ANÁLISIS DE DESEMPEÑO DE ALGORITMOS DE REGRESIÓN
USANDO SCIKIT-LEARN.

A. Florian[1] y J. Vélez[2]

[1] Red Tecnoparque Colombia, Nodo Medellín, Servicio Nacional de Aprendizaje SENA. Colombia.
[2] Universidad Pontificia Bolivariana, Medellín.

Palabras claves: Aprendizaje profundo, algoritmos de aprendizaje, predicción, optimización.

RESUMEN

El aprendizaje automático y el aprendizaje profundo son actualmente áreas de investigación extremadamente activas, ya que permiten proyecciones y estimaciones de variables basadas en datos históricos y la ejecución de algoritmos matemáticos; estos algoritmos están especializados para realizar regresiones, clasificación, predicción u otras aplicaciones donde se utilizan grandes volúmenes de datos y se requieren decisiones con un alto margen de precisión [1]. La gran cantidad de datos disponibles en la actualidad brinda grandes oportunidades y el potencial transformador de diferentes sectores como medicina, banca, ingeniería, deportes, entre otros; pero también presentan grandes desafíos en el uso efectivo de la información, ya que un análisis deficiente o erróneo de los datos conduce a una toma de decisiones equivocada. A medida que los datos crecen, el aprendizaje profundo toma un papel protagonista en el análisis y solución de problemas con un alto grado de complejidad que en un entorno natural no son fáciles de resolver porque presentan matices que solo con el uso de algoritmos de aprendizaje profundo se pueden observar. [2]. Las mejoras en el poder computacional, los grandes volúmenes de información, el almacenamiento rápido de datos y la paralelización han contribuido al análisis y la predicción de Big Data en áreas como la predicción de precios, el análisis de imágenes médicas, el control del tráfico e incluso el estudio del rendimiento del equipo de fútbol, entre muchas otras. Con todo lo anterior, se da una idea de la importancia actual de esta área de la ciencia y lo pertinente que es trabajar en este tema [3].

01 INTRODUCCIÓN

Python es un lenguaje de programación interpretado, orientado a objetos, multimodal el cual en la actualidad es un lenguaje de programación usado por desarrolladores, ingenieros y científicos de diferentes áreas, entre las que se destaca los analistas de datos tanto en el área financiera, como en la medicina y muchos otros campos, este lenguaje cuenta con una gran cantidad de herramientas entre las que se destaca scikit-learn que integra una amplia gama de algoritmos de aprendizaje automático de última generación para resolución de problemas supervisados y no supervisados; todo esto, aunado a la rápida incorporación de la tecnología con su poder predictivo y su capacidad para generar funciones optimizadas, ha hecho que el aprendizaje profundo haya despertado un interés creciente en la generación de modelos analíticos basados en datos [4].

El aprendizaje profundo y Big data son las dos tendencias más populares en el mundo digital con un rápido crecimiento, gracias a la amplia disponibilidad de datos que son incluso difíciles de administrar y analizar. Esta gran cantidad de datos ha conducido a un cambio dramático de paradigma en la investigación científica y el descubrimiento basado en datos [5]. El análisis de Big data ofrece un gran potencial para revolucionar aspectos de nuestra sociedad, es por esto que los diferentes modelos de análisis de información hoy toman gran importancia [6]; basado en todo lo anterior, surge este documento donde se hace un análisis comparativo de algunas técnicas de regresión lineal aplicada al análisis de datos, haciendo uso de las herramientas que ofrece Python y sus bases de datos (datasets - boston), en este documento se presentan resultados del entrenamiento de modelos de aprendizaje automático como lo son: regresión lineal, regresión múltiple, regresión lineal polinomial y vectores de soporte. Estas técnicas se aplicarán a la misma base de datos y se analizarán los resultados obtenidos para luego concluir sobre el ajuste y desempeño de cada una de ellas a este tipo de datos.

02 MATERIALES Y MÉTODOS

Una de las técnicas usadas para medir la precisión del modelo de regresión lineal (RL) es la determinación del coeficiente de determinación R^2, el cual calcula la calidad del modelo para replicar los resultados de la regresión. El rango de R^2 está entre 0 y 1, siendo 1, el resultado más óptimo para el entrenamiento del sistema [8].

Las Máquinas de Vectores Soporte (MVS) son sistemas de aprendizaje que utilizan como hipótesis funciones lineales en espacios característicos de grandes dimensiones [13], las MVS pertenecen a la categoría de los clasificadores lineales, ya que inducen separadores lineales llamados hiperplanos. La idea es seleccionar un plano de separación que equidista de los datos más cercanos para formar el margen máximo de cada hiperplano y, de esta manera, dar solución a los problemas de clasificación binaria y demás problemas de clasificación y regresión [14].

Los modelos de regresión polinomial (RLP) se usan cuando la variable de respuesta muestra un comportamiento no lineal. Los estimadores de los parámetros del modelo polinomial se obtienen por el método de los mínimos cuadrados, la RLP produce un polinomio que describe la relación entre cualquier conjunto de entradas y la salida correspondiente [10], en este método se crea una función basada en un polinomio de n-ésimo grado, tomando la suma de los cuadrados de los residuos y su derivada con respecto a cada uno de los coeficientes se puede obtener un sistema de ecuaciones [11]. El sistema puede reducirse hasta encontrar un polinomio de grado m con mínimos cuadrados, resolviendo un sistema de m+1 ecuaciones lineales simultáneas.

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m \quad (8)$$

$$s_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x - a_2 x_i^2 - \ldots - a_m x_i^m)^2 \quad (9)$$

2.1 REGRESIÓN LINEAL SIMPLE

El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) y un conjunto de variables independientes (variables explicativas) $x_1, \ldots x_n$ se busca una función que sea una buena apro-

ximación de una nube de puntos (x_i, y_i), mediante una línea, modelo de regresión lineal se puede expresar de la siguiente manera:

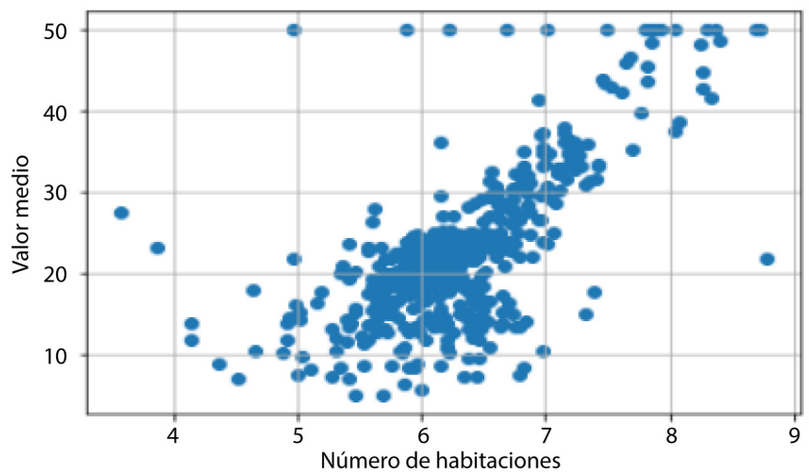
$$y = \alpha + \beta x + \epsilon \qquad (1)$$

En donde α es la ordenada en el origen, β es la pendiente de la recta, ϵ una variable que incluye diferentes factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud, a la que llamaremos error. x & y son variables aleatorias, por lo que no se puede establecer una relación lineal exacta entre ellas. Para hacer una estimación del modelo de regresión lineal simple, trataremos de buscar una recta de la forma:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x \qquad (2)$$

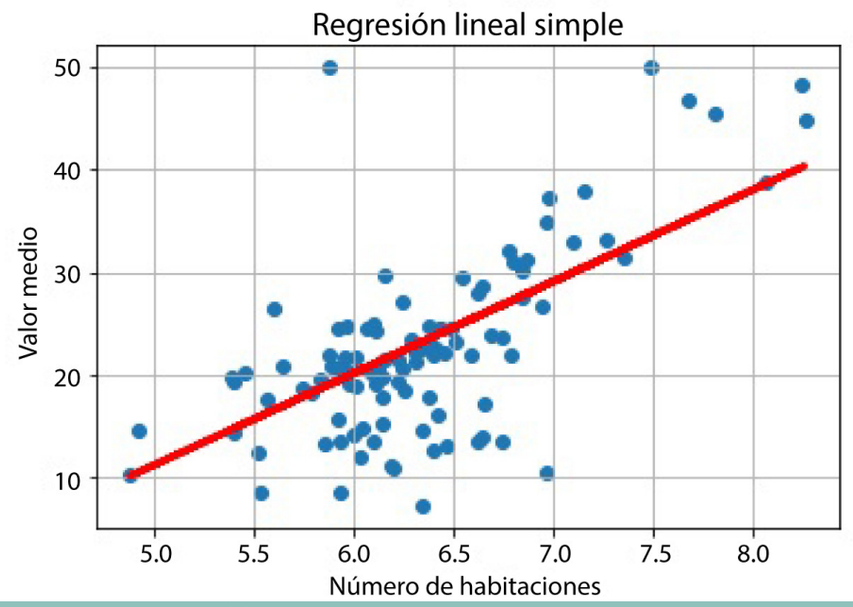
Tal que se ajuste lo mejor posible a la nube de puntos analizada [7].

Como se mencionó anteriormente, en este trabajo se realizará un análisis comparativo del rendimiento de algunas de las técnicas de regresión lineal (RL), y como ejemplo para esto, se toma una base de datos de la librería sklearn de Python la cual tiene entre sus parámetros los precios de las casas en Boston Estados Unidos, usando como criterio el número de habitaciones con que cuenta la vivienda. La gráfica 1 muestra de manera ilustrativa la base de datos.



Gráfica 1. Representación gráfica de la base de datos

Con esta base de datos se entrenará cada uno de los modelos, con el fin de analizar el rendimiento ante este tipo de información. En primer lugar, se ejecutará un algoritmo de regresión lineal simple y se procederá a calcular el coeficiente de determinación R^2 para determinar de manera cuantitativa como se ajusta este sistema a una base de datos de este tipo, como se observa en la gráfica 2 la cual muestra cómo se ajusta un modelo de RL simple a los datos en estudio.



Gráfica 2. Entrenamiento sistema RL

Datos del modelo:
Valor de la pendiente: 8.78979
Valor de la intersección: -32.91017

La ecuación de regresión lineal es:

$$\hat{y} = 8.789799$$
$$\hat{x} = \text{-}32.91017$$

La precisión del algoritmo es: 0.44225

Una de las técnicas usadas para medir la precisión del modelo de RL es la determinación del coeficiente de determinación R^2, el cual calcula la calidad del modelo para replicar los resultados de la regresión. El rango de R^2 está entre 0 y 1, siendo 1 el resultado esperado del entrenamiento del sistema [8]; lo anterior, brinda criterios de evaluación

sobre el desempeño del modelo ante este problema específico, con base en la teoría previamente descrita y los resultados obtenidos, se observa que al entrenar el modelo de RL con estos datos resulta una precisión de 0.442 en la escala del coeficiente de determinación. A todas luces se puede distinguir que este no es un buen desempeño de la RL para ajustarse a este tipo de datos, razón por la cual se debe recurrir al uso de otras técnicas de machine learning que puedan ajustarse mejor, con base en lo antes observado se procederá a ajustar el modelo de regresión lineal múltiple con el fin de determinar su desempeño ante este tipo de información.

2.2 REGRESIÓN LINEAL MÚLTIPLE (RLM)

El entrenamiento del modelo de RLM para este tipo de datos muestra un comportamiento mejor que el modelo de RL, esto puede ser observado en la precisión del modelo de RLM el cual es más cercano a la unidad; si bien la RLM presenta un mejor comportamiento que la RL, esto no significa que este modelo sea el más apropiado para estos datos, ya que la precisión sólo alcanza un poco más del 50% del valor máximo de precisión posible al que puede llegar este tipo de modelo.

En la regresión lineal múltiple (RLM), la variable endógena es una función lineal de k regresores, correspondientes a las variables de transformaciones de las mismas y una perturbación aleatoria. El modelo también incluye un término independiente.

Si se designa y a la variable endógena, $x_2, x_3, \ldots, x_k$ a los regresores y u al error o perturbación aleatoria, el modelo de regresión lineal múltiple se podrá expresar como:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u \qquad (3)$$

Los parámetros β son desconocidos.

En el segundo miembro de la ecuación (3) se pueden distinguir dos componentes: un componente sistemático $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$ y la perturbación aleatoria u. Llamando μ_y y el componente sistemático se puede expresar como:

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \qquad (4)$$

De acuerdo con (3) $\mu_y y$ es una función lineal en los parámetros $\beta_1, \beta_2, \beta_3, \ldots, \beta_x$ Ahora, supongamos que tenemos una muestra aleatoria de tamaño $n, \{y_i x_{2i}, x_{3i}, \ldots x_{ki}; \text{ con } i = 1, 2, 3, \ldots n\}$ extraída de la población estudiada. Si expresamos el modelo poblacional para todas las observaciones de la muestra, se obtiene el siguiente sistema:

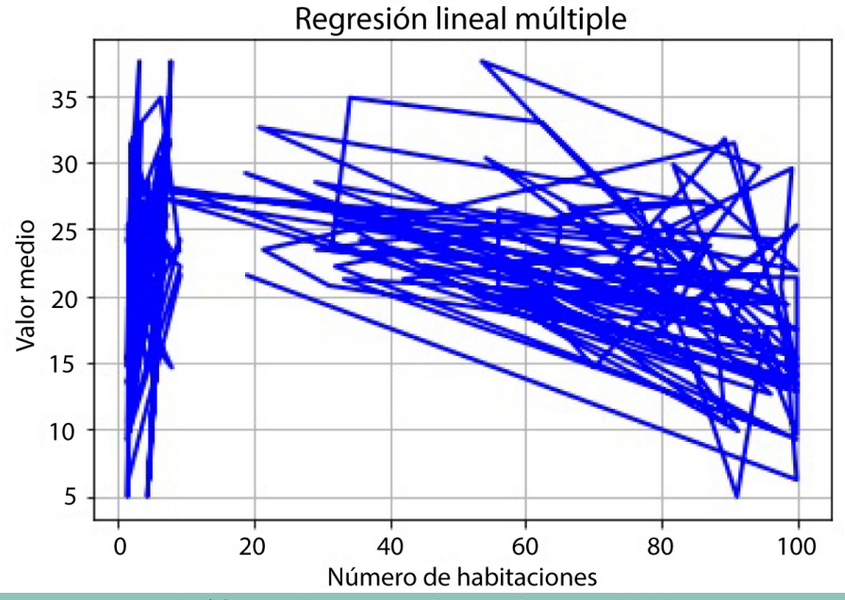
$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + u_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + u_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \ldots + \beta_k x_{kn} + u_n \end{aligned} \qquad (5)$$

Este sistema de ecuaciones puede expresarse de una forma matricial.

$$\begin{bmatrix} y_1 \, y_2 \, . \, y_n \end{bmatrix} = \begin{bmatrix} 1 \, x_{21} \, x_{31} \, . \, x_{k1} \;\; 1 \, . \, 1 \;\; x_{22} \, . \, x_{2n} \;\; x_{32} \, . \, x_{k2} \;\; \ldots \;\; x_{3n} \, . \, x_{kn} \end{bmatrix} + \begin{bmatrix} u_1 \, u_2 \, . \, u_n \end{bmatrix} \qquad (6)$$

Este sistema se puede expresar de manera general como lo muestra la ecuación (7).

$$y = x + u \qquad (7)$$



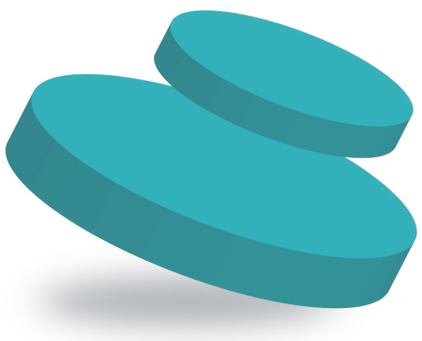
Gráfica 3. Entrenamiento sistema RLM

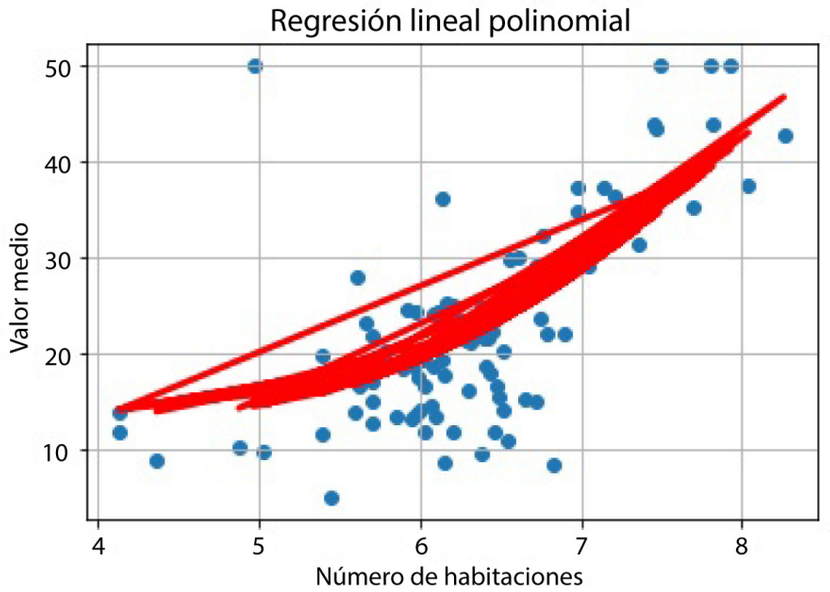
Donde y es un vector n 1, x una matriz n k, es un vector k 1 y u es un vector n 1 [9].
La RLM es otra herramienta con la cual se cuenta para hacer análisis de regresión y predicción, tal que se pueda tomar decisiones para la compra de vivienda o especulación de precios de productos con algunas características especiales que se puedan ajustar a este tipo de datos, en la gráfica 3 muestra el comportamiento de modelo de RLM aplicado a la base de datos en análisis.

Valor de las pendientes
A= [8.16924076, -0.1042, -0.48172265]
Valor de la intersección o coeficiente
B= -19.90609
La precisión del Algoritmo es: 0.522675

2.3 REGRESIÓN LINEAL POLINOMIAL

El modelo de regresión polinomial se ajusta a relaciones no lineales entre el valor de x (número de habitaciones) y la correspondiente media condicional de y (valor medio de venta) que son los datos usados para el análisis de este trabajo. La regresión polinómica se ajusta a modelos no lineales, a datos y a problemas de estimación estadística, es lineal en el sentido de que la función de regresión $E(y \mid x)$ es lineal en los parámetros desconocidos que se calculan a partir de los datos [12], es por esta razón que se entrena el sistema y se entregan los resultados en este documento, tal y como se muestra en la gráfica 4, con el fin de comparar su rendimiento y precisión con respecto a los métodos de regresión anteriormente analizados.





Gráfica 4. Entrenamiento sistema RLP

Valor de la pendiente o coeficiente A es:
[0.0, -21.74734527, 2.37548685]
Valor de la intersección o coeficiente B es: 64.1397
Precisión del modelo es: 0.548958

A medida que se utilizan modelos que se ajustan más a datos con alta dispersión y algún grado de no linealidad, es preciso la implementación de modelos que sean flexibles y adaptables, como se observa el modelo de RLP se ajusta de una mejor manera a los datos de entrenamiento, lo cual se refleja en la precisión del modelo que es algo mejor que la de los modelos anteriormente estudiados, esto podría dar indicios de que para este tipo de datos se deba usar la RLP en lugar de la RL y las RPM.

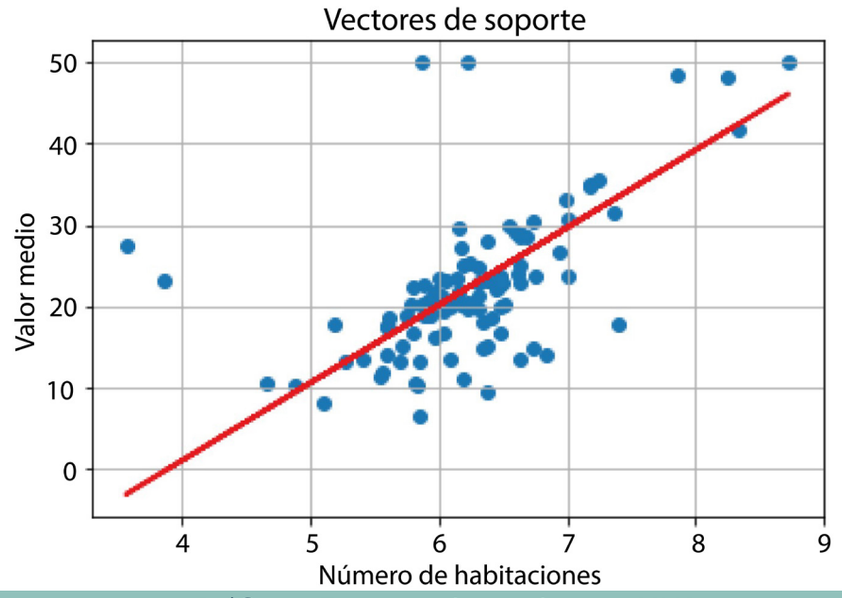
2.4 MÁQUINAS DE VECTORES DE SOPORTE (MVS)

Dados los datos de entrenamiento y sus etiquetas correspondientes $(x_n, y_n), n = 1, \ldots, N, x_n \in R^D$, $t_n \in \{-1, +1\}$, el aprendizaje de las MVS consiste en la siguiente optimización restringida:

$$\frac{1}{2}W^T W + C \sum_{n=1}^{N} \xi_n \qquad (10)$$

ξ_n son variables flojas que penalizan los puntos de datos que violan los requisitos de margen. Tenga en cuenta que podemos incluir el sesgo aumentando todos los vectores de datos x_n con un valor escalar de 1. El problema de optimización no restringido correspondiente es el siguiente:

$$\frac{1}{2}W^T W + C\sum_{n=1}^{N} max(1 - W^T x_n t_n, 0) \qquad (11)$$



Gráfica 5. Entrenamiento sistema MVS

03 RESULTADOS Y DISCUSIONES

Como resultado de este trabajo se aprecian las mejoras en el poder computacional. La paralelización en el manejo de información ha contribuido al análisis de datos en áreas como la predicción de precios, el análisis de imágenes médicas, el control del tráfico, entre muchas otras. Lo anterior, da una idea de la importancia de las nuevas técnicas de inteligencia artificial y la relevancia que tiene en los diferentes campos de la ciencia y la tecnología moderna.

Las MVS son una alternativa más en los sistemas de aprendizaje supervisado que se pueden utilizar en diferentes aplicaciones de aprendizaje automático y con diferentes tipos de datos, para este caso en particular, se entrenó el modelo tal y como hizo en los casos previamente estudiados. Se analizó su precisión utilizando el coeficiente de determinación R^2 que es una métrica apropiada para estos sistemas de regresión lineal, la cual arrojó que la precisión del modelo para este tipo de datos es de: 0.526243. Esto se hace visible en la gráfica 5, donde se muestra que muchos puntos de prueba quedan por fuera de la línea de regresión, evidenciando la razón por la cual la precisión de la MVS para esta aplicación es de poco más del 50%, valor que está significativamente lejos de la unidad que es el objetivo de un modelo de regresión, dado que esto indicaría que el modelo se ajusta completamente a los datos y que todas sus predicciones serían acertadas, cosa que no aplicaría en este caso debido a que este modelo daría una incertidumbre en la predicción de 50%, valor que es realmente bajo si se quiere tomar algún tipo de decisión basado en este modelo.

Por otro lado, el entrenamiento del modelo de RLM para este tipo de datos muestra un comportamiento mejor que el modelo de RL, esto puede ser observado en la precisión del modelo de RLM el cual es más cercano a la unidad; si bien la RLM presenta un mejor comportamiento que la RL, esto no significa que este modelo sea el más apropiado para estos datos, ya que la precisión sólo alcanza un poco más del 50% del valor máximo de precisión posible al que puede llegar este tipo de modelo.

04 CONCLUSIONES

El aprendizaje profundo y Big Data son las dos tendencias más populares en el mundo digital con un rápido crecimiento, gracias a la amplia disponibilidad de datos y a la necesidad de tener información depurada y analizada para la toma de decisiones en diferentes áreas de la ciencia, la economía y hasta en el sector de la salud.

Los diferentes modelos de regresión representan un papel muy importante en el análisis preliminar de datos y modelos de aprendizaje supervisado, ya que con estos se puede tener una idea general del comportamiento de la información en análisis, sus tendencias y sus posibles variaciones futuras.

Es de vital importancia contar con herramientas que permitan valorar el comportamiento, la precisión y demás parámetros de calidad de los modelos de aprendizaje automático dado que, del resultado del ajuste de estos, se tomarán decisiones que podrían afectar el futuro de una empresa, mejorar el tráfico de una ciudad y hasta salvar vidas.

Basados en las simulaciones, resultados y conclusiones alcanzadas en este documento, se propone como trabajos futuros, realizar estudios de otras técnicas de aprendizaje automático con el fin de encontrar modelos que puedan ser más apropiados para el tipo de datos anteriormente analizados, además de aplicar otro tipo de métricas de calidad a los modelos de aprendizaje automático que brinden más información sobre el rendimiento y demás parámetros de los modelos.

05 REFERENCIAS

[1] H. A. B. González, "Aplicación del análisis de regresión lineal simple para la estimación de los precios de las acciones de Facebook, Inc," REICE Rev. Electrónica Investig. En Cienc. Económicas, vol. 5, no. 10, Art. no. 10, 2017, doi: 10.5377/reice.v5i10.5535.
[2] X.-W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, vol. 2, pp. 514–525, 2014, doi: 10.1109/ACCESS.2014.2325029.
[3] Y. Tang, "Deep Learning using Linear Support Vector Machines," ArXiv13060239 Cs Stat, Feb. 2015, Accessed: May 27, 2020. [Online]. Available: http://arxiv.org/abs/1306.0239.
[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Mach. Learn. PYTHON, p. 6.
[5] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proceedings of the 2012 international conference on Management of Data - SIGMOD '12, Scottsdale, Arizona, USA, 2012, p. 793, doi: 10.1145/2213836.2213958.
[6] B. Panda, J. S. Herbach, S. Basu, and R. J. Bayardo, "PLANET: massively parallel learning of tree ensembles with MapReduce," Proc. VLDB Endow., vol. 2, no. 2, pp. 1426–1437, Aug. 2009, doi: 10.14778/1687553.1687569.
[7] "Diseño Experimental." http://www.virtual.unal.edu.co/cursos/ciencias/2000352/index.html (accessed Jun. 11, 2020).
[8] "Correlacion." http://www.sachile.cl/upfiles/revistas/54e63a1a778ff_15_correlacion-2-2014_edit.pdf (accessed Jun. 12, 2020).
[9] E. Uriel and U. de Valencia, "3 Regresión lineal múltiple: estimación y propiedades," p. 38.
[10] J. M. A. Gómez, R. A. A. Córdova, and Y. A. I. Salinas, "Estimación del factor K en transformadores de distribución usando modelos de regresión lineal," R A, vol. 20, no. 48, p. 13, 2016.
[11] H. Yue, E. G. Jones, and P. Revesz, "Local Polynomial Regression Models for Average Traffic Speed Estimation and Forecasting in Linear Constraint Databases," in 2010 17th International Symposium on Temporal Representation and Reasoning, Paris, France, Sep. 2010, pp. 154–161, doi: 10.1109/TIME.2010.24.
[12] J. E. G. Barajas, "Método de regresión polinomial aplicado a la estimación de la señal de referencia de un sistema de filtrado adaptativo por cancelación para el tratamiento del desplazamiento de la línea de base del electrocardiograma," vol. 1, no. 1, p. 10, 2009.
[13] L. G. Abril, "Modelos de Clasificación basados en Máquinas de Vectores Soporte," p. 20.
[14] E. J. C. Suárez, "Tutorial sobre Máquinas de Vectores Soporte (SVM)," p. 27.