Reconocimiento de Dígitos Manuscritos por Medio de Técnicas de Minería de Datos

Recognition of Manuscript Digits by Media of Data Mining Techniques

C. Paz, J. Ojeda, E. Badillo, J. Bonett & D. Heredia
Universidad Simon Bolivar, Barranquilla
Correo de correspondencia: dianahv@unisimonbolivar.edu.co

Resumen La necesidad de digitalizar documentos escritos de cualquier tipo impone el uso de sistemas de software capaces de interpretar correctamente los caracteres en ellos contenidos; el reto más grande se presenta cuando dichos documentos son manuscritos, pues cada persona tiene características propias de escritura. El presente trabajo es el resultado de un intento de aplicación de técnicas de minería de datos, particularmente árboles de clasificación, para el reconocimiento de dígitos manuscritos; a pesar de no obtener un reconocimiento pleno de este tipo de escrituras, los resultados son interesantes.

Palabras clave: Minería de Datos, Dígitos Manuscritos, Datasets, WEKA, Árboles de decisión.

Abstract The need to digitize written documents of any kind imposes the use of software systems capable of correctly interpreting the characters contained therein; The greatest challenge arises when these documents are manuscripts, because each person has their own writing characteristics. The present work is the result of an attempt to apply data mining techniques, particularlyclassification trees, for the recognition of manuscript digits; despite not obtaining a full recognition of this type of scriptures, the results are interesting.

Key words: Data Mining, Manuscript Digits, Datasets, WEKA, Decision trees.

I. Introducción

La gran influencia de los sistemas de información, los sistemas de sensores, el internet, teléfonos y dispositivos electrónicos "inteligentes", entre otros avances tecnológicos, incrementan la necesidad de tener mecanismos para digitalizar documentos e imágenes y poder así almacenarlos, transmitirlos y procesarlos de alguna forma para aprovechar la información que pueden contener.
El proceso de digitalización consiste en convertir un mensaje, documento, imagen en una sucesión de impulsos electrónicos, que equivalen a los dígitos 0 y 1 combinados (Código binario). Cuando se digitaliza un documento manuscrito se obtiene básicamente su imagen, la cual se puede almacenar y transmitir, pero no procesar o interpretar su contenido. Como humanos, es obvio que al ver la imagen podemos leer y entender su contenido, pero un computador o dispositivo electrónico cualquiera no es capaz de hacerlo. Una dificultad adicional resulta del hecho de que cada persona tiene una forma particular de escribir, sus caracteres manuscritos tienen formas distintas a los de otras personas. La información contenida en estos documentos puede ser muy valiosa en diversos ámbitos, por ejemplo: Digitalización de libros o revistas físicas, en la cual se puedan hacer búsquedas tal y como se hace en uno electrónico; captura de datos desde documentos manuscritos, firmas digitales, entre otros.

II. Generalidades de la minería de datos

La gran influencia de los sistemas de información, los sistemas de sensores, el internet, teléfonos y dispositivos electrónicos "inteligentes", entre otros avances tecnológicos, incrementan la necesidad de tener mecanismos para digitalizar documentos e imágenes y poder así almacenarlos, transmitirlos y procesarlos de alguna forma para aprovechar la información que pueden contener. El proceso de digitalización consiste en convertir un mensaje, documento, imagen en una sucesión de impulsos electrónicos, que equivalen a los dígitos 0 y 1 combinados (Código binario) [1]. Cuando se digitaliza un documento manuscrito se obtiene básicamente su imagen, la cual se puede almacenar y transmitir, pero no procesar o interpretar su contenido. Como humanos, es obvio que al ver la imagen podemos leer y entender su contenido, pero un computador o dispositivo electrónico cualquiera no es capaz de hacerlo. Una dificultad adicional resulta del hecho de que cada persona tiene una forma particular de escribir, sus caracteres manuscritos tienen formas distintas a los de otras personas. La información contenida en estos documentos puede ser muy valiosa en diversos ámbitos, por ejemplo: Digitalización de libros o revistas físicas, en la cual se puedan hacer búsquedas tal y como se hace en uno electrónico; captura de datos desde documentos manuscritos, firmas digitales, entre otros.

De ahí surge un gran reto y es la pregunta que guía la presente propuesta: ¿Cómo interpretar correctamente cada carácter escrito a mano, los dígitos en particular, de tal manera que un dispositivo electrónico sea capaz de "leer" el contenido no

simplemente como imagen, sino como texto? Además, ¿qué tipo de herramientas informáticas y técnicas son necesarias para dicha interpretación?.

Definición.
[2] La minería de datos tiene varias definiciones, veamos algunos a continuación:

"Es el proceso de descubrir nuevas correlaciones, patrones y tendencias, utilizando grandes cantidades de datos almacenados en repositorios, aplicando tecnologías de reconocimiento de patrones, así como herramientas matemáticas y estadísticas."
- *"Es un proceso más amplio que tiene como objetivo el descubrimiento de conocimiento en grandes bases de datos"*
- *"Es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos".*

Para concluir la minería de datos es el proceso de extraer conocimiento a partir de grandes cantidades de datos, mediante el uso de diferentes técnicas de aprendizaje de máquina. La utilidad de la minería de datos ya no se pone en duda, por lo cual esta tecnología está siendo aplicada por muchas herramientas de software.

Técnicas de Minería de Datos.

[3]Las técnicas de minería de datos permiten la extracción de conocimiento. Actualmente, existe un amplio abanico de técnicas de minería de datos que se pueden clasificar en predictivas (las variables se pueden clasificar en dependientes e independientes), descriptivas (se agrupan a partir de características similares) y auxiliares (herramientas de apoyo a la verificación).

Las técnicas predictivas en las que las variables pueden clasificarse inicialmente en dependientes e independientes en base a un conocimiento teórico previo, algunos algoritmos son los de tipo de regresión, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas.
- Las técnicas descriptivas en el que todas las variables tiene inicialmente el mismo estatus o grado de pertenencia. Estas técnicas, se crean automáticamente partiendo del reconocimiento de patrones. Entre este grupo tenemos técnicas de agrupación (clustering), segmentación, reducción de la dimensionalidad, etc.
- Las técnicas auxiliares son herramientas de apoyo superficial y más limitadas. Basadas en técnicas de estadísticas descriptivas, consultas e informes enfocados generalmente a la verificación y presentación.

Clustering.
Consiste en agrupar un conjunto de datos, sin tener clases predefinidas, basándose en la similitud de los valores de los atributos de los distintos datos. Esta agrupación, a diferencia de la clasificación, se realiza de forma no supervisada, ya que no se conoce de antemano las clases del conjunto de datos de entrenamiento. El clustering identifica clusters, o regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [4] El clustering se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters.

Dentro las técnicas de clustering se tiene, el algoritmo K-means, el cual fue creado en 1967 por MacQueen y es el algoritmo de clustering más conocido y utilizado, siendo de simple aplicación y eficaz. La idea básica del algoritmo es obtener los K centros iniciales y formar clusters, asociando todos los objetos de X a los centros más cercanos, después se recalculan los centros. Si esos centros no difieren de los centros anteriores, entonces el algoritmo termina; caso contrario, se repite el proceso de asociación con los nuevos centros hasta que no haya variación en los centros, o se cumpla algún otro criterio de parada como poco número de reasignaciones de los objetos. Para obtener los centroides, se calcula la media o la moda según se trate de atributos numéricos o simbólicos. Las acciones o pasos a seguir son las siguientes:

- Primero se especifica por adelantado cuantos clusters se van a crear, éste es el parámetro k, para lo cual se seleccionan k elementos aleatoriamente, que representarán el centro o media de cada cluster.
- A continuación, cada una de las instancias, ejemplos, es asignada al centro del cluster más cercano de acuerdo con la distancia Euclidiana que le separa de él.
- Para cada uno de los clusters así construidos se calcula el centroide de todas sus instancias y estos centroides son tomados como los nuevos centros de sus respectivos clusters.
- Finalmente se repite el proceso completo con los nuevos centros de los clusters.
- La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración.
- Arboles de decisión. Los árboles de decisión son uno de los algoritmos más sencillos y fáciles de implementar y a su vez de los más poderosos. Este algoritmo genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información, es decir, elige al atributo que mejor clasifica a los datos. [5]

Las características más importantes en el trabajo con árboles de decisiones son la especificación de los criterios para minimizar los costes, la selección del método de división y la elección del tramo del árbol adecuado o problema del sobreajuste [3]
- Redes Neuronales. Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra. [6].

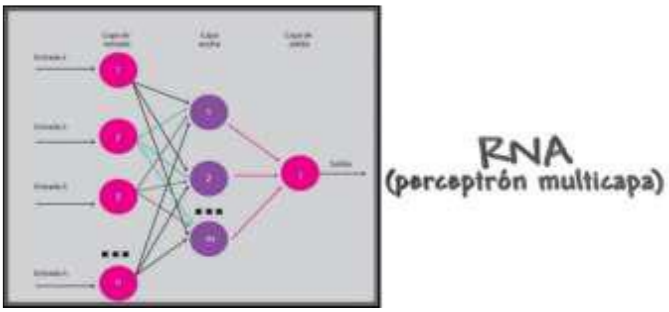


Figura 1. Redes Neuronales.

Redes Bayesianas
[7]Formalmente, una Red Bayesiana es un grafo dirigido cíclico cuyos nodos representan variables y los arcos que los unen codifican dependencias condicionales entre las variables. El grafo proporciona una forma intuitiva de describir las dependencias del modelo y define una factorización sencilla de la distribución de probabilidad conjunta consiguiendo un modelo manejable que es compatible con las dependencias codificadas. Existen algoritmo eficiente para aprender modelos gráficos probabilísticos a partir de datos, permitiendo así la aplicación automática de esta metodología en problemas complejos. Las Redes Bayesianas que modelizan secuencias de variables (por ejemplo, series temporales de observaciones) se

denominan Redes Bayesianas Dinámicas. Una generalización de las Redes Bayesianas que permiten representar y resolver problemas de decisión con incertidumbre son los Diagramas de Influencia.

Pre procesamiento de Datos.

"El Pre procesamiento de Datos" o "La Preparación de Datos" engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas con más completitud, etc.) [8].

Outliers.
Los Outliers, también conocidos como anormalidades en los datos, son observaciones dentro de una secuencia de datos que es anómala respecto a la conducta presente en la mayoría de las observaciones dentro del mismo conjunto de datos. La existencia de outliers es una complicación común en los análisis de datos y es necesario tomar ciertas medidas sobre ellos para evitar generar sesgos en el análisis.

MATLAB: Es un entorno de computación y desarrollo de aplicaciones totalmente integrado. Siendo útil para proyectos con elevados cálculos matemáticos y la visualización gráfica.

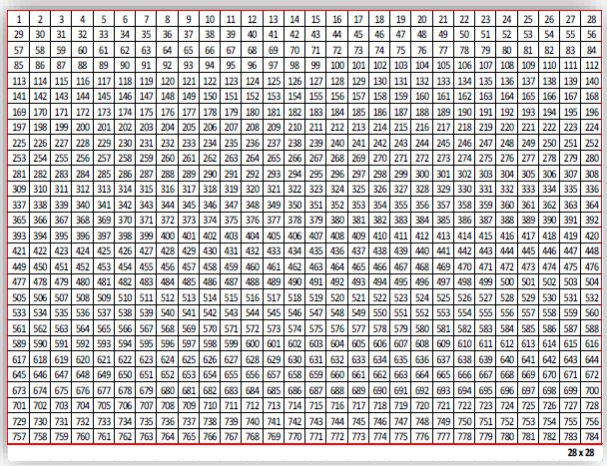
IBM SPSS Modeler: Es un producto de la empresa IBM SPSS, que, mediante un entorno gráfico, permite aplicar técnicas de minería de datos para descubrir patrones y tendencias en datos estructurados o no estructurados. En este sentido, se puede visualizar gráficamente el proceso llevado a cabo. Además, provee de una serie de módulos que permiten soportar un análisis con grandes volúmenes de datos.

SAS Enterprise Miner: Está basada en la metodología SEMMA (Sample, Explore, Modify, Model, Assess), y agiliza el proceso de minería de datos para crear modelos predictivos y descriptivos de alta precisión para grandes volúmenes de datos.
- Salford Systems Data Mining: De la empresa Salford Systems, la cual es especializada, entre otras tareas, en la elaboración de software de minería de datos y consultoría.
- Orange: Permite el pre-procesamiento de datos, características de puntuación y filtrado, modelado, evaluación del modelo y técnicas de exploración.
- Oracle Data Mining (ODM): Es una herramienta desarrollada por la empresa Oracle para aplicar técnicas de minería de datos a grandes volúmenes de datos.
- Rapid Miner: Es una herramienta que forma parte del proyecto Rapid-i. Cuenta con dos componentes RapidMine para operadores de minería de datos y RapidAnalytics que permite trabajo colaborativo, escalable y concurrente de múltiples usuarios.

WEKA: Es un conjunto de librerías JAVA para la extracción de conocimiento desde bases de datos. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de pre-procesado, clasificación, agrupamiento, asociación y visualización. Debido a que es de código abierto, permite a su vez la fácil integración a través de su API de desarrollo y la gran aceptación por la comunidad científica.

III. Construcción del calsificador de dígitos manuscritos

En el manejo de la información, con el fin de lograr identificar los números manuscritos se consiguió un conjunto de datos (dataset) que consta de 42.000 datos en un archivo .csv, cada número manuscrito digitalizado, que llamaremos etiqueta, está en una matriz cuadrada de 28x28, es decir, cada etiqueta cuenta con 784 bits.



Los datos de la matriz original van de 0 hasta 255 que es la escala de grises, dependiendo a la fuerza trazada, en cada bits se obtiene un valor, siendo 0 el blanco absoluto (no se trazó nada ese bits) y 255 negro absoluto (se usó todo el bits para el dibujo del número)
Al cargar los 42.000 datos en el software WEKA este no soporto dicho cargue; razón por la cual se buscó reducir la matriz.
Reducción de la Matriz: El problema original es la gran cantidad de datos así que lógicamente se pensó en reducirlos pero para esto habría que reducir la matriz sin perder la lógica. Se Intentó reducir la matriz de 28x28 a 7x7 para que en vez de tener 784 bits por etiqueta tuviera 49 y le fuera más fácil a WEKA dar un resultado. Estos primeros 16 bits serán nuestro primer bit en la matriz reducida, al terminar la reducción de la matriz nos quedó algo así (figura 1)

Figura1-matriz reducida
Se calculó el promedio por cada submatriz de 4x4 para tener un aproximado real de la matriz original y este promedio será el valor que se coloque en la primera posición y así hasta llegar a la posición 49 (los números en la matriz de la imagen son Bposiciones y no los valores de la digitalización). En este punto todo estaba bien pero al momento de pasar estos datos a WEKA se puede percibir se estaba perdiendo mucha información por eso se decidió no utilizar esta opción y binarizando la matriz. Binarización de Matriz: Si bien es cierto el paso anterior no dio la respuesta que se esperó no todo fue malo, ya que la idea de reducción con esa lógica era buena solo había que buscar un mejor camino, por eso se decidió intentar la binarización. Teniendo en cuenta los valores obtenidos en la matriz anterior lo que se hizo fue tomar este promedio obtenido y si era mayor a 127 se colocaría 1 al bit y si era menor a 127 el valor del bit seria 0, de esta manera esta nueva matriz solo quedaría con 1 y 0.
Una vez evaluado, reducido y binarizado la matriz original intentaremos realizar otro procedimiento el cual consta de sumar las líneas horizontales, las líneas verticales, las diagonales izquierdas y las diagonales derechas, para la muestra y desarrollo de esta documentación se redujo la cantidad de tuplas a 500.
Dichas sumas nos dará una nueva matriz la cual es apta para una lectura óptima por parte de WEKA. Árbol J48: Resultado obtenido, en el siguiente caso WEKA informa que de las 500 tuplas sólo clasificó correctamente el 49.6% por el árbol J48 e incorrecto el 50.4% por el mismo árbol J48 (figura 2)

Clasificación Correcta: 49.6
Clasificación Incorrecta: 50.4

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49

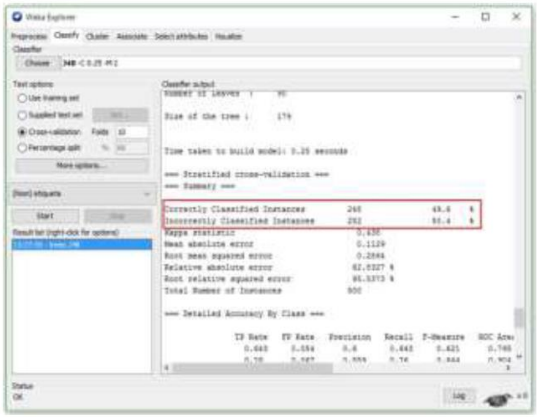


Figura2-resultado obtenido árbol j48

Adicionalmente podemos observar la matriz de confusión (imagen 3), antes de verla se explicará brevemente, esta matriz clasifica con letras los números, es decir, el 0 lo clasifica como a, el 1 como b, el 2 como c y así sucesivamente.
La diagonal señalada indica que para la columna a (numero 0), clasificó correctamente 36 números 0, 38 números 1 los clasifico correctamente, etc.

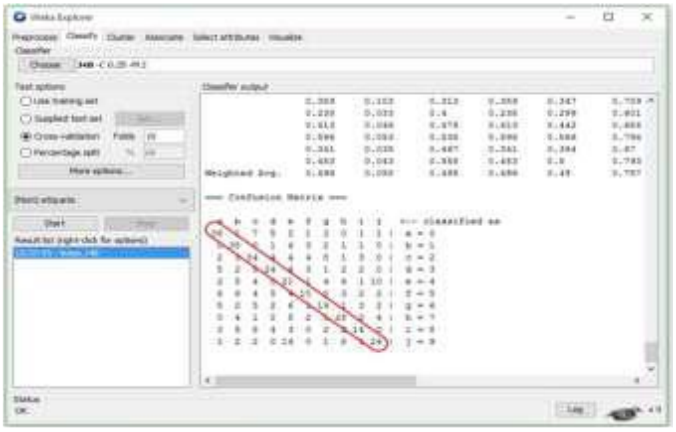


Figura3-matriz de confusión para j48

RandomForest: Resultados obtenidos tanto los porcentajes de efectividad como la matriz de confusión (figura 4 y 5)

Clasificación Correcta: 52.6
Clasificación Incorrecta: 47.4

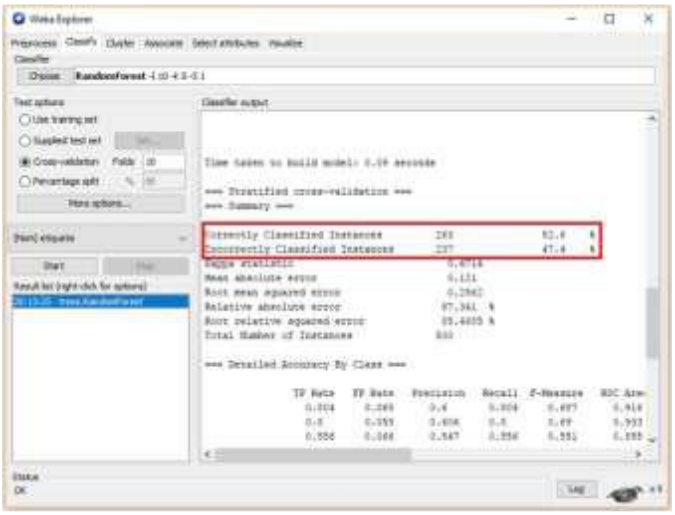


Figura4-resultado obtenido RandomForest

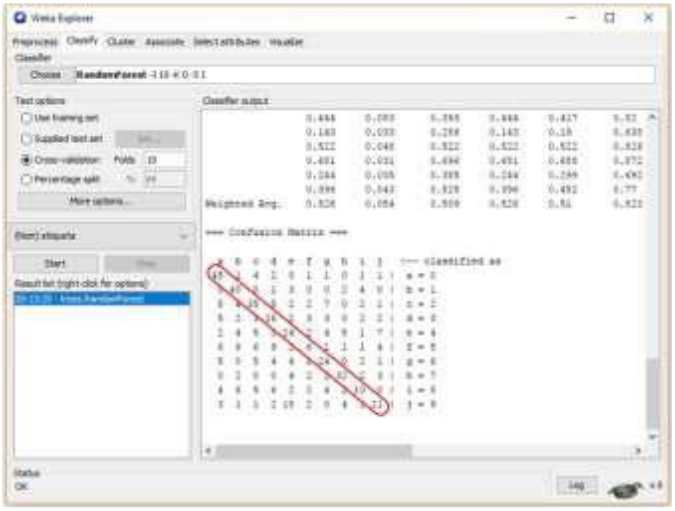


Figura5-matriz de confusión para RandomForest

Al comprar los resultados obtenidos en el árbol J48 y en el RandomForest podemos observar que sus resultados no son muy diferentes, su margen de error es pequeño, aunque en teoría el RandomForest tiene que ser mucho más efectivo que el árbol J48.

IV. Conclusión

Una vez se ha realizado el trabajo de aplicar técnicas de minería de datos para identificar o clasificar dígitos manuscritos, se evidencia que ellas logran resultados aceptables, más no ideales para la correcta interpretación y digitalización eficiente de textos manuscritos [13-15]. Sin embargo, se obtuvieron resultados interesantes, como los siguientes:

- Los árboles obtenidos por Random Forest son más eficientes en la clasificación que los obtenidos por J48., como se puede observar en la figura a continuación.
- Se logró mejor porcentaje de tuplas bien clasificadas (Dígitos correctamente identificados) con los datos originales, dejando que la herramienta WEKA realice los pre-procesamientos adecuados, como discretización.
- Como ocurre con los humanos, los modelos obtenidos por minería de datos, confunden en gran medida los dígitos manuscritos 4 y 9, 5 y

0, 5 y 6, entre otros. Esto puede deberse a las características propias de la escritura de cada individuo.

V. Referencias bibliográficas

[1] L. C. Rodríguez, «MINERIA DE DATOS,» España, 2015.

[2] D. S. G. CESAR PEREZ LOPEZ, Minería de datos. Técnicas y herramientas, España: Paraninfo, 2007.

[3] J. H. P. S. Y. Ming-Syan Chen, Data mining: An Overview From a Database Perspective, Estados Unidos: IEEE, 1996.

[4] A. S. V. M. G. A. Sergio Valero Orea, «Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos,» *Revista de la Universidad Tecnológica de Izúcar de Matamoros*, p. 8, 2009.

[5] T. ALUJA, de *LA MINERÍA DE DATOS, ENTRE LA ESTADÍSTICA Y LA INTELIGENCIA ARTIFICIAL*, Catalunya, 2001, pp. 479 - 498.

[6] J. B. J. S.-M. D. Gutiérrez, «Minería de Datos. Redes Bayesianas y Neuronales,» Grupo de Meteorología de Santander, España, 2011.

[7] C. Z. Q. Y. S. Zhang, «Data preparation for data mining,» *Special Issue Data Cleaning and Preprocessing*, pp. 17:5-6, 375-381, 2003.

[8] «Multimedia,» [En línea]. Available: http://www2.udec.cl/~lsalazarv/digitalizacion.html.

[9] O. B. B. M. O. Piragauta, «RECONOCIMIENTO ÓPTICO DE NÚMEROS ESCRITOS A MANO USANDO FUNCIONES DE BASE RADIAL Y SISTEMA MEMÉTICO DIFERENCIAL,» 2014.

[10] L. Rokach y O. Maimon, Data mining with decision trees - Theory and Applications, World Scientific, 2015.

[11] N. Bhargava, G. Sharma, R. Bhargava y M. Mathuria, «Decision Tree Analysis on J48 Algorithm for Data Mining,» *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, nº 6, pp. 1114 - 1119, 2013.

[12] N. Rodríguez, A. Jiménez, L. Reyes y B. Suarez, «Data mining: a scholar dropout predictive model,» de *Humanitarian Technology Conference (MHTC), IEEE Mexican*, 2017.

[13] M. Molina Cárdenas, P. Pedroza Barrios, K. M. Gaitán Moreno, J. F. Salgado Arismendy y M. C. Ordóñez Ávila, «Diseño y Construcción del Prototipo de un Brazo Robótico con Tres Grados de Libertad, como Objeto de Estudio,» Ingeniare, vol. 10, nº 18, pp. 87-94, 2015.

[14] P.A. Sánchez-Sánchez and J.R. García-González, "A New Methodology for Neural Network Training Ensures Error Reduction in Time Series Forecasting", Journal of Computer Science, 13 (7), pp. 211.217, 2017. DOI : 10.3844/jcssp.2017.211.217

[15] R. Feo, "Epistemología y práctica de la investigación sobre el aprendizaje estratégico en América Latina", *Revista Educación y Humanismo*, vol. 17(29), 220-235., 2015.