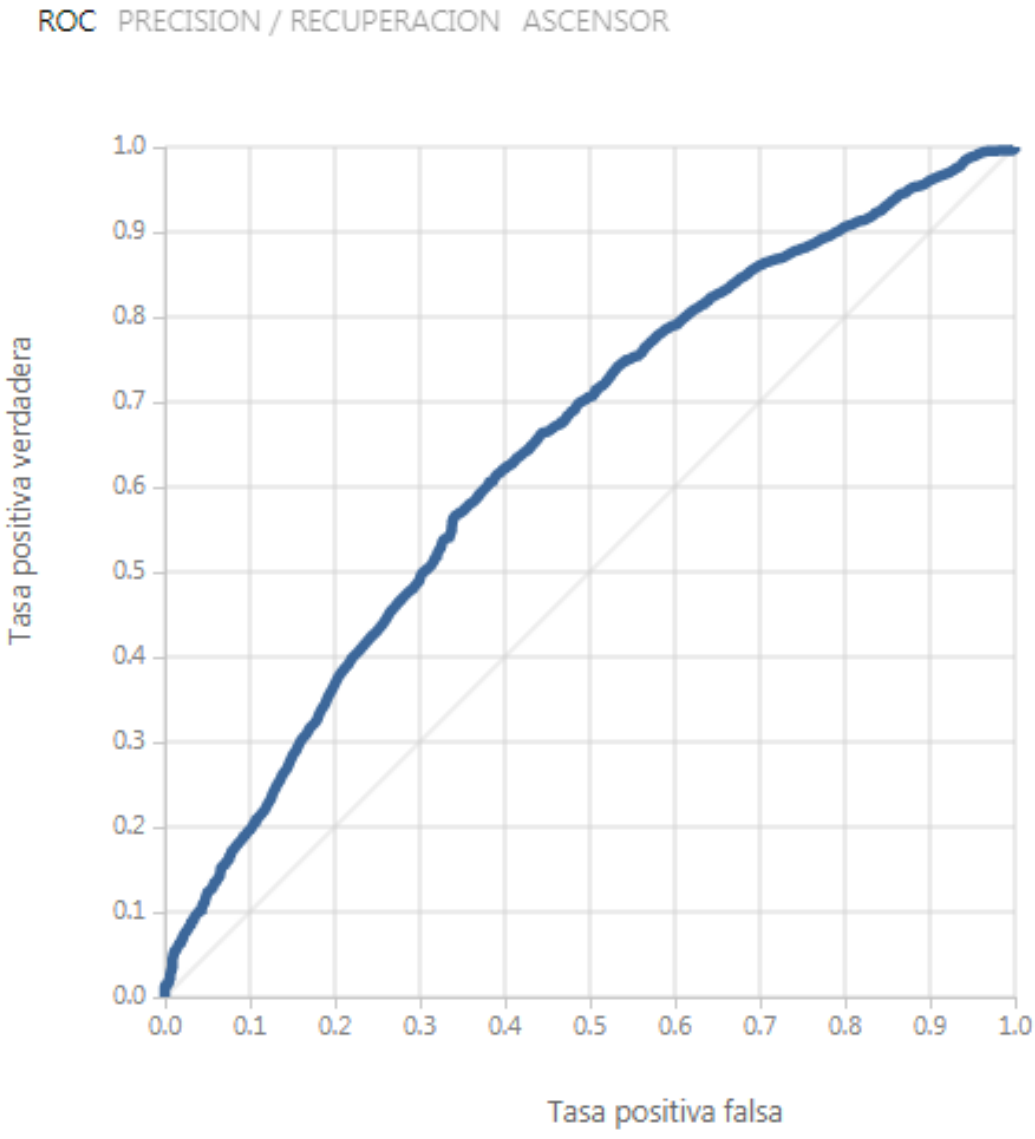


Figura 5-17.: Precisión de la técnica de redes neuronales

Tabla de puntuación	Ejemplos positivos	Ejemplos negativos	Fracción por encima del umbral	Exactitud	Puntuación F1	Precisión	Recordar	Precisión negativa	Recuerdo negativo	AUC acumulativa
(0.900,1.000]	0	0	0	0.436	0	1	0	0.436	1	0
(0.800,0.900]	3	0	0	0.437	0.001	1	0.001	0.437	1	0
(0.700,0.800]	578	180	0.074	0.475	0.177	0.763	0.1	0.452	0.96	0.003
(0.600,0.700]	2041	1009	0.371	0.576	0.546	0.688	0.453	0.51	0.735	0.067
(0.500,0.600]	1260	865	0.578	0.614	0.662	0.654	0.671	0.56	0.542	0.177
(0.400,0.500]	941	915	0.759	0.617	0.71	0.619	0.834	0.611	0.337	0.332
(0.300,0.400]	963	1511	1	0.564	0.721	0.564	1	1	0	0.643
(0.200,0.300]	0	0	1	0.564	0.721	0.564	1	1	0	0.643
(0.100,0.200]	0	0	1	0.564	0.721	0.564	1	1	0	0.643
(0.000,0.100]	0	0	1	0.564	0.721	0.564	1	1	0	0.643

Figura 5-18.: Variables generadas en la evaluación de la técnica de redes neuronales

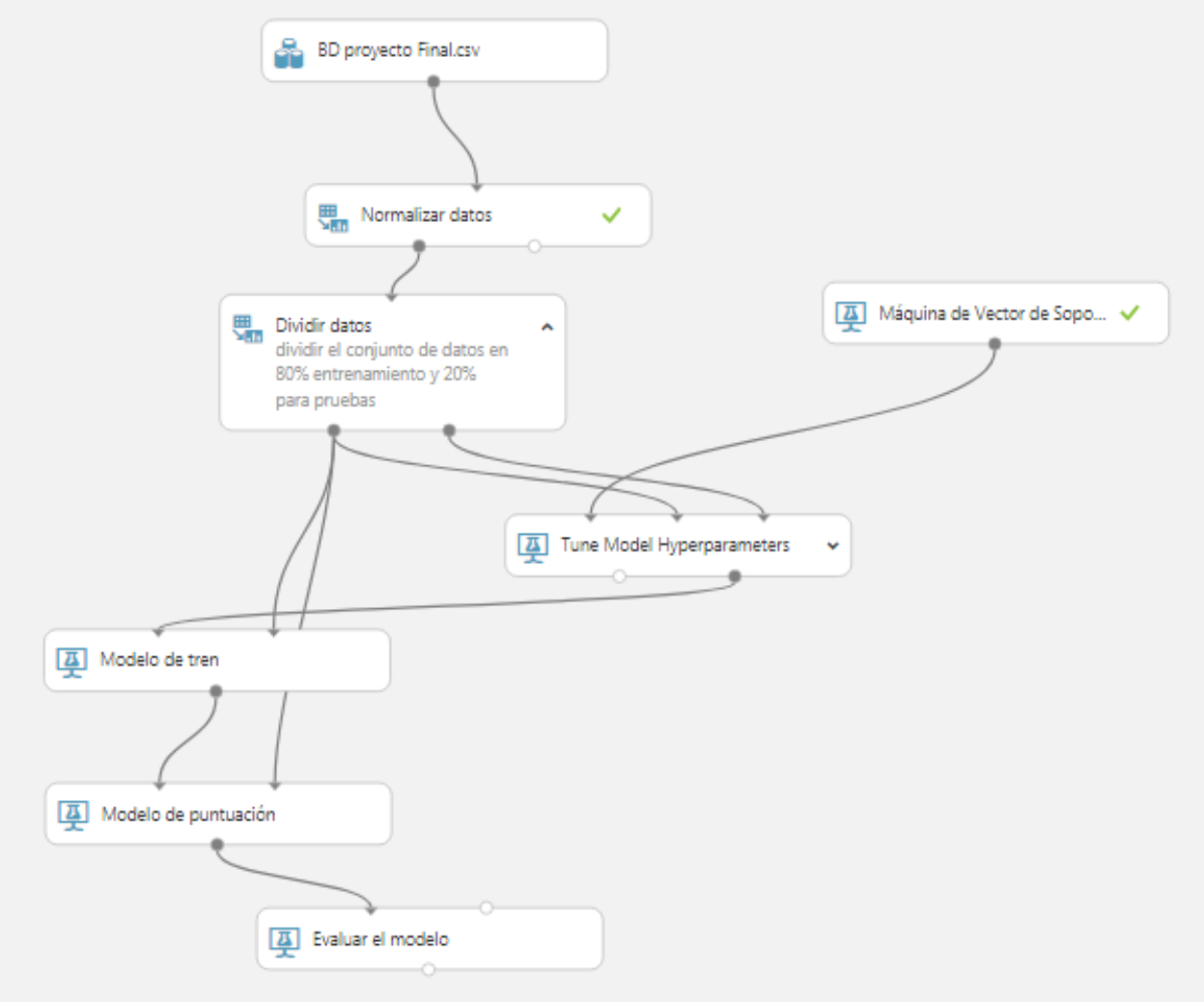
En la siguiente tabla se define cada variable de medición, donde el verdadero positivo es la proporción de casos positivos que están bien detectadas por la prueba.
El falso positivo, es la proporción de casos negativos que una prueba detecta como positivos.
El falso negativo, es la proporción de casos positivos que una prueba detecta como negativo.
El verdadero negativo, es la proporción de casos negativos que son bien detectados en una prueba.
Exactitud, es el sesgo de la estimación de casos negativos que sin bien detectados en una prueba.
El recall (sensibilidad), es la fracción de instancias relevantes que han sido recuperadas.
La precisión, siendo esta la dispersión del conjunto de valores obtenidos de mediciones repetidas en una magnitud.

Técnica	Verdadero Positivo	Falso Positivo	Falso Negativo	Verdadero Negativo	Exactitud	Recall (Sensibilidad)	Precisión
Redes Neuronales	3882	2054	1904	2426	0.614	0.671	0.654

Figura 5-19.: Tabla matriz de confusión, técnica de redes neuronales

Teniendo en cuenta los resultados obtenidos se evidencia que la proporción de casos positivos que están bien detectadas por la técnica es de 3882, los negativos que la técnica detecta como positivos 2054, positivos que la técnica detecta como negativo 1904, los negativos que son bien detectados en la técnica de 2426 y teniendo en cuenta las mediciones del sesgo de la estimación o cuan cerca el valor real se encuentra el valor medio tiene exactitud de 0.614, al igual que la sensibilidad o fracción de instancias relevantes que han sido recuperadas de 0.671 y el 0.654 de dispersación del conjunto de valores obtenidos de mediciones repetidas en la magnitud o precisión; teniendo en cuenta lo anterior se evidencia que los resultados son aceptables al aplicar la técnica a datos hidrológicos que evalúan los fenómenos presentados.

5.3.7. Maquinas De Soporte Vectorial



También conocidas por las siglas SVM, estas son una clase de métodos de aprendizaje supervisados, el cual se encuentra entre los primeros algoritmos de aprendizaje automático el cual requiere de datos etiquetados. En el proceso de capacitación, el algoritmo analiza los datos de entrada y reconoce los patrones en un espacio de funciones multidimensionales denominado hiperplano; para la paarametrizacion del modelo se siguen los siguientes pasos:

1. Se proporcionan un conjunto de datos que contienen los resultados y con ello capacitar al modelo; para la ejecución del estudio e investigación del mismo, lo primero que se realiza es la organización de los datos según la estructura que es solicitada por la herramienta, un archivo CSV separado por comas (,).

2. Se normalizan los datos, lo que consiste en una técnica que es aplicada como parte de la separación de los datos para el aprendizaje automático; teniendo como objetivo cambiar los

valores de las columnas numéricas en el conjunto de datos para usar una escala común, sin que se distorsionen las diferencias en los rangos de valores o perder información. El método de transformación que se utiliza es Zscore el cual convierte todos los valores en un z-score. La media y la desviación estándar se calculan para cada columna por separado.

3. Se procede a dividir los datos, lo cual permite la separación de los datos en conjuntos de entrenamiento y prueba; el modo de división utilizado es dividir filas, el cual se utiliza para dividir los datos en 2 partes en 80-20.

4. El modelo o método de entrenamiento de los datos es configurado como parámetro único, se ingresa como tolerancia de optimización 1E-07, peso de regularización L1 y L2 con valores 1 y Tamaño de memoria con valor en 20.

5. El siguiente paso es la utilización de un Tune Model Hyperparameters el cual crea y prueba varios modelos, utilizando diversas combinaciones de configuraciones y compara las métricas de todos los modelos para obteniendo la combinación de configuraciones; el método utilizado para el barrido es cuadricula aleatorio, el cual desde el punto de vista computacional es considerado como más eficiente. Se especifica el número 5 como el máximo de ejecuciones en el barrido aleatorio, el cual indica cuántas veces debe entrenarse el modelo, utilizando una combinación aleatoria de valores de parámetros, se indica en la columna de etiqueta el dato Fenómeno como el valor a evaluar, se utiliza la métrica AUC para la clasificación que representa el área bajo la curva cuando los falsos positivos se trazan en el eje x y los verdaderos positivos se trazan en el eje y. La métrica utilizada para la regresión es el error absoluto medio en el cual se Promedia todo el error en el modelo, donde error significa la distancia del valor predicho del valor verdadero.

6. Se procede con el modelo de tren el cual permite el entrenamiento del modelo de manera supervisad, utilizando el conjunto de datos que contienen datos históricos para aprender patrones (Los datos contienen tanto el resultado (etiqueta) que está tratando de predecir, como los factores relacionados (variables). Ya que el modelo de aprendizaje automático precisa los resultados para determinar las características que mejor predicen los resultados.); se indica como etiqueta el dato fenómeno.

7. Se implementa el modelo de puntuación el cual generar predicciones utilizando una clasificación capacitada ingresando el modelo entrenado y el conjunto de datos obtenidos de Tune Model Hyperparameters.

8. Como paso final se procede a evaluar el modelo el cual permite medir la precisión del modelo capacitado mediante el cálculo de conjunto de métricas de evaluación con métricas estándar.

5.3.8. Resultados de la técnica maquinas de soporte vectorial

A continuación se indica la matriz de resultados arrojada por la técnica aplicada.

Año	Mes	día	nivel1	nivel2	fenómeno	Etiquetas puntuadas	Probabilidades anotadas
-1.524438	-0.442454	1.396148	-2.155806	-1.719531	0.878788	-1.13793	0.443894
-0.151368	0.722207	-0.877539	-1.400096	-0.192251	0.878788	0.878788	0.560774
1.319779	1.595702	-0.309117	0.489179	0.741088	-1.13793	0.878788	0.669828
-0.837903	1.595702	1.055095	0.25302	-0.998315	-1.13793	0.878788	0.579517
0.142862	1.595702	1.737201	0.819802	1.589577	0.878788	0.878788	0.617991
-1.524438	0.431042	1.737201	-1.49456	-1.125589	-1.13793	-1.13793	0.479476
-0.151368	-0.442454	0.941411	-0.975009	-0.65892	0.878788	0.878788	0.521918
1.712085	-1.607115	-1.104908	0.25302	-0.828618	0.878788	0.878788	0.597062
1.221702	0.722207	-1.67333	1.103194	1.33503	0.878788	0.878788	0.659882
-1.524438	-1.024784	1.623517	-2.297502	-2.228625	0.878788	-1.13793	0.423201
-1.328285	1.013372	0.486673	-0.927777	0.104721	-1.13793	0.878788	0.521866
0.73132	-1.31595	1.16878	0.01686	-0.828618	0.878788	0.878788	0.54789
0.829397	-0.442454	1.396148	1.197657	0.995634	-1.13793	0.878788	0.594663
0.829397	-1.024784	0.14562	-0.786082	-1.083164	-1.13793	0.878788	0.553015
0.927473	1.304537	-1.218592	2.142295	1.759275	-1.13793	0.878788	0.679151
1.221702	-1.024784	1.737201	0.583643	0.019872	0.878788	0.878788	0.581034
1.614008	1.013372	1.16878	0.394715	0.444116	0.878788	0.878788	0.654351
-1.426361	0.139876	-0.309117	-0.408227	-0.616495	-1.13793	0.878788	0.507803
0.142862	1.595702	0.714042	1.292121	1.589577	0.878788	0.878788	0.632858
0.927473	-0.442454	0.941411	0.347484	0.95321	0.878788	0.878788	0.585994
1.319779	1.595702	-0.195433	0.441947	0.486541	-1.13793	0.878788	0.668822
-1.328285	0.431042	-0.081749	-1.636256	-1.380136	0.878788	-1.13793	0.49752
0.829397	-0.151289	0.714042	1.433817	1.292606	-1.13793	0.878788	0.611348
1.712085	-1.607115	0.259305	0.394715	-0.489222	0.878788	0.878788	0.590038
-0.151368	-0.442454	0.031936	-0.691618	-0.531646	0.878788	0.878788	0.533145
0.829397	0.722207	-0.65017	1.622744	1.462303	0.878788	0.878788	0.648007
1.712085	-0.442454	-0.195433	1.055962	0.231994	-1.13793	0.878788	0.637307
-1.524438	-0.733619	1.623517	-2.014111	-1.761956	0.878788	-1.13793	0.436277

Figura 5-20.: Tabla de resultados, técnica de maquinas de soporte vectorial

Se detalla la tabla estadística,

Estadística	
Media	0.5636
Mediana	0.5645
Min.	0.3646
Max	0.7039
Desviación estándar	0.0613
Valores únicos	10248
Valores faltantes	0
Tipo de función	Puntuación Numérica

Figura 5-21.: Tabla de estadística detallada, técnica de maquinas de soporte vectorial

Histograma de Probabilidades anotadas,

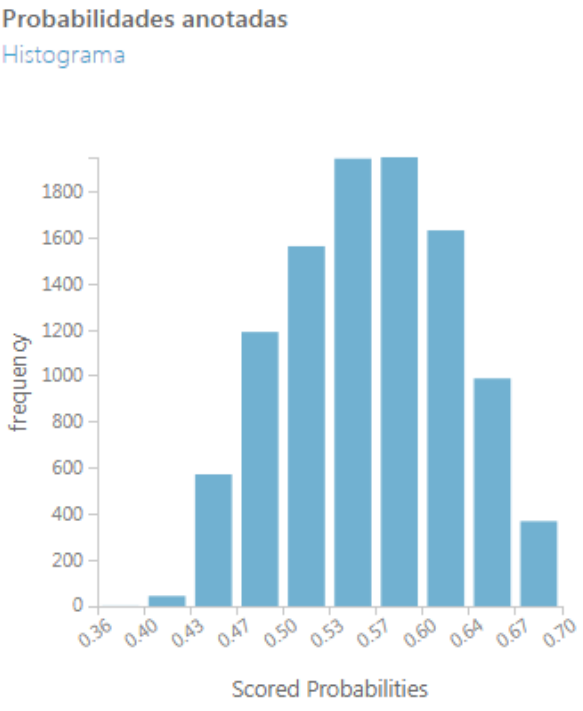


Figura 5-22.: Histograma de Probabilidades anotadas, técnica de maquinas de soporte vectorial

5.3.9. Evaluación del modelo de maquinas de soporte vectorial

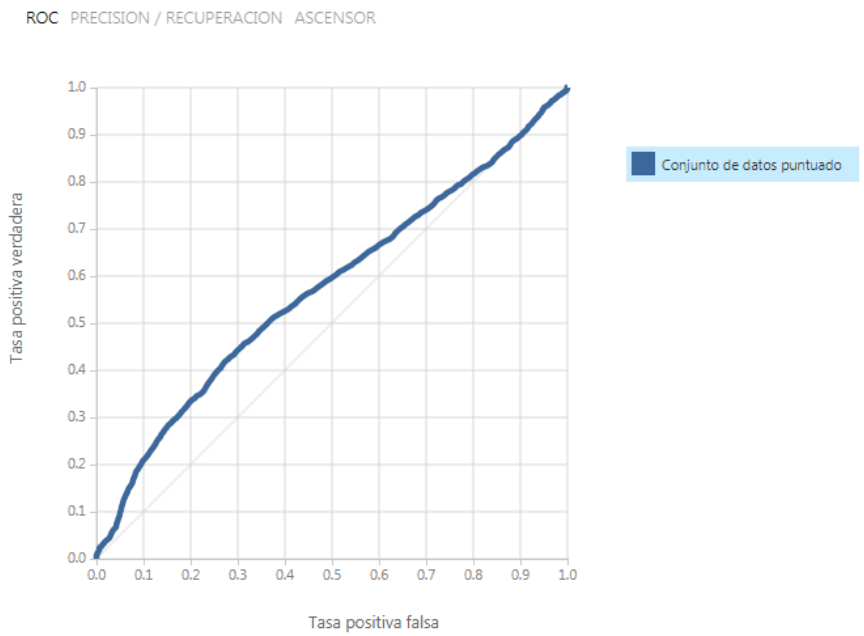


Figura 5-23.: Precisión de la técnica de maquinas de soporte vectorial

Tabla de puntuación	Ejemplos positivos	Ejemplos negativos	Fracción por encima del umbral	Exactitud	Puntuación F1	Precisión	Recordar	Precisión negativa	Recuerdo negativo	AUC acumulativa
(0.900,1.000]	0	0	0	0.436	0	1	0	0.436	1	0
(0.800,0.900]	0	0	0	0.436	0	1	0	0.436	1	0
(0.700,0.800]	6	0	0.001	0.437	0.002	1	0.001	0.437	1	0
(0.600,0.700]	2061	1023	0.301	0.538	0.466	0.669	0.357	0.482	0.772	0.048
(0.500,0.600]	2734	2644	0.825	0.547	0.674	0.567	0.83	0.452	0.181	0.408
(0.400,0.500]	985	812	1	0.564	0.721	0.564	1	1	0	0.573
(0.300,0.400]	0	1	1	0.564	0.721	0.564	1	1	0	0.573
(0.200,0.300]	0	0	1	0.564	0.721	0.564	1	1	0	0.573
(0.100,0.200]	0	0	1	0.564	0.721	0.564	1	1	0	0.573
(0.000,0.100]	0	0	1	0.564	0.721	0.564	1	1	0	0.573

Figura 5-24.: Tabla de variables generadas en la evaluación de la técnica de maquinas de soporte vectorial

Técnica	Verdadero Positivo	Falso Positivo	Falso Negativo	Verdadero Negativo	Exactitud	Recall (Sensibilidad)	Precisión
Máquinas De Soporte Vectorial	4801	3667	985	813	0.547	0.830	0.567

Figura 5-25.: Tabla matriz de confusión, técnica de maquinas de soporte vectorial

Teniendo en cuenta los resultados obtenidos se evidencia que la proporción de casos positivos que están bien detectadas por la técnica es de 4801, los negativos que la técnica detecta como positivos 3667, positivos que la técnica detecta como negativo 985, los negativos que son bien detectados en la técnica de 813 y teniendo en cuenta las mediciones del sesgo de la estimación o cuan cerca el valor real se encuentra el valor medio tiene exactitud de 0.547, al igual que la sensibilidad o fracción de instancias relevantes que han sido recuperadas de 0.830 y el 0.567 de dispersión del conjunto de valores obtenidos de mediciones repetidas en la magnitud o precisión; teniendo en cuenta lo anterior se evidencia que los resultados son buenos al aplicar la técnica a datos hidrológicos que evalúan los fenómenos presentados.

5.4. Comparación de técnicas aplicadas

Se realiza gráficamente la comparación de técnicas, teniendo como referencia la precisión de las técnicas al procesar los datos hidrológicos suministrados.

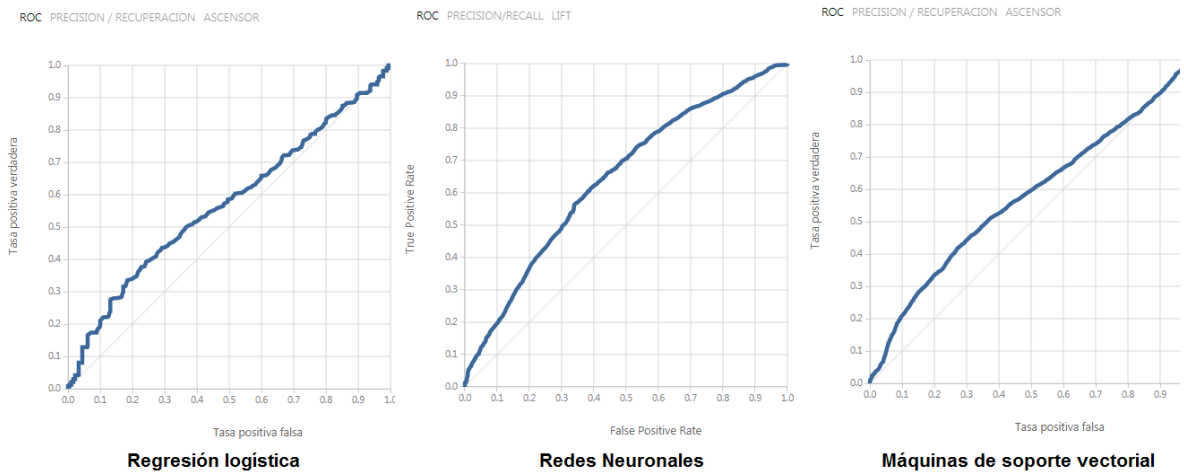


Figura 5-26.: Comparación de técnicas aplicadas

En la siguiente tabla se evidencian los resultados de forma agrupada de las técnicas utilizadas, teniendo en cuenta las variables relevantes de evaluación.

Técnica	Verdadero Positivo	Falso Positivo	Falso Negativo	Verdadero Negativo	Exactitud	Recall (Sensibilidad)	Precisión
Regresión logística	4872	3638	914	842	0.557	0.842	0.573
Redes Neuronales	3882	2054	1904	2426	0.614	0.671	0.654
Máquinas De Soporte Vectorial	4801	3667	985	813	0.547	0.830	0.567

Figura 5-27.: Comparación de matriz de confusión entre las técnicas aplicadas

Se grafican las variables a considerar, para realizar la comparación de las técnicas de Regresión logística, Redes neuronales y Maquinas de soporte vectorial:

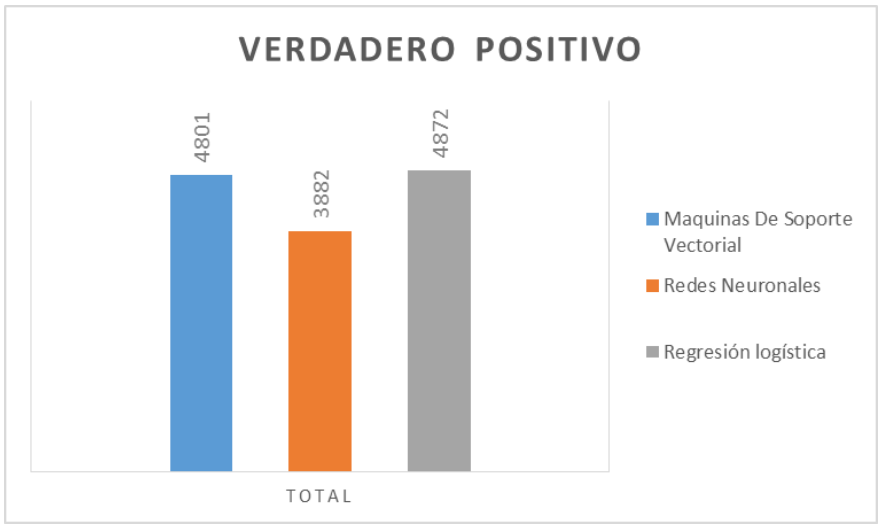


Figura 5-28.: Comparación de técnicas utilizando como referencia la variable de Verdadero Positivo, haciendo énfasis en la proporción de casos positivos que están bien detectadas por la técnica

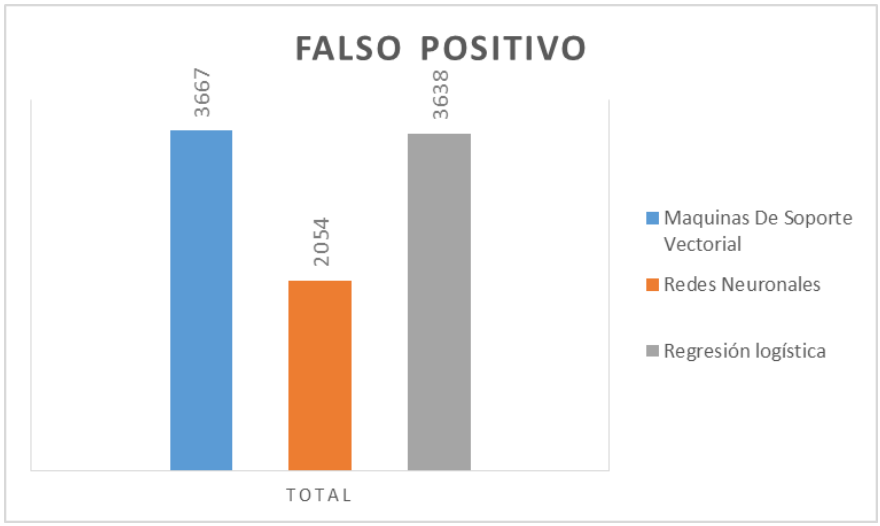


Figura 5-29.: Comparación de técnicas utilizando como referencia la variable de Falso Positivo, haciendo énfasis en la proporción de casos negativos que la técnica detecta como positivos

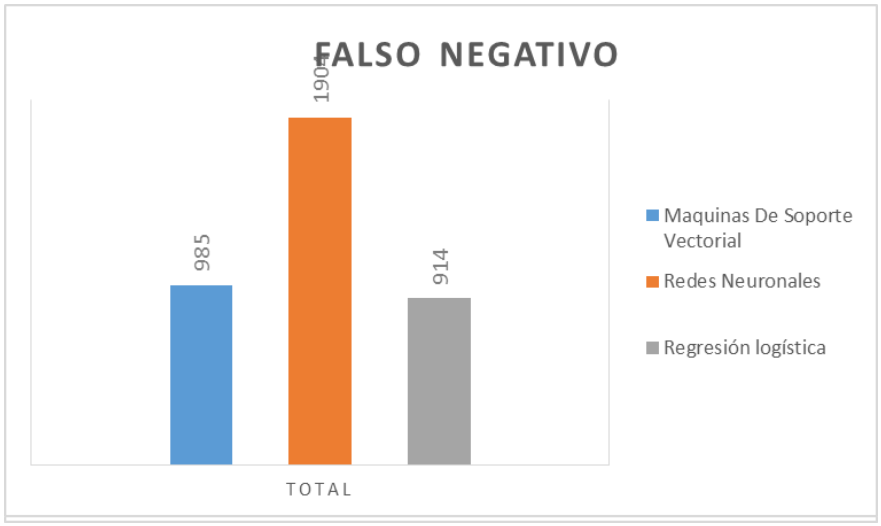


Figura 5-30.: Comparación de técnicas utilizando como referencia la variable de Falso Negativo, haciendo énfasis en la proporción de casos positivos que la técnica detecta como negativo

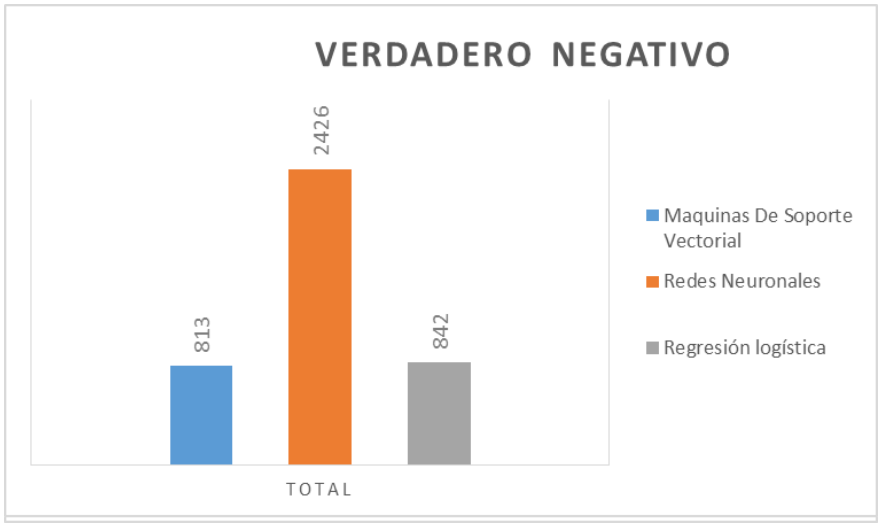


Figura 5-31.: Comparación de técnicas utilizando como referencia la variable de Verdadero Negativo, haciendo énfasis en la proporción de casos negativos que son bien detectados en la técnica

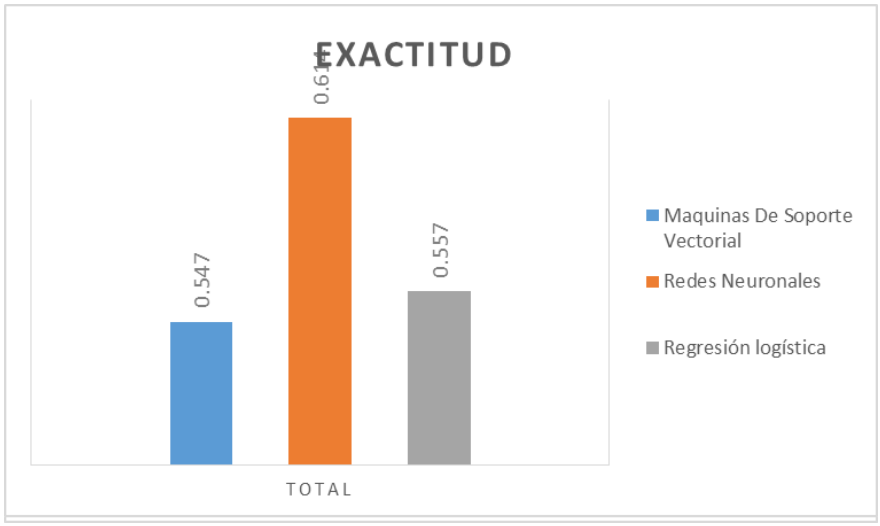


Figura 5-32.: Comparación de técnicas utilizando como referencia la variable de Exactitud, haciendo énfasis en el sesgo de la estimación o cuan cerca del valor real se encuentra el valor medio

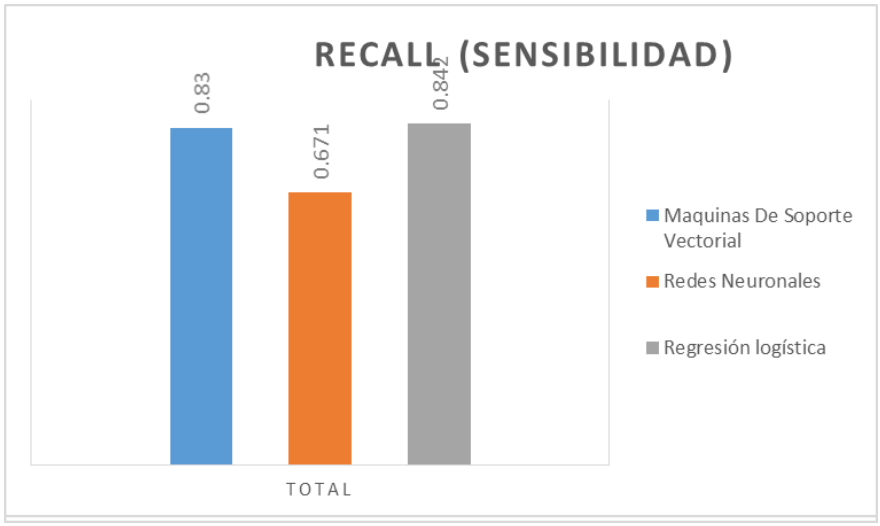


Figura 5-33.: Comparación de técnicas utilizando como referencia la variable de Recall o Sensibilidad, haciendo énfasis en la fracción de instancias relevantes que han sido recuperadas

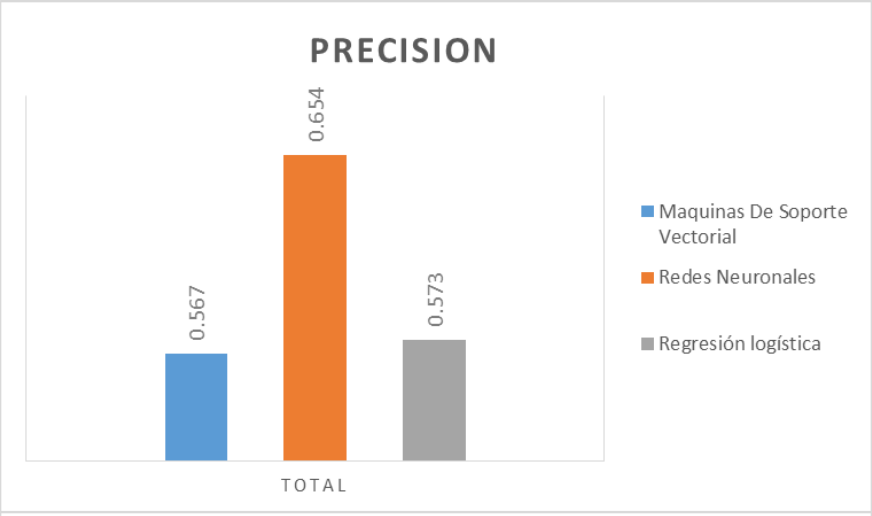


Figura 5-34.: Comparación de técnicas utilizando como referencia la variable de Precisión, haciendo énfasis en la dispersacion del conjunto de valores obtenidos de mediciones repetidas en una magnitud

Con el fin de realizar la comparación de modelo, específicamente de las técnicas de aprendizaje automático supervisado Maquinas de soporte vectorial, Redes neuronales y Regresión logística, se evalúan datos hidrológicos en las fechas del 10/04/1981 hasta el 31/08/2016; se seleccionan como variables principales para este proceso: La Variable Exactitud, Recall o Sensibilidad y Precisión.

Técnica	Exactitud	Recall (Sensibilidad)	Precisión
Regresión logística	0.557	0.842	0.573
Redes Neuronales	0.614	0.671	0.654
Máquinas De Soporte Vectorial	0.547	0.830	0.567

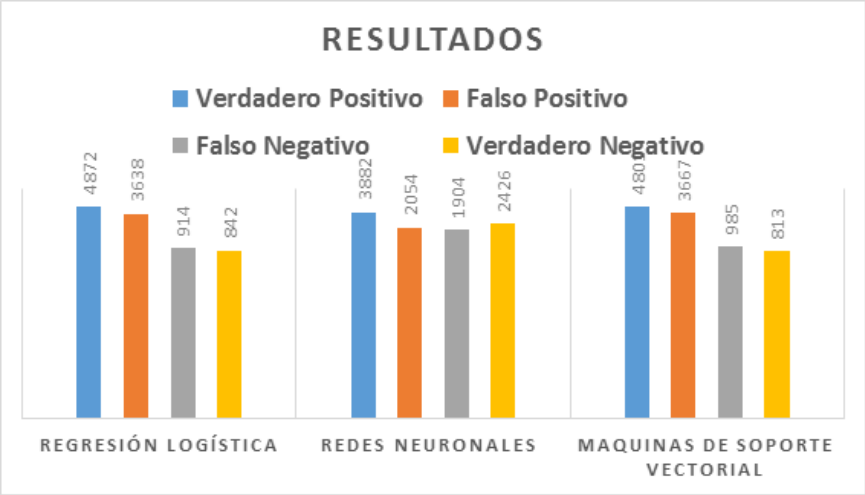


Figura 5-35.: Resultados matriz de confusión

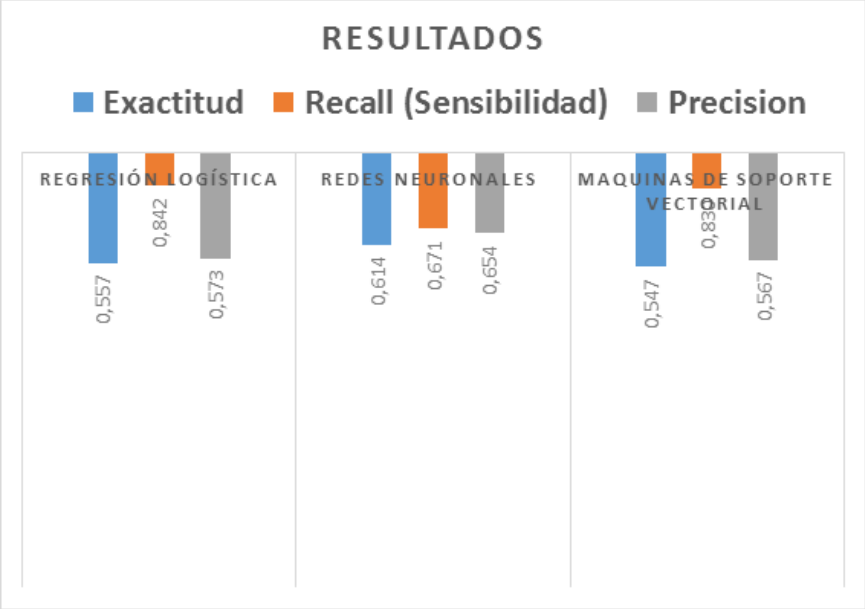


Figura 5-36.: Métricas de exactitud, sensibilidad y precisión

6. Conclusiones y recomendaciones

6.1. Conclusiones

Técnica	Exactitud	Recall (Sensibilidad)	Precisión
Regresión logística	0.557	0.842	0.573
Redes Neuronales	0.614	0.671	0.654
Máquinas De Soporte Vectorial	0.547	0.830	0.567

En base a la información suministrada por el IDEAM se tomaron como variables para el análisis las 2 mediciones de los niveles diarios tomados en la estación Hidrológica 29037610 Kilómetro 107 de corriente canal del dique y adicional la información del fenómeno del niño y niña; con lo cual se realizó la construcción del conjunto de datos a analizados. A raíz de la realización de pruebas con diversas técnicas, se llegó a la conclusión que las que presentan comportamientos deseados son Maquinas de soporte vectorial, redes neuronales y regresión logística; por lo que se aplican dichas técnicas en el conjunto de datos construido.

Con el propósito de realizar la aplicación de modelos de inteligencia artificial , específicamente en técnicas de aprendizaje automático supervisado Maquinas de soporte vectorial, Redes neuronales y Regresión logística, se evalúan datos hidrológicos registradas por la estación Hidrológica 29037610 Kilómetro 107 de corriente canal del dique, cuya información fue suministrada por el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia (IDEAM) en las fechas del 10/04/1981 hasta el 31/08/2016, con el fin de comparar las técnicas aplicadas se toman como variables de comparación la Exactitud, Precisión y Sensibilidad; dando como resultado que la técnica que la técnica Máquinas de soporte vectorial, obtuvo un mejor desempeño a nivel de Exactitud con 0.547, Precisión de 0.567 y obteniendo en Recall o Sensibilidad 0.830 presentando una leve diferencia de puntuación de 0.012 con la técnica Regresión Logística la cual obtuvo mejores resultados.
Los resultados de la aplicación de la técnica Redes neuronales, aunque sus resultados fueron aceptables, no superó en ninguno de las variables evaluadas con respecto a las otras técnicas evaluadas.

6.2. Recomendaciones

Posterior a la conclusión del proyecto investigativo realizado, se considera interesante investigar sobre los aspectos de la hidrología y fenómenos que influyen en la misma, por lo que se propone:

- Extender los estudios expuestos en este proyecto con datos hidrológicos de la ciudad de Cartagena y así ver el comportamiento de las mismas con relación a los fenómenos de la niña y el niño a mayor escala.
- Evaluar técnicas adicionales con el fin de buscar mejores resultados.
- Extender los estudios y comparaciones de las técnicas comparadas a los diversos datos climáticos.

A. Anexo: Cronograma

Figura A-1.: Cronograma de actividades

B. Anexo: Presupuesto

UNIVERSIDAD DEL SINÚ
Elías Bechara Zainúm
Seccional Cartagena

PROCESO: INVESTIGACIÓN, CIENCIA E INNOVACIÓN
TITULO: PRESUPUESTO PROYECTO DE INVESTIGACIÓN
CODIGO: R-INVE-030
VERSIÓN: 002

Titulo del proyecto:

Nombre del grupo:

Rubro	Recursos Unisinu Cartagena		Recursos Externos		Total
	Especie	Frescos	Especie	Frescos	
Personal	$ 1.440.000,00	$ -	$ 6.090.000,00	$ -	$ 7.530.000,00
Servicios técnicos	$ -	$ -	$ -	$ -	$ -
Equipos de uso propio	$ -		$ 1.800.000,00		$ 1.800.000,00
Compra de equipos	$ -	$ -	$ -	$ -	$ -
Materiales / insumos / reactivos	$ -	$ -	$ -	$ -	$ -
Salidas de campo	$ -	$ -	$ -	$ -	$ -
Software	$ -	$ -	$ -	$ -	$ -
Viajes	$ -	$ -	$ -	$ -	$ -
Gastos de publicación	$ -	$ -	$ -	$ -	$ -
Gastos de patentes	$ -	$ -	$ -	$ -	$ -
Total	$ 1.440.000,00	$ -	$ 7.890.000,00	$ -	
TOTAL					$ 9.330.000,00

	Entidades	Total	Especie	Frescos
Caracterización de la inversión	Inversión unisinu	15%	15%	0%
	Inversión externa	85%	85%	0%

Figura B-1.: Presupuesto

C. Anexo: Solicitud de información de periodo de fenómenos

Gmail

Yesenia Castillo <ycastillo07@gmail.com>

Solicitud de Información
1 mensaje

Subd. de Meteorologia (Subd. de Meteorologia) de Meteorologia (Subd. de Meteorologia) <meteorologia@ideam.gov.co>
Para: ycastillo07@gmail.com

5 de septiembre de 2018, 16:22

Buenas tardes,

Señora
YESENIA CASTILLO ROMÁN

Reciba un cordial saludo,

De manera atenta envío la respuesta a su solicitud mediante radicado No 20184000009661. Se envía un pdf firmado y el otro a color para que pueda apreciar mejor las gráficas.

Quedamos atentos a sus comentarios.

Cordialmente,

Subdirección de Meteorología
Instituto de Hidrología, Meteorología y Estudios Ambientales
PBX (571) 352 7160 Ext. 1403 - Línea nacional 018000 110 012
Calle 25 D No. 96 B - 70 Bogotá D.C.
www.ideam.gov.co

2 archivos adjuntos

20184000009661-Color.pdf
324K

20184000009661.pdf
529K

Figura C-1.: Solicitud de información de periodo de fenómenos

D. Anexo: Resultados de redes neuronales

-1.52444	1.304537	1.509833	-1.16394	-1.29529	0.878788	0.878788	0.624554
-0.34752	1.595702	-0.08175	0.819802	0.825937	0.878788	0.878788	0.601076
-1.23021	-0.73362	-0.76386	-2.10857	-1.93165	-1.13793	-1.13793	0.456429
-0.64175	0.722207	-1.55965	-0.2193	0.147145	-1.13793	0.878788	0.526707
-1.03406	0.139876	1.16878	-0.31376	-0.40437	0.878788	-1.13793	0.328968
0.927473	-1.60712	-0.53649	1.197657	1.122908	0.878788	0.878788	0.744807
-1.42636	-1.60712	-1.33228	-1.06947	-1.46499	0.878788	0.878788	0.633563
1.221702	-1.31595	1.055095	0.394715	0.316843	0.878788	0.878788	0.681547
0.73132	-0.73362	1.16878	-0.12484	0.231994	-1.13793	-1.13793	0.426717
1.221702	-0.44245	1.737201	2.047831	1.844124	0.878788	0.878788	0.660694
0.633244	-0.73362	-1.67333	-0.92778	-1.29529	-1.13793	-1.13793	0.355357
0.240938	0.139876	0.372989	-0.88055	-0.4468	-1.13793	-1.13793	0.363434
-0.34752	-0.44245	-0.08175	-0.78608	-0.53165	-1.13793	-1.13793	0.331596
1.319779	0.139876	-1.10491	0.914266	0.741088	-1.13793	0.878788	0.643247
1.515932	0.431042	1.737201	0.205788	0.316843	0.878788	0.878788	0.716119
-1.13213	0.431042	0.259305	-0.45546	-0.19225	0.878788	-1.13793	0.362763
0.535167	-0.15129	0.714042	1.197657	0.528965	-1.13793	-1.13793	0.394724
-0.54367	0.431042	0.486673	-0.59715	-0.1074	-1.13793	-1.13793	0.40053
-0.64175	-0.73362	1.623517	-1.44733	-0.95589	0.878788	-1.13793	0.37323
0.339014	-1.60712	1.737201	-0.97501	0.019872	-1.13793	0.878788	0.514747
-0.73983	-1.02478	1.282464	-1.68349	-1.88923	-1.13793	-1.13793	0.425739
0.142862	-1.02478	0.372989	-0.40823	-0.1074	0.878788	-1.13793	0.402375
0.927473	0.139876	0.031936	1.292121	1.33503	-1.13793	0.878788	0.541293
-0.05329	-1.02478	0.031936	-1.30563	-1.12559	0.878788	-1.13793	0.377342
0.142862	0.722207	-0.30912	0.819802	1.122908	0.878788	0.878788	0.58053
0.240938	1.595702	-1.67333	0.347484	0.656239	-1.13793	0.878788	0.593989
-0.15137	1.304537	-1.21859	-0.45546	-0.02255	0.878788	0.878788	0.603025
0.535167	1.013372	-0.99122	0.630875	1.207757	0.878788	0.878788	0.623603
-1.42636	1.304537	-1.10491	0.347484	0.274418	0.878788	0.878788	0.629459
1.02555	-1.02478	1.623517	0.01686	-0.19225	-1.13793	0.878788	0.548768
-0.64175	0.139876	0.714042	-0.88055	-0.78619	-1.13793	-1.13793	0.326104
-0.34752	0.431042	1.509833	0.630875	0.528965	0.878788	-1.13793	0.390482
-0.34752	-1.31595	-1.55965	-1.54179	-1.16801	0.878788	-1.13793	0.437906

Figura D-1.: Resultados de redes neuronales

E. Anexo: Resultados de regresión logística

-1.328285	0.431042	-0.081749	-1.636256	-1.380136	0.878788	0.878788	0.516954
0.829397	-0.151289	0.714042	1.433817	1.292606	-1.13793	0.878788	0.595012
1.712085	-1.607115	0.259305	0.394715	-0.489222	0.878788	0.878788	0.580016
-0.151368	-0.442454	0.031936	-0.691618	-0.531646	0.878788	0.878788	0.538279
0.829397	0.722207	-0.65017	1.622744	1.462303	0.878788	0.878788	0.628016
1.712085	-0.442454	-0.195433	1.055962	0.231994	-1.13793	0.878788	0.624013
-1.524438	-0.733619	1.623517	-2.014111	-1.761956	0.878788	-1.13793	0.461467
0.535167	-0.733619	1.282464	0.01686	-0.192251	-1.13793	0.878788	0.558759
-1.524438	1.304537	1.509833	-1.163937	-1.295287	0.878788	0.878788	0.541421
-0.347521	1.595702	-0.081749	0.819802	0.825937	0.878788	0.878788	0.607623
-1.230209	-0.733619	-0.763855	-2.108574	-1.931654	-1.13793	-1.13793	0.476071
-0.64175	0.722207	-1.559645	-0.219299	0.147145	-1.13793	0.878788	0.561196
-1.034056	0.139876	1.16878	-0.313763	-0.404373	0.878788	0.878788	0.519217
0.927473	-1.607115	-0.536486	1.197657	1.122908	0.878788	0.878788	0.543966
-1.426361	-1.607115	-1.332277	-1.069473	-1.464985	0.878788	-1.13793	0.433293
1.221702	-1.31595	1.055095	0.394715	0.316843	0.878788	0.878788	0.568295
0.73132	-0.733619	1.16878	-0.124835	0.231994	-1.13793	0.878788	0.567881
1.221702	-0.442454	1.737201	2.047831	1.844124	0.878788	0.878788	0.6016
0.633244	-0.733619	-1.67333	-0.927777	-1.295287	-1.13793	0.878788	0.564132
0.240938	0.139876	0.372989	-0.880545	-0.446797	-1.13793	0.878788	0.5789
-0.347521	-0.442454	-0.081749	-0.786082	-0.531646	-1.13793	0.878788	0.529087
1.319779	0.139876	-1.104908	0.914266	0.741088	-1.13793	0.878788	0.628328
1.515932	0.431042	1.737201	0.205788	0.316843	0.878788	0.878788	0.646461
-1.132132	0.431042	0.259305	-0.455458	-0.192251	0.878788	0.878788	0.526274
0.535167	-0.151289	0.714042	1.197657	0.528965	-1.13793	0.878788	0.581523
-0.543673	0.431042	0.486673	-0.597154	-0.107402	-1.13793	0.878788	0.55376
-0.64175	-0.733619	1.623517	-1.447328	-0.955891	0.878788	0.878788	0.503138
0.339014	-1.607115	1.737201	-0.975009	0.019872	-1.13793	0.878788	0.515192
-0.739826	-1.024784	1.282464	-1.683487	-1.889229	-1.13793	-1.13793	0.487166
0.142862	-1.024784	0.372989	-0.408227	-0.107402	0.878788	0.878788	0.529296
0.927473	0.139876	0.031936	1.292121	1.33503	-1.13793	0.878788	0.610576
-0.053291	-1.024784	0.031936	-1.305632	-1.125589	0.878788	0.878788	0.520028
0.142862	0.722207	-0.309117	0.819802	1.122908	0.878788	0.878788	0.597039
0.240938	1.595702	-1.67333	0.347484	0.656239	-1.13793	0.878788	0.634185
-0.151368	1.304537	-1.218592	-0.455458	-0.022553	0.878788	0.878788	0.605717
0.535167	1.013372	-0.991224	0.630875	1.207757	0.878788	0.878788	0.625665
-1.426361	1.304537	-1.104908	0.347484	0.274418	0.878788	0.878788	0.547122
1.02555	-1.024784	1.623517	0.01686	-0.192251	-1.13793	0.878788	0.570172
-0.64175	0.139876	0.714042	-0.880545	-0.786193	-1.13793	0.878788	0.537725
-0.347521	0.431042	1.509833	0.630875	0.528965	0.878788	0.878788	0.562755
-0.347521	-1.31595	-1.559645	-1.541792	-1.168013	0.878788	-1.13793	0.495223

Figura E-1.: Resultados de regresión logística

F. Anexo: Resultados de maquinas de soporte vectorial

0.535167	-0.733619	1.282464	0.01686	-0.192251	-1.13793	0.878788	0.55575
-1.524438	1.304537	1.509833	-1.163937	-1.295287	0.878788	0.878788	0.514457
-0.347521	1.595702	-0.081749	0.819802	0.825937	0.878788	0.878788	0.612549
-1.230209	-0.733619	-0.763855	-2.108574	-1.931654	-1.13793	-1.13793	0.462561
-0.64175	0.722207	-1.559645	-0.219299	0.147145	-1.13793	0.878788	0.567238
-1.034056	0.139876	1.16878	-0.313763	-0.404373	0.878788	0.878788	0.514855
0.927473	-1.607115	-0.536486	1.197657	1.122908	0.878788	0.878788	0.576195
-1.426361	-1.607115	-1.332277	-1.069473	-1.464985	0.878788	-1.13793	0.45046
1.221702	-1.31595	1.055095	0.394715	0.316843	0.878788	0.878788	0.572725
0.73132	-0.733619	1.16878	-0.124835	0.231994	-1.13793	0.878788	0.560762
1.221702	-0.442454	1.737201	2.047831	1.844124	0.878788	0.878788	0.621095
0.633244	-0.733619	-1.67333	-0.927777	-1.295287	-1.13793	0.878788	0.564045
0.240938	0.139876	0.372989	-0.880545	-0.446797	-1.13793	0.878788	0.560474
-0.347521	-0.442454	-0.081749	-0.786082	-0.531646	-1.13793	0.878788	0.524293
1.319779	0.139876	-1.104908	0.914266	0.741088	-1.13793	0.878788	0.641553
1.515932	0.431042	1.737201	0.205788	0.316843	0.878788	0.878788	0.627668
-1.132132	0.431042	0.259305	-0.455458	-0.192251	0.878788	0.878788	0.522913
0.535167	-0.151289	0.714042	1.197657	0.528965	-1.13793	0.878788	0.597306
-0.543673	0.431042	0.486673	-0.597154	-0.107402	-1.13793	0.878788	0.542035
-0.64175	-0.733619	1.623517	-1.447328	-0.955891	0.878788	-1.13793	0.480384
0.339014	-1.607115	1.737201	-0.975009	0.019872	-1.13793	-1.13793	0.498991
-0.739826	-1.024784	1.282464	-1.683487	-1.889229	-1.13793	-1.13793	0.46732
0.142862	-1.024784	0.372989	-0.408227	-0.107402	0.878788	0.878788	0.529222
0.927473	0.139876	0.031936	1.292121	1.33503	-1.13793	0.878788	0.625267
-0.053291	-1.024784	0.031936	-1.305632	-1.125589	0.878788	0.878788	0.50885
0.142862	0.722207	-0.309117	0.819802	1.122908	0.878788	0.878788	0.606591
0.240938	1.595702	-1.67333	0.347484	0.656239	-1.13793	0.878788	0.636761
-0.151368	1.304537	-1.218592	-0.455458	-0.022553	0.878788	0.878788	0.597528
0.535167	1.013372	-0.991224	0.630875	1.207757	0.878788	0.878788	0.630712
-1.426361	1.304537	-1.104908	0.347484	0.274418	0.878788	0.878788	0.561136
1.02555	-1.024784	1.623517	0.01686	-0.192251	-1.13793	0.878788	0.564033
-0.64175	0.139876	0.714042	-0.880545	-0.786193	-1.13793	0.878788	0.52374
-0.347521	0.431042	1.509833	0.630875	0.528965	0.878788	0.878788	0.564589
-0.347521	-1.31595	-1.559645	-1.541792	-1.168013	0.878788	-1.13793	0.49452
-0.151368	-1.607115	0.600358	-0.408227	-1.04074	-1.13793	-1.13793	0.499944
-0.837903	-0.733619	-0.763855	-1.447328	-1.295287	-1.13793	-1.13793	0.489406
-0.64175	-1.31595	-0.536486	-0.786082	-1.380136	0.878788	-1.13793	0.490501
-1.524438	0.722207	-0.536486	-1.541792	-0.998315	0.878788	0.878788	0.502682
1.319779	0.431042	1.055095	0.489179	1.038059	0.878788	0.878788	0.628248
-0.445597	-0.442454	0.827726	0.25302	0.316843	-1.13793	0.878788	0.532046
0.240938	1.013372	-0.991224	-0.219299	0.401692	-1.13793	0.878788	0.605987

Figura F-1.: Resultados de maquinas de soporte vectorial

Bibliografía

[1] Gustavo A. Betancourt. Las maquinas de soporte vectorial, 2005. Pereira.

[2] Jorge Matich CDamián. Redes Neuronales: Conceptos Básicos y aplicaciones. page 55, March 2001.

[3] Sancho Caparrini Fernando. Introducción al Aprendizaje Automático.

[4] Ovando G., Bocco M., and Sayago S. Redes Neuronales para modelar predicción de heladas. Argentina.

[5] José Antonio Gallego, Álvaro Ortiz, and Juan Carlos Plaza. Las 5 licencias de software libre más importantes que todo desarrollador debe conocer.

[6] Lozada J. Investigación Aplicada: Definición, propiedad Intelectual e Industria. page 6. Quito, Ecuador.

[7] Resendiz Trejo Juan Angel. *Las maquinas de vectores de soporte para identificación en línea*. PhD thesis.

[8] Machine Learning. Regresión logística, September 2017.

[9] Microsoft. Algoritmo de regresión logística de Microsoft.

[10] Microsoft. Qué es Azure Machine Learning Studio.

[11] Herrera Batista Miguel Ángel. Diseño: entre el diseño científico y las ciencias de lo artificial.

[12] Fernando Pavón Pérez. *Generación de Conocimiento basado en Aprendizaje Automático y Aplicación en Diferentes Sectores.* PhD thesis, Universidad Nacional de Educación a Distancia, Madrid, February 2016.

[13] Rodriguez S. and Vidal Martinez A. Clasificación de células cervicales con máquinas de soporte vectorial empleando rasgos del núcleo. June 2015. Ciudad de la Habana.

[14] Lara V. H., Martín Isabel M. P., and Martínez Vega J. Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional. page 30. Madrid.

[15] Vargas Cordero Z. R. *La investigación aplicada: una forma de conocer las realidades con evidencia científica.* PhD thesis. Costa Rica.