Congr. Int. en Ing. Electrónica. Mem. ELECTRO, Vol. 41, pp. 108-113, Oct 2019, Chihuahua, Chih. México
http://electro.itchihuahua.edu.mx/memorias_electro/MemoriaElectro2019.zip
ISSN 1405-2172

CLASIFICACIÓN DE DATOS NO BALANCEADOS EN ESPECTROSCOPIA DE EMISIÓN ÓPTICA "ELECTRO 2018"

Rosales Martínez Octavio
Universidad Autónoma del Estado de México
Kilómetro 60 carretera Toluca-Atlacomulco, Atlacomulco, Estado de México
(712) 122 04 46
tavo@hotmail.com

RESUMEN.

La espectroscopia óptica de emisión es una técnica en la que se caracterizan especies de elementos mediante de la identificación y clasificación de líneas vistas a través de en un espectro adquirido. Este procedimiento está acompañado de problemas en la adquisición de los datos, que produce un desplazamiento óptico y un fondo continuo en el espectro. Este trabajo se centra en la corrección del desplazamiento óptico utilizando técnicas regresión lineal, y en una comparativa de algoritmos de clasificación, alcanzado una métrica F1 promedio de hasta 99% con datos sobremuestreados. Además, como conclusión se exponen una comparativa de los algoritmos implementados con datos balanceados y no balanceados. Finalmente, el trabajo presentado tiene aplicación en el área de espectroscopia óptica, para la caracterización de especies de elementos como mercurio, argón, oxígeno y nitrógeno.
Palabras Clave: espectroscopia óptica, desplazamiento óptico, regresión lineal, árbol de decisión, métrica F1.

ABSTRACT.

Emission optical spectroscopy is a technique in which species of elements are characterized by the identification and classification of lines seen through an acquired spectrum. This procedure is accompanied by problems in the acquisition of the data, which produces an optical shift and a continuous background in the spectrum. This work focuses on the correction of optical displacement using linear regression techniques, and in a comparison of classification algorithms, achieved an average F1 metric of up to 99% with oversampled. In addition, as a conclusion, a comparison of the algorithms implemented with balanced and unbalanced data is presented. Finally, the presented work has application in optical spectroscopy, for the characterization of species of elements such as mercury, argon, oxygen and nitrogen.
Keywords: optical spectroscopy, optical displacement, linear regression, decisión tree, F1 metric.

1. INTRODUCCIÓN

La espectroscopia óptica de emisión es un método que permite capturar la radiación luminosa de un plasma para determinar las especies o líneas que componen un elemento, a este procedimiento se le conoce como caracterización [1]. En esta técnica los datos con los que se cuenta son longitud de onda e intensidad, adquiridos de una corrida o ejecución experimental con una lámpara de calibración de mercurio y argón modelo HG-1 de Ocean Optics para un rango de longitud de onda de 177nm a 891nm, así como las especies de elementos reportados por el Instituto Nacional de Estándares y Tecnología (NIST por sus siglas en inglés).

Trabajos encontrados en la literatura se centran en la caracterización de especies con técnicas de Machine Learning. Los estudios realizados presentan el análisis de espectros de rayos gamma mediante el uso de redes neuronales artificiales para el análisis de 28 radioisótopos. Esta red neuronal tiene 47 neuronas de entrada, 52 neuronas en la capa oculta y 28 neuronas de salida, los resultados presentados muestran la influencia del ruido en la identificación de especies y la aplicación de la segunda derivada para realizar la detección de picos [2].

Las técnicas de Machine Learning de este trabajo se enfocan en el aprendizaje supervisado con algoritmos de clasificación, alcanzando predicciones altas, en datos con características similares. Las especies estudiadas son: N 7 (Nitrógeno), O 8 (Oxígeno), así como Ar 18 (Argón) y Hg 80 (Mercurio), Ar 18 y Hg 80. Estos son los elementos que componen una lámpara de calibración HG-1 y cuyas longitudes de ondas de las especies son conocidas, y utilizados como datos de corrección de desplazamiento óptico y prueba. Los datos que se usaron para entrenamiento son las especies de elementos reportadas por el NIST, en las cuales la longitud de onda por cada especie también se conoce, con la desventaja de que los datos no están balanceados.

La caracterización manual de especies es un procedimiento que puede llevar desde horas hasta días, por esta razón se usan técnicas de Machine Learning para agilizar esta acción. En este trabajo se detalla el procedimiento empleado para corregir el desplazamiento óptico, además de una comparativa de algoritmos de Machine Learning para caracterizar automáticamente especies de espectroscopia de emisión óptica en datos no balanceados. El lenguaje de programación utilizado para implementar la solución es Python, porque es simple, rápido, portable, y porque es el lenguaje empleado por empresas que se dedican a la investigación en el campo de la

inteligencia artificial como Intel y Google. Además, se utilizaron bibliotecas con funciones específicas en el desarrollo de software científico como son: pandas, numpy, sklearn, scipy, matplotlib, seaborn entre otras [3, 4].

2. METODOLOGÍA

El procedimiento desarrollado, que inicia con la adquisición de datos hasta caracterización automática de especies se muestra en la Figura 1.

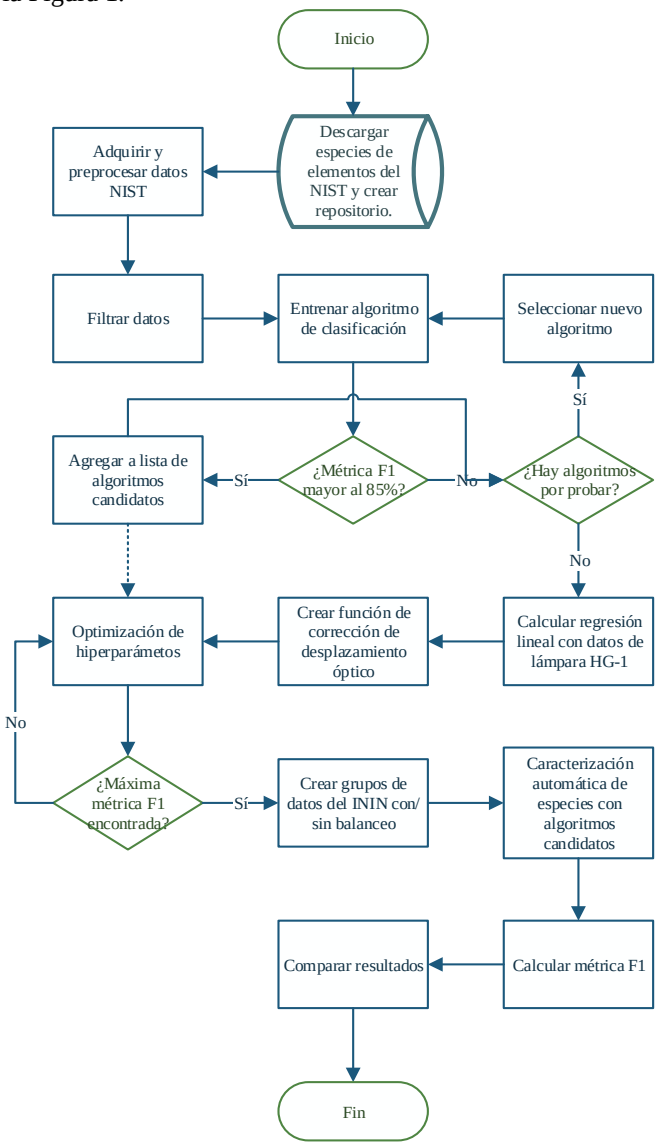

Figura 1. Diagrama de flujo para le metodología empleada en este trabajo.

A continuación, se detallan los principales procesos que permitieron dar solución a la caracterización automática de especies con técnicas de Machine Learning.

2.1. Adquisición y tratamiento de datos.

En primer lugar, se creó un repositorio local con las líneas reportadas en el NIST, para esto se descargaron de éste; 99 archivos que corresponden a los primeros 99 elementos, cada archivo contiene de 12 a 20 columnas de tipo numérico y cadena, de las cuales solo 3 son de interés; la primera columna es "element" de tipo cadena, que indica el elemento al que corresponde toda la información de la columna y es la clase a predecir. La segunda columna es "sp_num" de tipo entero, que indica los niveles de ionización en un rango del 1 al 72, la tercera columna es la longitud de onda observada de tipo flotante en un rango del 0.039639nm a 60,640,000,000nm, con la peculiaridad de que el encabezado "obs_wl_vac(nm)" se intercala a lo largo de los archivos con "obs_wl_air(nm)". La diferencia radica en que "obs_wl_vac(nm)" es de muestras tomadas en condiciones al vacío, mientras que "obs_wl_air(nm)" es para las muestras que se tomaron a presión atmosférica local, por este motivo se agregó una columna numérica nombrada "tipo" que permite diferenciar con un **0** las muestras tomadas al vacío y con un **1** las muestras tomadas a presión atmosférica local. Esta información se condensa en un único archivo tipo CSV (Comma Separated Values) dando un total de 207,991 especies, a los que se les eliminó los caracteres **"=** que acompañaban a datos numéricos. Estos archivos fueron utilizados como datos de entrenamiento.

Los datos experimentales de la espectroscopía de emisión óptica fueron proveídos por el Instituto Nacional de Investigaciones Nucleares (ININ) mediante tres corridas experimentales de una lámpara de calibración HG-1 de Ar 18 y HG 80, cada una con 33 archivos con extensión .dat, cada uno de estos archivos contiene 1024 filas de datos sin etiquetar, la primer columna es la longitud de onda que tiene un rango de 177nm a 891nm, y la segunda columna corresponde a la intensidad y la cual sirve para identificar picos de una especie respecto a otra Estos archivos se usaron con dos finalidades: corregir el desplazamiento óptico y ajustar los parámetros de los algoritmos.

2.2. Selección de algoritmos de Machine Learning.

La selección de algoritmos de Machine Learning se basó en técnicas de clasificación para aprendizaje supervisado con uso de la biblioteca sklearn; como datos de entrenamiento se usó el repositorio local obtenido de las especies reportadas por el NIST. Posteriormente se emplearon esos mismos datos como prueba y se seleccionaron aquellos algoritmos con una métrica F1 mayor al 40%. De los 4 algoritmos que se obtuvieron, el primero que fue objeto de estudio es k-Nearest Neightbors [5, 6], el cual requiere que el número de categorías k sea conocido a priori, la heurística de este algoritmo se puede resumir con el siguiente procedimiento:

1. Iniciar con un conjunto de datos dónde las categorías sean conocidas a priori.
2. Agregar una muestra cuya categoría es desconocida.
3. Calcular la distancia de la muestra respecto a otros puntos.
4. Clasificar la muestra asignándola a la categoría con la distancia más corta.

El siguiente algoritmo es el denominado Decision Tree [7, 8, 9]. En este, el nodo raíz y los nodos internos se autoconstruyen a partir de la pureza de cada característica, y las respuestas con los nodos hoja. Este algoritmo funciona con datos categóricos y numéricos; sin embargo, se debe tener cuidado en limitar el número de características y controlar la profundidad del árbol para evitar el sobre ajuste.

Posteriormente se estudió el algoritmo Random Forest. Este algoritmo tiene la peculiaridad de construir conjuntos de datos con muestras aleatorias tomadas del conjunto de datos original; se construye un árbol decisión con un subconjunto de características en cada nodo. Este proceso se repite hasta crear un conjunto de árboles al que se denomina bosque, después, cada muestra se pasa como entrada a cada árbol generado, se cuentan las respuestas y se muestra aquella que apareció mayoritariamente [10].

Finalmente se estudió también el algoritmo AdaBoost. En este, cada árbol consta únicamente de un nodo y 2 hojas, al conjunto de todos los árboles se le conoce como “forest of stumps” y el orden de construcción empieza de manera descendente desde el árbol con la pureza más alta [11, 12].

2.3. Corrección de desplazamiento óptico.

Los datos provenientes de la lámpara de calibración tienen un desplazamiento óptico inducido por el monocromador, esto provoca que los picos de un espectro aparezcan en otra posición, lo que provoca predicciones incorrectas. En la Figura 2 se observan 2 especies del elemento Ar 18 en las longitudes de onda 750.3869nm y 763.5106nm, el pico a la derecha con la línea punteada corresponde al espectro del archivo 25 de la lámpara de calibración sin aplicar algún ajuste, el pico a la izquierda con la línea continua es el mismo pico de la izquierda cuando se aplica la corrección de desplazamiento óptico, es de resaltar que la especie de Ar 18 reportada en el NIST corta los picos aproximadamente a la mitad, lo que facilita su predicción.

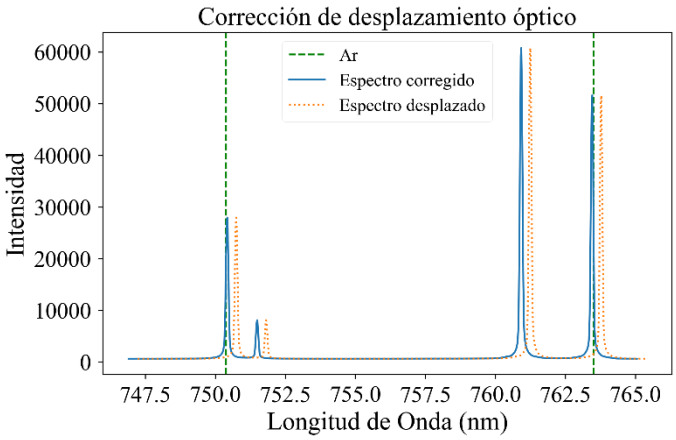

Figura 2. Corrección de desplazamiento óptico en espectro de lámpara HG-1

La corrección de desplazamiento óptico se realizó de la siguiente manera:

a) Detectar los picos de mercurio y argón en los archivos provenientes de la lámpara de calibración.
b) Calcular la diferencia (yD) entre las especies Ar 18 y HG 80 reportadas en el NIST (xNIST) y los picos de las especies de la lámpara HG-1 (xHG1)
$yD_i = xNIST_i - xHG1_i$
c) Calcular regresión lineal considerando, $y_i = yD_i$ y $x_i = xNIST_i$:

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i x_i - \left(\sum_{i=i}^{n} x_i\right)^2} \quad (1)$$

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i x_i - \left(\sum_{i=i}^{n} x_i\right)^2} \quad (2)$$

$$y_i = b + m x_i \quad (3)$$

Dónde:

b es la intersección.
m es la pendiente.

d) La grafica resultante se muestra en la Figura 3

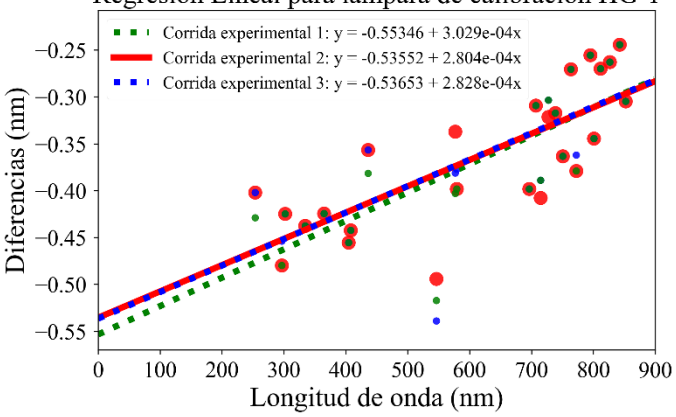

Figura 3. Regresión lineal para cada corrida experimental de la lámpara de calibración HG-1.

e) Las intersecciones b y las pendientes m de las corridas experimentales 2 y 3 son similares, por lo tanto, se calcula el promedio para obtener la siguiente ecuación:
$y_i = -0.536024 + 2.815753^{-4} x_i$ (4)

f) Reemplazar cada pico detectado x_i en la ecuación 4 para posteriormente predecir con los modelos de Machine Learning.

2.4. Optimización de hiperparámetros.

Un problema recurrente en los árboles de decisión es el sobreajuste porque el algoritmo busca en cada nodo disminuir la entropía (grado de dispersión de datos) hasta 0. Por esto se grafica profundidad contra precisión, con el objetivo de buscar aquella profundidad en la que la predicción en los datos de prueba y entrenamiento ya no mejora [13]. En la Figura 4 se observa que a partir de la profundidad en 22 para Random Forest y 24 para Decision Tree, la métrica F1 ya no mejora, por lo tanto, esta es la profundidad buscada [14].

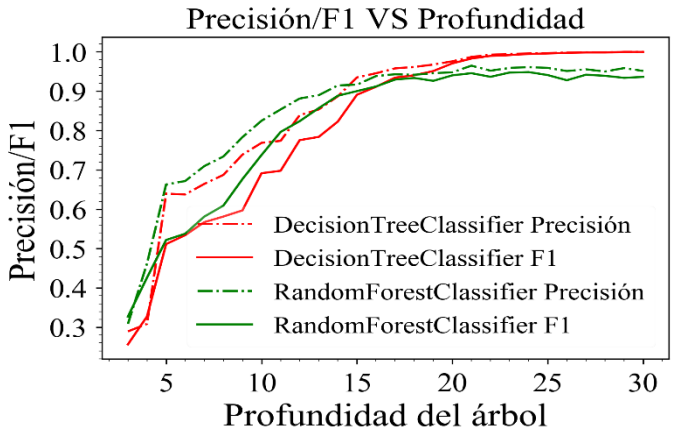

Figura 4. Comparativa de Precisión VS Profundidad para los algoritmos Decision Tree y Random Forest.

2.5. Balanceo de datos.

El balanceo de datos consiste en equilibrar la cantidad de atributos en una clase, de no hacerlo, un modelo de Machine Learning podría tener una tendencia sobre un atributo mayoritario, lo que podría desembocar en un rendimiento pobre [9]. En la Figura 5 se muestran las 4 especies por elemento correspondientes a la lámpara de calibración y se observa que Ar 18 es la clase mayoritaria con 1,859 especies, mientras que Hg 80, O 8 y N 7 son las clases minoritarias.

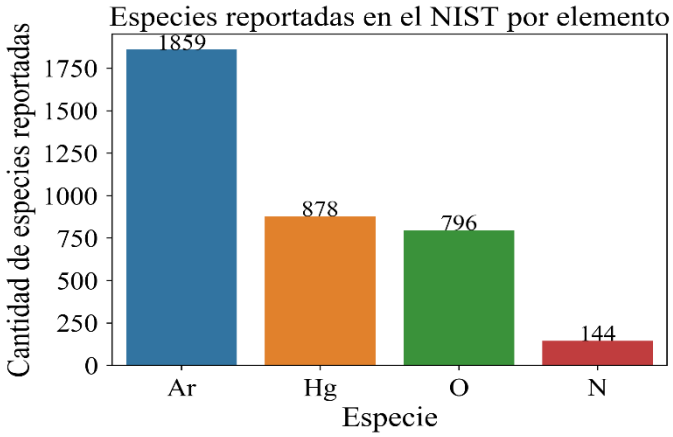

Figura 5. Especies reportadas en el NIST para los elementos Ar 18, Hg 80, O 8 y N 7, correspondiente a los grados de ionización 1 y 2, observados al vacío y a presión atmosférica local, con una longitud de onda comprendida en el rango de 177nm a 891nm.

El desbalanceo de datos se trató de la siguiente manera: a) se obtuvo un conjunto de datos mediante sobremuestreo para las especies de los elementos Hg 80, O 8, N 7, duplicando líneas espectrales al azar hasta alcanzar las 1,859 e igualar así las especies de Ar 18, y posteriormente se determinó la métrica F1 obtenida con los algoritmos de clasificación con el conjunto de los datos desbalanceados; b) se obtuvo otro conjunto mediante submuestreo de los elementos Ar 18, Hg 80, O 8, hasta llegar a las 144 especies e igualar así las especies del elemento N 7, eliminando al azar especies de estos elementos, y se obtuvo la métrica F1 con los datos desbalanceados, y c) se compararon las métricas obtenidas de ambos conjuntos. La Figura 6 se observa que el sobremuestreo tiende a mejorar la métrica F1 en todos los algoritmos estudiados, por lo tanto, este ajuste será parte de la solución.

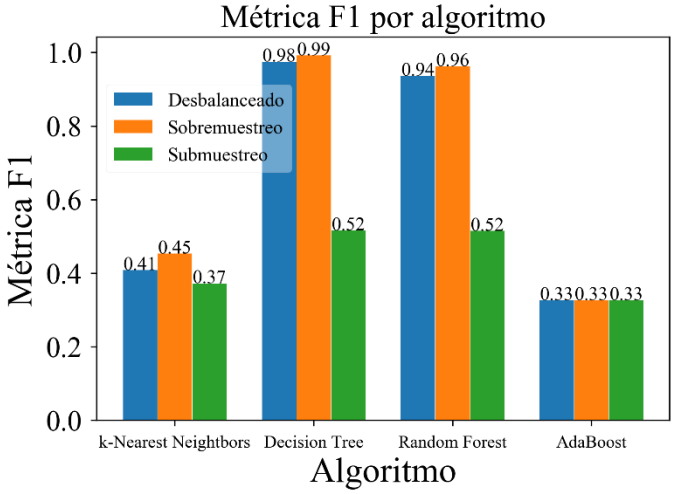

Figura 6. Métrica F1 para cada algoritmo estudiado en condiciones de con datos desbalanceados a los que se les aplicó sobremuestreo y submuestreo

2.6. Caracterización automática de especies.

La caracterización automática de especies se compone de las siguientes etapas: a) integrar los bloques de un espectro en un Data Frame, b) aplicar corrección de desplazamiento óptico y detectar los picos, c) aplicar técnicas de balanceo de datos y almacenarlas en Data Frames independientes, d) entrenar modelos de Machine Learning con los hiperparámetros óptimos encontrados y los datos de entrenamiento, e) usar datos de prueba con sobremuestreo para predicciones, f) graficar y etiquetar predicciones, g) calcular métrica F1 sí las etiquetas son conocidas. En la Figura 7, con círculos se representan los picos detectados, y la etiqueta indica el elemento predicho.

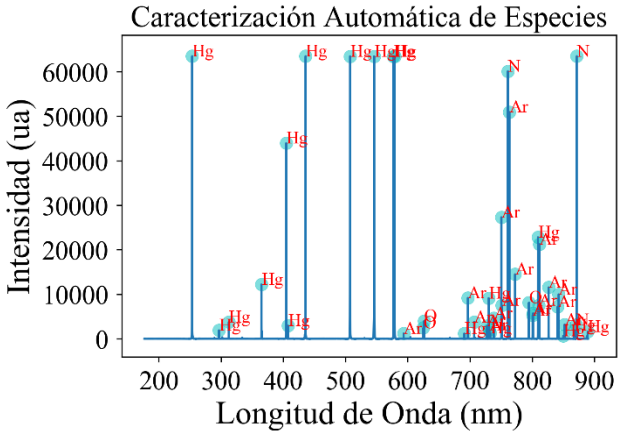

Figura 7. Caracterización automática de especies para la corrida experimental número 2 de la lámpara de calibración HG-1.

3. RESULTADOS.

La caracterización automática de especies se realizó con el algoritmo Decisión Tree, sobremuestreo de datos y corrección de desplazamiento óptico, el conjunto de datos de prueba corresponde a una lámpara de calibración modelo HG-1, se sabe que las especies de Hg 80 y Ar 18 son mayoría, el resto son especies que no vienen descritas en el manual y que el clasificador identificó como N 7 y O 8. Las especies de la Figura 7 se buscó una a una en las líneas reportadas en el NIST y se determinó que 37 de 38 especies son correctas, con lo que se alcanza una precisión del 97.36%

3.1. Conclusiones y trabajos a futuro.

Para el problema particular que se está estudiando, que consiste en la caracterización automática de líneas espectrales resultantes de una espectroscopía de emisión óptica en un plasma frío con técnicas de machine learning, se propone como mejor solución el uso de árboles de decisión, con datos previamente tratados mediante sobremuestreo y corrección de desplazamiento óptico con regresión lineal.

Se buscarán otros algoritmos de regresión que permitan corregir el desplazamiento óptico, ya que, aunque el utilizado funciona, presenta el inconveniente de presentar mucha sensibilidad a los valores atípicos que influyen en la pendiente y el coeficiente.

La optimización de hiperparámetros está planeada en una segunda etapa con la implementación de un algoritmo genético y validación cruzada, con la finalidad de aumentar la precisión en las predicciones.

Se tiene contemplando buscar más algoritmos, afinarlos y usarlos como predictores por votación para mostrar la etiqueta predictora junto con un porcentaje de certeza en la predicción.

Referencias

[1] E. Restrepo y A. Devia, «Caracterización de materiales utilizando espectroscopia óptica de emisión,» *Revista Colombiana de Física*, vol. 34, nº 2, pp. 478-483, 2002.

[2] E. Yoshida, K. Shizuma, S. Endo y T. Oka, «Application of neural networks for the analysis of gamma-ray spectra measured with a Ge spectrometer,» *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 484, nº 1-3, pp. 557-563, 2002.

[3] T. Oliphant, «Python for Scientific Computing,» *Computing in Science & Engineering*, vol. 9, pp. 10-20, 2007.

[4] R. Kumar, «Future For Scientific Computing Using Python,» *International Journal of Engineering* , vol. 2, nº 1, pp. 30-41, 2015.

[5] M.-L. Zhang y Z.-H. Zhou, «A k-nearest neighbor-based algorithm for multi-label classification,» *GrC*, vol. 5, pp. 718-721, 2005.

[6] S. A. Dudani, «The distance-weighted k-nearest-neighbor rule,» *IEEE Transactions on Systems, Man, and Cybernetics*, nº 4, pp. 325-327, 1976.

[7] W. Lui, S. Chawla, D. A. Cieslak y N. V. Chawla, «A robust decision tree algorithm for imbalanced data sets,» *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 766-777, 2010.

[8] J. R. Quinlan, «Induction of decision trees,» *Machine learning*, vol. 1, nº 1, pp. 81-106, 1986.

[9] D. Cieslak y N. V. Chawla, «Learning Decision Trees for Unbalanced Data.,» *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, vol. I, nº 5211, pp. 241-256, 2008.

[10] L. Breiman, «Random Forest,» *Machine learning*, vol. 45, nº 1, pp. 5-32, 2001.

[11] Y. Sun, M. S. Kamel, A. K. Wong y Y. Wang, «Cost-sensitive boosting for classification of imbalanced data,»

Pattern Recognition, vol. 40, nº 12, pp. 3358-3378, 2007.

[12] Y. Freund y R. E. Schapire, «Experiments with a new boosting algorithm,» *icml*, vol. 96, pp. 148-156, 1996.

[13] M. Sebban, R. Nock, J. H. Chauchat y R. Rakotomalala, «Impact of learning set quality and size on decision tree performances,» *IJCSS*, vol. 1, nº 1, pp. 85-105, 2000.

[14] N. Patel y S. Upadhyay, «Study of various decision tree pruning methods with their empirical comparison in WEKA,» *International journal of computer applications*, vol. 60, nº 12, pp. 20-25, 2012.

[15] V. Van Asch, «Macro-and micro-averaged evaluation measures [[basic draft]],» *Belgium: CLiPS*, pp. 1-27, 2013.