Congr. Int. en Ing. Electrónica. Mem. ELECTRO, Vol. 41, pp. 171-176, 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

Rosales-Martínez Octavio[1,], Flores-Fuentes A. A.[1], Granda-Gutiérrez E. E.[1], Mercado-Cabrera A.[2], Alejo E. R.[3], Mendoza-Méndez R.[4], y García-Mejía J. F.[1]*

[1]Universidad Autónoma del Estado de México/C.U. Atlacomulco, km 60 carr. Toluca-Atlacomulco, Atlacomulco, Edo. Méx. C.P. 50450
[2]Instituto Nacional de Investigaciones Nucleares/km 70 carr. Toluca-México, Ocoyoacac, Edo. Méx. 52750.
[3]Instituto Tecnológico de Toluca/TecNM, Av. Tecnológico 100, La Virgen, Metepec, Edo. Méx. 52149.
[4]Universidad Autónoma del Estado de México/C.U. Temascaltepec. Km. 67.5 Carr. Toluca-Tejupilco, Temascaltepec, Edo. Méx. 51300
*tavo_rmx@hotmail.com; orosalesm001@alumno.uamex.mx

RESUMEN.

La espectroscopia óptica de emisión es una técnica de caracterización de especies producidas por los elementos químicos, mediante la identificación y clasificación de líneas vistas a través de en un espectro adquirido. Este procedimiento está acompañado de algunos problemas en la adquisición de los datos, como resultado, se presenta un desplazamiento óptico y un fondo continuo en el espectro. Este trabajo se enfoca no sólo en la corrección del desplazamiento óptico mediante el uso de técnicas de regresión línea, sino también en realizar una comparación de los algoritmos de clasificación; *k-Nearest Neightbors*, *Decision Tree, Random Forest* y *AdaBoost,* para identificar de manera automática las especies, ^{7}N (Nitrógeno), ^{8}O (Oxígeno), ^{18}Ar (Argón) y ^{80}Hg (Mercurio). Los resultados presentados con datos balanceados y no balanceados, establecen que el uso del algoritmo de clasificación nombrado *Decision Tree,* es el que presenta un mejor desempeño para la caracterización automática de especies. Además, los resultados de desempeño muestran un valor de hasta 99% utilizando la métrica F1 con datos sobremuestreados.

Palabras Clave*:* Espectroscopia óptica, desplazamiento óptico, regresión lineal, árbol de decisión, métrica F1.

ABSTRACT.

Optical emission spectroscopy is a technique for characterization of species produced by chemical elements by means of the identification and classification of their spectral lines. Some issues during the data acquisition causes the optical shifting and the presence of a continuous background in the spectra. This work focuses on the correction of optical displacement through the implementation of linear regression techniques. Moreover, a comparison of the classification algorithms: *k-Neighbors, Decision Tree, Random Forest and AdaBoost,* is presented, in order to automatically identify the species of ^{7}N (nitrogen), ^{8}O (oxygen), ^{18}Ar (Argon) and ^{80}Hg (Mercury). Results from using both: balanced and unbalanced data, determinate that the implementation of Decision Tree classification algorithm presents the better performance for the automatic characterization of species. Also, the performance results show a value up to 99% using the F1 metric with oversampled data.

Keywords: Optical spectroscopy, optical displacement, linear regression, decision tree, F1 metric.

1. INTRODUCCIÓN

La espectroscopia óptica de emisión es un método que permite capturar la radiación luminosa de un plasma (gas altamente ionizado) para determinar las especies o líneas espectrales que componen un elemento químico; a este procedimiento se le conoce como caracterización [1]. En este proceso los parámetros más importantes son: longitud de onda e intensidad. Estos son adquiridos de una corrida o ejecución experimental utilizando una lámpara de calibración de mercurio y argón modelo *HG-1* de *Ocean Optics*, que tiene un rango de longitud de onda de 177 nm a 891 nm. Además, otros datos considerados son las especies de elementos reportados por el Instituto Nacional de Estándares y Tecnología (NIST por sus siglas en inglés) [2].

Trabajos en la literatura presentan la caracterización de especies con técnicas de *Machine Learning*. Los estudios realizados muestran el análisis de espectros de rayos gamma mediante el uso de **Redes Neuronales Artificiales (RNA)** para el análisis de 28 radioisótopos. La red neuronal propuesta 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 [3].

Las técnicas de *Machine Learning* en este trabajo se enfocan en el aprendizaje supervisado con algoritmos de clasificación, alcanzando predicciones altas, en datos con características similares, como se detalle en la sección 3.1. Las especies estudiadas son: ^{7}N (Nitrógeno), ^{8}O (Oxígeno), así como ^{18}Ar (Argón) y ^{80}Hg (Mercurio). Estas últimas, ^{18}Ar y ^{80}Hg componen una lámpara de calibración *HG-1*, donde las longitudes de ondas de las especies son conocidas, por lo que son utilizadas 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. La desventaja es que los datos no están balanceados, como se trata en la sección 2.5.

La caracterización de forma manual de especies es un procedimiento que puede llevar desde horas hasta días, por esta razón se propone el uso de técnicas de *Machine Learning* para

agilizar este proceso. 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 con datos no balanceados. El lenguaje de programación utilizado para la implementación es *Python*, debido a sus características de: rapidez de ejecución, portabilidad, y es potencialmente un lenguaje empleado por empresas que se dedican a la investigación en el área de la Inteligencia Artificial (IA) como lo es Intel y Google. De esta manera, se utilizaron bibliotecas con funciones específicas en el desarrollo de software científico tales como: *pandas*, *numpy*, *sklearn*, *scipy*, *matplotlib* y *seaborn* [4, 5].

2. METODOLOGÍA

El procedimiento desarrollado inicia con la adquisición de datos de un espectrómetro para crear un repositorio local. En segundo lugar, se realiza la corrección de datos, que consiste en un filtro, y en la corrección del fondo continuo. Posteriormente se realiza el entrenamiento de los algoritmos de clasificación con el conjunto de datos. Finalmente, la caracterización automática de especies, tiene como fases la comprobación mediante la métrica F1, optimización y regresión de parámetros, hasta obtener los resultados esperados. La Figura 1, muestra la metodología empleada mediante un diagrama de flujo. A continuación, se detallan cada uno de los procesos.

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; 99 archivos que corresponden a los primeros 99 elementos. Cada archivo contiene de 12 a 20 columnas de tipo numérico y alfanumérico, de las cuales sólo 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 por 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 0.039639 nm a 60,640,000,000 nm, con la característica de que el encabezado *obs_wl_vac(nm)* se intercala a lo largo de los archivos con *obs_wl_air(nm)*. Esta 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 esta razón 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 alfanuméricos que acompañaban a datos numéricos, estos archivos fueron utilizados como datos de entrenamiento.

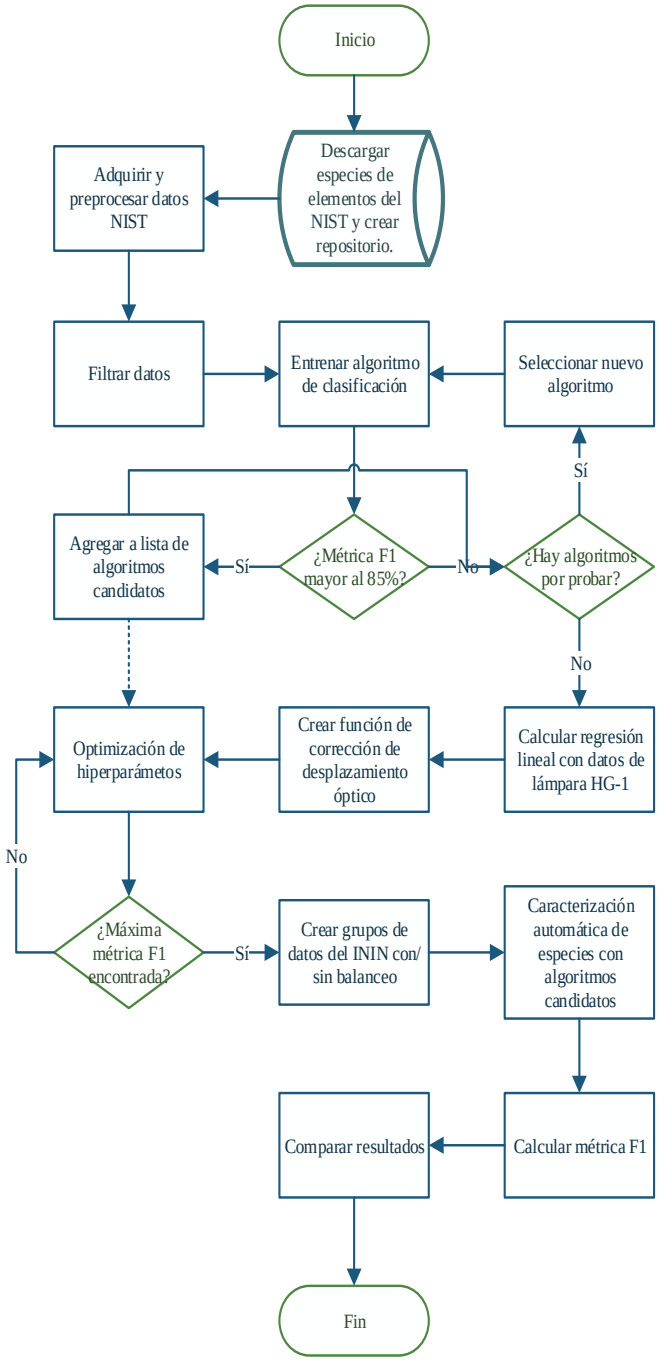

Figura 1. Diagrama de flujo para le metodología empleada en este trabajo.

Los datos experimentales de la espectroscopia de emisión óptica son adquiridos mediante tres corridas experimentales de una lámpara de calibración HG-1 de ^{18}Ar y ^{80}HG, en el Instituto Nacional de Investigaciones Nucleares (ININ). Cada una con 33 archivos con extensión .dat; estos archivos contienen 1,024 filas de datos sin etiquetar. En primer lugar, la columna es la longitud de onda que tiene un rango de 177 nm a 891 nm, y en segundo

lugar la otra columna corresponde a la intensidad, 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*; los datos de entrenamiento provienen del 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% [6]. Cuatro algoritmos fueron estudiados, el primero como objeto de estudio es **k-Nearest Neightbors** [7, 8], este 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 segundo algoritmo es el denominado **Decision Tree**. 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 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 [9-11].

El tercer algoritmo **Random Forest**, 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 de prueba se pasa como entrada a cada árbol generado, y se cuentan las predicciones, mostrando aquella que apareció mayoritariamente [12].

Finalmente, el último algoritmo es **AdaBoost**, en donde cada árbol consta únicamente de un nodo y dos 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 [13, 14].

2.3. Corrección de fondo continuo y desplazamiento óptico.

El fondo continuo es una señal de ruido sobre la que se montan los espectros adquiridos con el monocromador, provocado deformaciones. La corrección del fondo continuo se realizó con una implementación en Python de mínimos cuadrados asimétricos [15].

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, produciendo predicciones incorrectas. En la Figura 2 se observan dos especies del elemento ^{18}Ar, ubicadas en las longitudes de onda 750.3869 nm y 763.5106 nm que corresponde al espectro del archivo 25 de la lámpara de calibración, y se observa que los picos del espectro ^{18}Ar corregido, coinciden con las líneas de la especie ^{18}Ar, al aplicar la corrección del desplazamiento óptico.

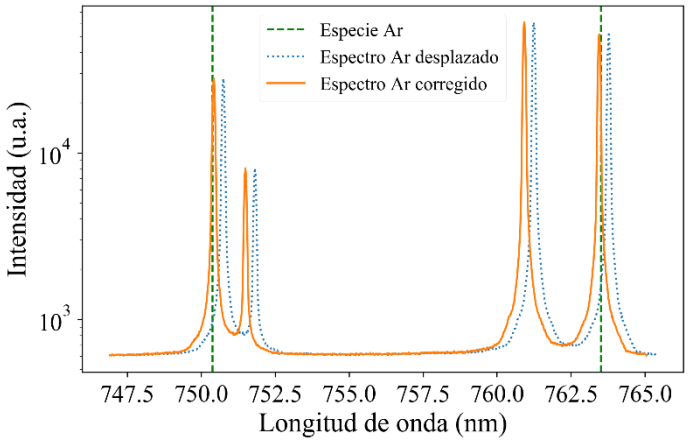

Figura 2. Corrección de desplazamiento óptico en espectro de lámpara HG-1

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

a) Detectar los picos de mercurio ^{80}Hg y argón ^{18}Ar en los archivos provenientes de la lámpara de calibración.

b) Calcular la diferencia (yD) entre las especies ^{18}Ar y ^{80}Hg 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 $x_i = xNIST_i$, como:

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

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

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

Dónde, b es la intersección y m es la pendiente.

d) La grafica resultante se muestra en la Figura 3 dónde se aprecia que la corrida 2 y 3 son similares, y la corrida 3 está ligeramente desplazada a causa de valores atípicos.

e) Las intersecciones b y las pendientes m de las corridas experimentales 2 y 3 son similares (ver Figura 3), por lo tanto, se calcula el promedio para obtener la siguiente ecuación:

$$y_i = -0.536024 + 2.815753^{-4} x_i \quad (4)$$

f) Reemplazar cada pico detectado origen (ININ) x_i en la ecuación 4, para corregir el desplazamiento óptico y

posteriormente predecir con los algoritmos de *Machine Learning* propuestos en la sección 2.2.

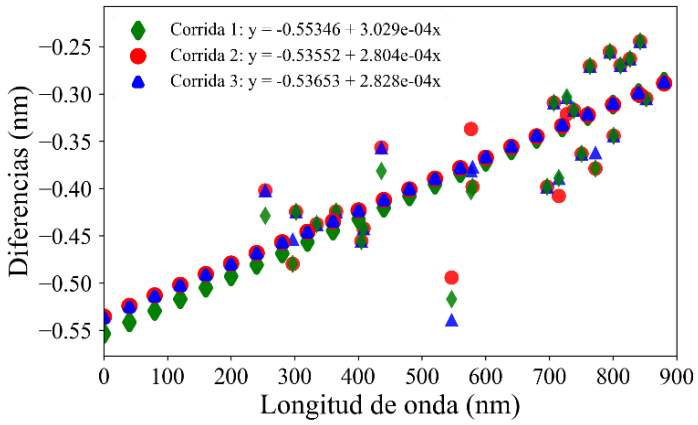

Figura 3. Regresión lineal para cada corrida experimental de la lámpara de calibración HG-1.

2.4. Optimización de hiperparámetros.

Un problema recurrente en los árboles de decisión es el sobreajuste, debido a que el algoritmo busca en cada nodo disminuir la entropía (grado de dispersión de datos) hasta 0. Por esta razón se grafica profundidad contra precisión, y efectividad, con el objetivo de encontrar la profundidad en donde la predicción en los datos de prueba y entrenamiento convergen [16]. En la Figura 4 se observa que a partir de la profundidad ubicada en la magnitud 22, para **Random Forest** y en 24 para **Decision Tree**, la métrica F1 converge, por lo tanto, es la profundidad indicada [17].

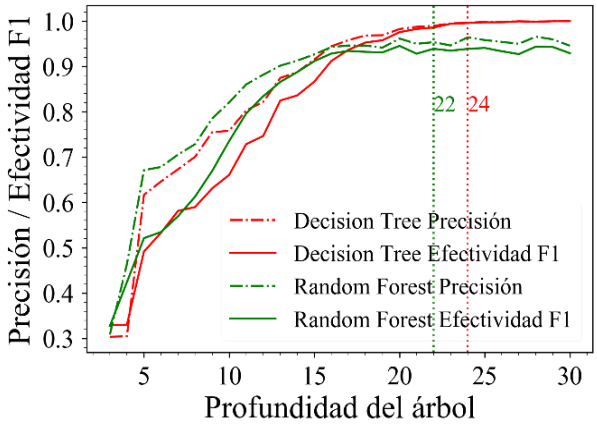

Figura 4. Comparativa de Precisión vs Profundidad para los algoritmos Decision Tree y Random Forest.

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 resultar en un rendimiento pobre [11]. En la Figura 5 se muestran las cuatro especies por elemento correspondientes a la lámpara de calibración y se observa que ^{18}Ar es la clase mayoritaria con 1,859 especies, mientras que ^{80}Hg, ^{8}O y ^{7}N son las clases minoritarias. Especies reportadas en el NIST, 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.

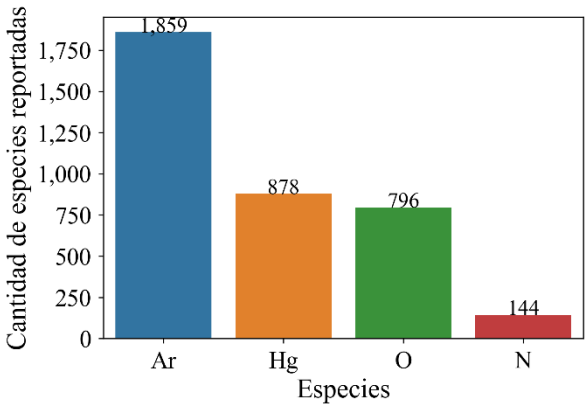

Figura 5. Resultados de métrica F1, para para datos no balanceados.

El desbalanceo de datos se trató así: A) cálculo de la métrica F1 de los datos si balancear. B) obtención de un conjunto de datos mediante sobremuestreo para las especies de los elementos ^{80}Hg, ^{8}O, ^{7}N, duplicando líneas espectrales al azar hasta alcanzar las 1,859 líneas e igualar así las especies de ^{18}Ar. Posteriormente se determinó la métrica F1 con los algoritmos de clasificación. C) obtención de otro conjunto de datos mediante submuestreo de los elementos ^{18}Ar, ^{80}Hg, ^{8}O, hasta llegar a las 144 especies e igualar así las del elemento ^{7}N, eliminando al azar especies de estos elementos, y se obtuvo la métrica F1. D) comparación de las métricas F1 obtenidas de todos los conjuntos. En la Figura 6 se observa que el sobremuestreo tiende a mejorar la métrica F1 en todos los algoritmos estudiados, por lo tanto, este ajuste es parte de la solución.

3. RESULTADOS.

3.1. 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 los algoritmos 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, y G) calcular métrica F1 sí las etiquetas son conocidas.

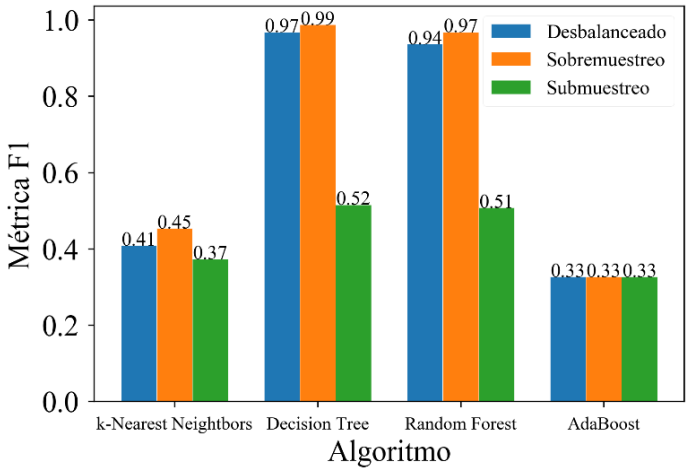

Figura 6. Métrica F1 para cada algoritmo estudiado en condiciones de con datos desbalanceados, que se les aplicó sobremuestreo y submuestreo

En la Figura 7, se observa la corrida experimental número 2 de la lámpara de calibración HG-1 en el rango de longitud de onda de los 200 nm hasta los 900 nm..

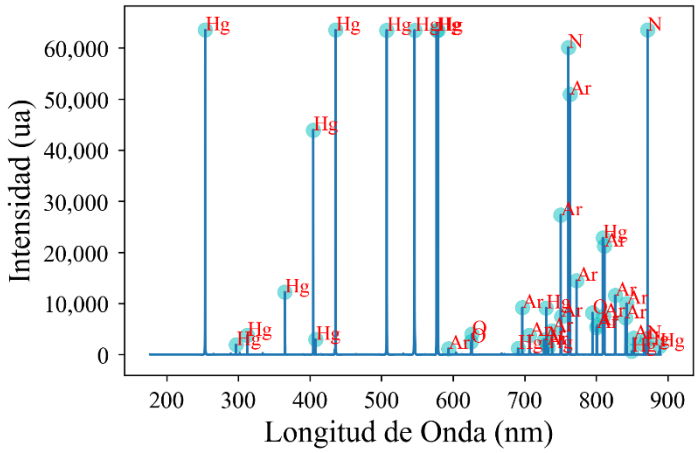

Figura 7. Caracterización automática de especies. Los círculos representan los picos detectados, donde la etiqueta indica el elemento predicho.

En la Figura 8 se muestra un acercamiento de la Figura 7 en el rango de 760 nm hasta 810 nm, donde se aprecia con mayor claridad las especies elementos ^{7}N, ^{8}O, ^{18}Ar y ^{80}Hg.

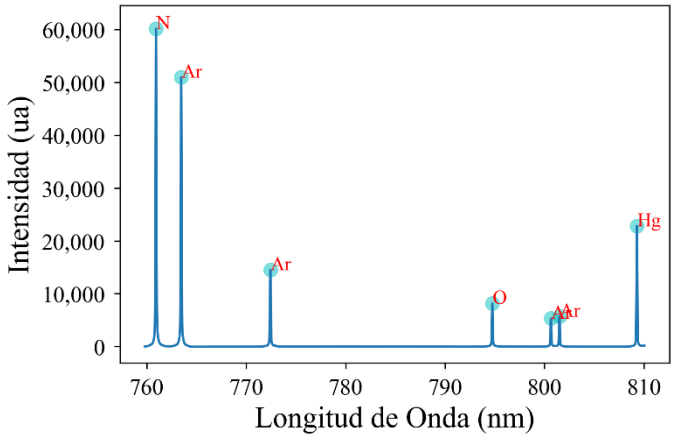

Figura 8. Acercamiento en la caracterización automática de especies para la lámpara de calibración HG-1 en el rango de longitud de onda de los 760 nm hasta los 810 nm, donde se aprecian las especies de los elementos ^{7}N, ^{8}O, ^{18}Ar y ^{80}Hg.

La corrección del desplazamiento óptico aproxima cada pico detectado con la especie de elemento que le corresponde. La optimización de hiperparámetros junto con el balanceo de datos contribuye en predicciones más precisas y efectivas. En este caso, el sobremuestreo influye en los algoritmos estudiados para generar mejores predicciones.

Los puntajes más altos se alcanzaron con los algoritmos **Decisión Tree** y **Random Forest**, utilizando las correcciones y optimizaciones presentadas en las secciones 2.3 y 2.4. El conjunto de datos de prueba corresponde a una lámpara de calibración modelo HG-1, y se sabe a priori que las especies de ^{80}Hg y ^{18}Ar son mayoría, el resto son especies que no vienen descritas en el manual y que el clasificador identificó como ^{7}N y ^{8}O.

Finalmente, la caracterización automática de especies se llevó a cabo para las especies elementos ^{7}N, ^{8}O, ^{18}Ar y ^{80}Hg, en un rango de 200 nm hasta los 900 nm, donde cada elemento es etiquetado en su respectivo valor de longitud de onda para identificarlo.

CONCLUSION Y TRABAJO A FUTURO.

El problema estudiado consiste en la caracterización automática de líneas espectrales resultantes de una espectroscopía de emisión óptica con técnicas de *Machine Learning*; se propone como mejor solución el uso de árboles de decisión como estimador del *random forest*, con datos previamente tratados mediante sobremuestreo y corrección de desplazamiento óptico con regresión lineal.

Se propuso el uso de algoritmos de regresión que permitan corregir el desplazamiento óptico, y como se muestra en la Figura 2, los picos detectados se aproximan a la especie que le corresponde.

La optimización de hiperparámetros se plantea en una segunda etapa con la implementación de un algoritmo genético y

validación cruzada, con la finalidad de aumentar la precisión en las predicciones.

Finalmente se propone ensamblar los algoritmos con las puntuaciones más altas, es decir, obtener la predicción de cada algoritmo y por votación mostrar la etiqueta predictora con la puntuación más alta.

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] National Institute of Standars and Technology, «NIST Atomic Spectra Database Lines Form,» 09 04 2019. [En línea]. Available: 19. [Último acceso: 11 10 2018].

[3] 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.

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

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

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

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

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

[9] 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.

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

[11] 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.

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

[13] 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.

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

[15] P. H. Eilers y H. F. Boelens, «Baseline correction with asymmetric least squares smoothing,» *Leiden University Medical Centre Report,* vol. 1, nº 1, pp. 1-24, 2005.

[16] 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.

[17] 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.