Marvik - Trabajo final DL: Cíclope

Tabaré Pérez

8 de agosto de 2022

El mito del Cíclope

Figura 1: Cíclope

En la mitología griega, los Cíclopes (en griego Kýklopes, que viene de "kyklos", rueda, círculo y "ops", 'ojo') eran los miembros de una raza de gigantes con un solo ojo en mitad de la frente.

El mito del Cíclope podría haber surgido a partir de un gremio de forjadores de metal de la Edad del Bronce que probablemente tenían tatuados en la frente anillos concéntricos como muestra de homenaje al sol por ser su fuente de energía.

El instrumento usado para relevar los espectros con los cuales desarrollaremos este trabajo, tiene "un sólo ojo" con el cual "observa" a las muestras en longitudes de onda del NIR (Near Infrared) entre 950nm y 1650nm. Por esta característica es que uso el nombre "Ciclope" en este trabajo.

Introducción

La espectroscopía agrupa a un conjunto de diferentes técnicas que usan radiación electromagnética para obtener datos de la estructura y propiedades de la materia que son usadas para la resolución de una amplia variedad de problemas analíticos.

El término deriva de la palabra "spectron" en Latín, que significa espíritu o fantasma, y de la palabra griega "skopein" que significa "mirar dentro del mundo".

En resumen, la espectroscopía se ocupa de medir e interpretar los espectros que surgen de la interacción de la radiación electromagnética con la materia. Se refiere a la absorción, emisión o dispersión de la radiación electromagnética por átomos o moléculas.

Desde su inserción como técnica de relevamiento de datos en la segunda mitad del siglo XIX, se ha desarrollado para incluir todas las regiones del espectro electromagnético y todos los procesos atómicos o moleculares alcanzables. En consecuencia, la mayoría de los ingenieros y científicos trabajan directa o indirectamente con espectroscopía en algún momento de su carrera.
.

La espectroscopía representa un enfoque metodológico general, mientras que los métodos pueden variar con respecto a:

- Las especies analizadas (átomos, moléculas).
- La región del espectro electromagnético (UV, visible, infrarrojo, etc)
- El tipo de interacción radiación-materia (como emisión, absorción, reflexión, transmisión o difracción).

El principio fundamental compartido por todas las diferentes técnicas es iluminar a la muestra deseada con un haz de radiación electromagnética para observar cómo responde a tal estímulo. La respuesta se registra típicamente

como una función de la longitud de onda de radiación, y una gráfica de tales respuestas representa un espectro. Cualquier energía de luz (desde ondas de radio de baja energía hasta rayos gamma de alta energía) puede producir un espectro.

Los objetivos generales de la espectroscopía son comprender cómo interactúa exactamente la luz con la materia y cómo se puede usar esa información para comprender cuantitativamente y/o cualitativamente las propiedades químicas de la muestra. Además, la espectroscopía también debe ser apreciada como un conjunto de herramientas que pueden emplearse juntas para comprender diferentes sistemas y resolver problemas químicos complejos.

En este trabajo vamos a utilizar técnicas especroscópicas en las regiones del NIR (Near Infrared - infrarrojo cercano) para resolver el tipo de problemas enunciados anteriormente.

Veamos a continuación dónde se ubica el NIR en el espectro electromagnético:

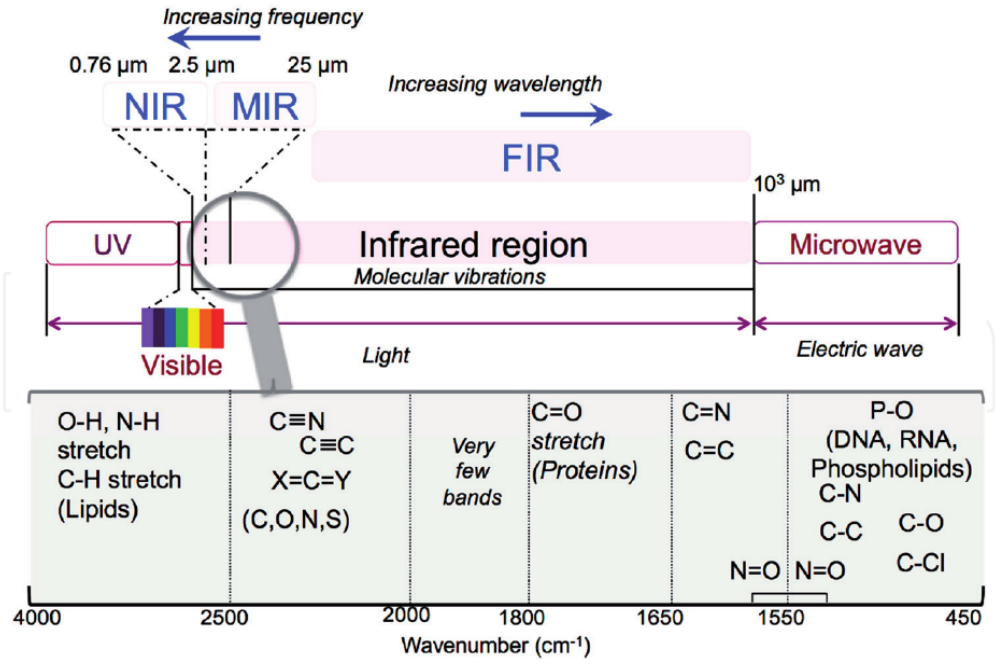

Figura 2: NIR en el espectro electromagnético

Antecedentes: Quimiometría

La quimiometría se desarrolló en la década de 1960. Extrae información de los sistemas químicos mediante el uso de métodos como la estadística multivariada, las matemáticas aplicadas y la informática, para abordar problemas de química, bioquímica, medicina, biología e ingeniería química.

Svante Wold (Umeå Universitet, Suecia) inventó la palabra **"quimiometría"** para una solicitud de subvención a fines de 1971. En 1974, junto con Bruce Kowalski, Universidad de Washington, Seattle, EE. UU., crean la "Sociedad Internacional de Quimiometría (ICS)". El primer artículo con la palabra quimiometría fue publicado por Wold en 1972. Sorprendentemente, solo se cita siete veces según "Web of Science".

En la década de 1980, aparecen las primeras revistas dedicadas al tema como por ejemplo "Chemometrics and Intelligent Laboratory Systems" y "Journal of Chemometrics", el primer libro con la palabra quimiometría en el título, varios simposios de ACS, la primera serie de libros sobre el tema (Research Studies Press), el primer software dedicado (ARTHUR, SIMCA, y UNSCRAMBLER), y los primeros talleres. Una reunión celebrada en Cosenza, Italia, en 1983 fue probablemente el primer gran intento de reunir a una amplia gama internacional de científicos que trabajaban en quimiometría. El desarrollo de la quimiometría está fuertemente relacionado con la incorporación de instrumental con capacidad de digitalización de los resultados analíticos y de conectividad con computadoras.

Personalmente fui muy afortunado de vivir todo el proceso de evolución de la quimiometría nombrado anteriormente. Fui protagonista de la transformación de nuestro laboratorio de un laboratorio de "química húmeda" clásico pasando por un "laboratorio de instrumentación electrónica" y en esta etapa final a un "laboratorio de tratamiento y generación de información". Nos queda mucho camino por recorrer y este curso es parte del inicio del viaje.

Actualmente, el uso de técnicas de "Machine Learning", "Deep Learning" y "Artificial Inteligence" son cada vez más usadas en este campo que acompaña tanto el desarrollo de los nuevos métodos de cálculo como el aprovechamiento de hardware cada vez más poderoso para el tratamiento de grandes volúmenes de datos.

Este trabajo es un modesto ejemplo en pequeña escala de este tipo de técnicas para resolver, en este caso, problemas de clasificación y de estimación de la concentración de analitos de interés a partir de espectros NIR.

Objetivo del trabajo

Usar los espectros NIR de muestras de diferentes tipos de tabaco obtenidos por un espectrofotómetro Perten 7250 para implementar técnicas de DL para la modelización, predicción y clasificación de dichas muestras. Se dispondrá de una base de datos de espectros con sus respectivas etiquetas correspondientes a los tipos de tabacos presentes en el set y concentraciones de nicotina y azúcar determinadas usando métodos analíticos de referencia y acreditados ISO17025.

Figura 3: Perten 7250 y cápsula de medición

La estimación de la concentración de un analito de interés en una muestra usando métodos espectrales tiene las siguientes ventajas:

- Poca manipulación para la preparación de la muestra. En el caso de los tabacos lo único que se hace es moler la muestra.
- Sustituye, para determinaciones de rutina, a los métodos tradicionales húmedos con lo cual se baja el costo por análisis (no se usan reactivos) y el tiempo de respuesta (de horas a pocos segundos).

Para el caso de la clasificación de las muestras por métodos espectrales, en el caso del tabaco, complementa el trabajo del catador experto que, por sus años de experiencia, es capaz de clasificar las muestras según sus percepciones organolépticas (color, aroma, textura). Este tipo de conocimiento, muy valioso, es difícil de transmitir y de aprender y depende además de las capacidades innatas del catador. Los métodos espectrales son objetivos y basados en medidas y métodos matemáticos.

Para la estimacion de las concentraciones de los analitos de interés, usaremos la "Ley de Lambert y Beer" que dice que:

La intensidad de un haz de luz monocromática, que incide perpendicular sobre una muestra, decrece exponencialmente con la concentración de la muestra según esta ley:

$$A = k * C$$

En donde:

- A = Absorbancia
- k = Constante
- C = Concentración del analito

Esto permite deducir que:

$$C = \sum_1^n k_i * A_i$$

- n = Número de longitudes de onda del espectro NIR
- A_i = Absorbancia en la longitud de onda i de espectro
- k = Coeficiente i del modelo correspondiente a la longitud de onda i del espectro.
- C = Estimacion de la concentración del analito

En los dos problemas de estimación (nicotina y azúcar), lo que hay que resolver es calcular los k_i de tal forma de obtener las concetraciones de los analitos de interés con un error apropiado para la tarea asignada.

Se plantea como objetivo secundario (se implementará si es posible hacerlo en el tiempo disponible para este trabajo final), el uso de transformaciones de los datos originales con el objetivo de reducir las dimensiones del problema y evaluar si las predicciones y las clasificaciones implementadas mejoran (corrección línea de base, FFT, PCA, algoritmos genéticos para la selección de las mejores longitudes de onda, etc)

Stack de herramientas de trabajo

Configuración del laptop personal:

- Lenovo T490
- Distribución Linux Mint

Composición del texto del informe y la presentación asociada:

- Emacs: Herramienta de composición para la generación del informe final:
  - En modo LaTeX(ver TUG) para los ajustes finos del texto y la generación del entregable en PDF.
  - En modo orgmode (ver orgmode.org) para la composición del texto y exportación a LaTeX.
  - Uso del Beamer LaTeX package, de la distribución TeXLive, para la generación de las filminas de las presentación.

Limpieza de datos:

- Emacs: Limpieza de los archivos .csv.
- LibreOffice Calc: formateo de números y textos para la exportación del set limpio a .csv

Entorno de procesamiento:

- Google Colab

Librerías Python

- Keras
- Tensorflow
- Skitlearn
- Numpy
- Pandas
- Matplotlib
- StatsModels
- BaselineRemoval
- PyCompare
- Pingouin

Repositorio de datos y programas:

- Github
- Cliente git de Linux para interactuar con el respositorio.

Set de datos

Los datos se obtuvieron a partir de muestras de tabacos pasadas por un espectrofotómetro Perten 7250.

Son archivos exportados desde el soft del espectrofotómetro en formato .csv:

- tabaco-all.csv
- burley-tostado-gral.csv
- burley-verde-01.csv
- burley-verde-02.csv
- oriental.csv
- venas.csv
- virginia-blend-01.csv
- virginia-blend-02.csv
- virginia-verde-01.csv
- virginia-verde-02.csv

Descripción de las variables:

El archivo exportado desde el instrumento genera un conjunto de variables de las cuales nos vamos a quedar con:

- Sample ID: Es un número que identifica a la muestra.
- Espectro NIR (son 141 longitudes de onda desde 950nm a 1650nm con paso de avance de 5nm). Esto corresponde a las 141 variables que vamos a usar como las X de entrada para todos los procesos.
- Product Name: Se identifica el tipo de tabaco de la muestra.

- Las Y de la matriz de datos son los valores de nicotina y azúcar determinados por métodos de referencia certificados por norma ISO17025 (acreditacipón de laboratorios de análisis) bajo la cual nuestro laboratorio está certificado:
  - NIC: Porcentaje de nicotina en la muestra.
  - AZ: Porcentaje de azúcar en la muestra.

A partir de esos archivos, se generaron dos sets para los dos problemas a resolver:

- Para la clasificación: clasificacion-02.csv
  - Es es un archivo en el cual vamos a agregar una variable binaria que identificará si una muestra es de tabaco virginia o no. Vamos a usar todos los datos de tabaco virginia, burley, oriental y venas que disponemos.

    Tenemos en total 1140 registros:
    - 637 tabacos tipo VIRGINIA.
    - 503 tabacos tipo BURLEY, ORIENTAL y VENAS.
- Para la estimación de nicotina y azúcar: nic-az.csv
  - Para este problema, vamos a usar, como origen de datos, el archivo tabacos.csv que contiene 3294 muestras de tabaco con sus respectivos análisis de laboratorio de nicotina y azúcar. Una vez revisado, se genera el archivo de trabajo nic-az.csv.

Problemas encontrados:

Se detecta que en la exportación desde el soft de gestión del espectrofotómetro, los nombres de algunas variables incluyen una coma lo cual provoca la aparición de variables extra que en realidad no existen. Detectado el problema, se usa el editor emacs para sustituir y corregir los nombre de variables problemáticas.

Luego de corregidos estos problemas, usamos la planilla electrónica Calc, del paquete ofimático LibreOffice, para proceder al borrado de las variables irrelevantes para el problema y obtener así un set de datos con estrictamente las variables a usar. Se trabaja sobre la cantidad de dígitos de los valores de absorción del espectro (950nm a 1650nm en paso de a 5nm) redondeando dichos valores a 8 dígitos después de la coma. Los valores de nicotina usan 2 decimales y para el azúcar 1 decimal. A partir de cada planilla se generan dos archivos:

- Uno en formato .ods (formato estandard para planillas) en el cual se tienen en cuenta algunos aspectos estéticos para facilitar la visualización de los datos y su inspección.
- Uno en formato .csv que es el que se usa para el porcesamiento.

Datos

El primer paso es mirar los datos que tenemos disponibles.

Veamos primero el aspecto que tienen los primeros 20 espectros NIR de las muestras que nos provee el instrumento:

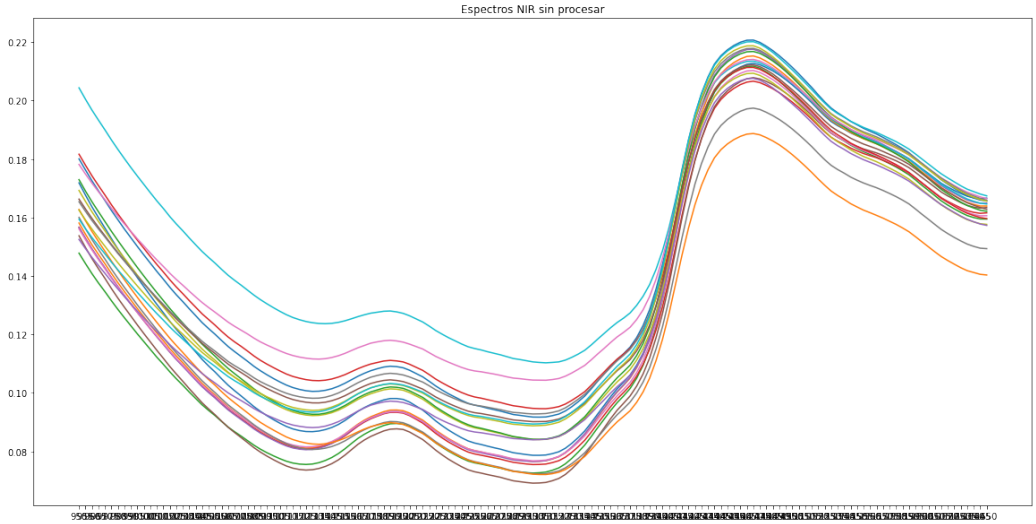

Figura 4: Espectros crudos

Esta es la forma típica de un espectro NIR del tabaco. Son muy similares y podemos ver diferencias en pendientes y alturas. Estos espectros van a ser procesados para evaluar si, tanto la clasificación como la estimación de las concentraciones de los analitos, mejora.

Veamos ahora cómo se distribuyen los valores de la nicotina y azúcar:

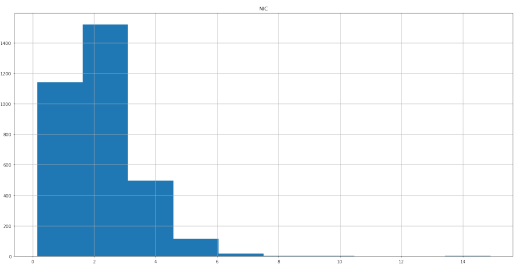

Figura 5: Distribución NICOTINA

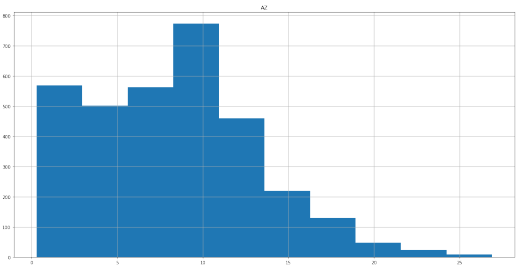

Figura 6: Distribución AZÚCAR

Lo ideal es poder disponer de distribuciones uniformes de las variables a ser estimadas. Para el caso de los tabacos, es una dificultad importante poder disponer de un set acondicionado de esa manera ya que son de productos naturales y muchas veces no se dispone de muestras que cubran algunos rangos de interés desde el punto de vista de las concentraciones de los analitos.

Estimación de Nicotina usando ANN densas

Para este primer problema de regresión, nos planteamos dos escenarios para explorar el comportamiento de los espectros crudos y de los espectros con corrección de línea de base.

Veamos la diferencia de estos dos juegos de datos:

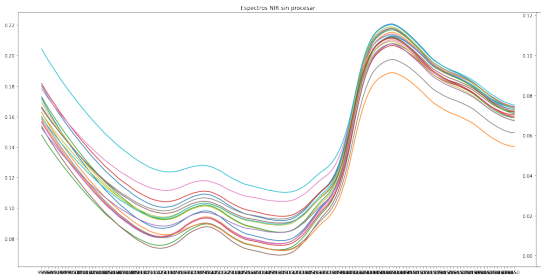

Figura 7: Espectros crudos

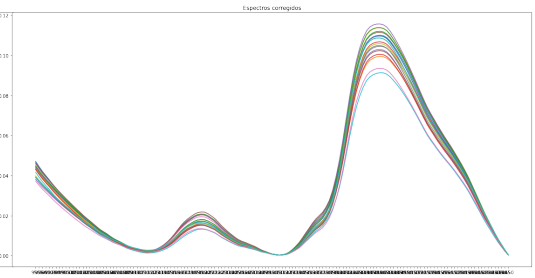

Figura 8: Espectros corregidos

Es de destacar que, dada la naturaleza de los espectros y de la información contenida en los mismos, es importante conservar la forma o sea la relación de alturas entre todas las longitudes de onda. Cualquier procesamiento que se haga debe de respetar esto ya que el espectro de una sustancia es su huella digital. El procedimiento de corrección de línea de base respeta esta condición.

Pero, ¿cuál el significado de la corrección de línea de base?:

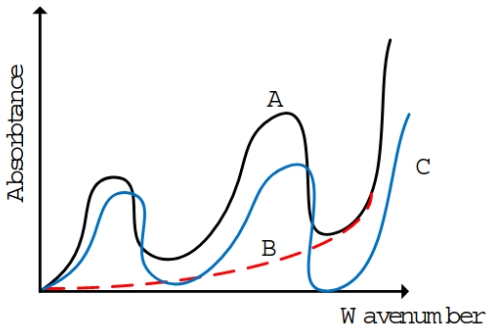

Figura 9: Corrección de la línea de base

A es el espectro con línea base desviada que se contradice con la condición de linealidad de la ley de Lambert - Beer.

B es la línea de base estimada que,

al restarla a A obtenemos **C** que es el espectro corregido que cumple con la condición de linealidad deseada.

Al corregir todos los espectros, estamos mejorando la relación señal ruido del set de datos y, de esta forma, aspiramos a obtener un mejor modelo de predicción. Esta es la hipótesis planteada.

En el caso de los espectros NIR, se usó un polinomio de segundo grado para estimar la línea de base a restar.

Se construyó una red densa con la siguiente arquitectura:

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_4 (Dense)              (None, 64)                9088

 dense_5 (Dense)              (None, 32)                2080

 dense_6 (Dense)              (None, 16)                528

 dense_7 (Dense)              (None, 1)                 17

=================================================================
Total params: 11,713
Trainable params: 11,713
Non-trainable params: 0
_________________________________________________________________
```

Figura 10: Arquitectura de la red usada para estimar nicotina

- Learning rate = 0.01
- epochs: 200

Se obtuvieron las siguientes métricas para cada unos de los juegos de espectros:

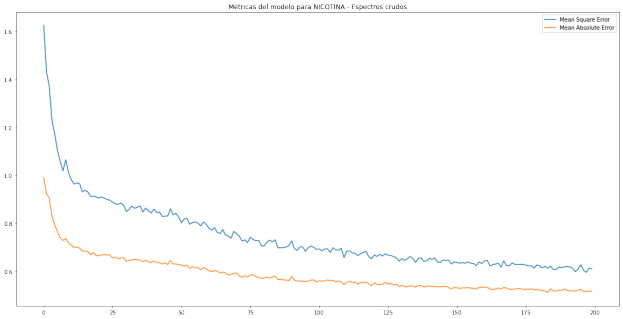

Figura 11: Métricas del modelo para los espectros crudos

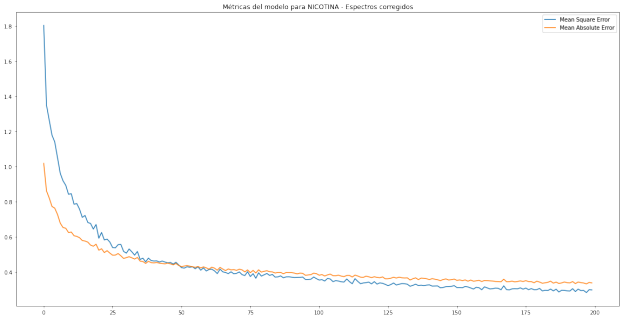

Figura 12: Métricas del modelo para los espectros corregidos

Vemos una mejora en las métricas para los espectros corregidos en su línea de base.

Para evaluar si podemos usar el método óptico en sustitución del método por química húmeda vamos a ver los gráficos denominados Bland - Altman usados muy a menudo en la clínica médica.

Los investigadores médicos a menudo necesitan comparar dos métodos de medición, o un nuevo método con uno establecido, para determinar si estos dos métodos se pueden usar indistintamente o si el nuevo método puede reemplazar al establecido. Para muchas aplicaciones no alcanza con evaluar la correlación. Es importante que el método analítico en estudio nos brinde valores lo más similares posibles al de referencia.

El método de Bland-Altman calcula la diferencia media entre dos métodos de medición (el "sesgo" o "bias") y los límites de concordancia del 95 % (1,96 desvíos estandard).

Se espera que los límites del 95 % incluyan el 95 % de las diferencias entre los dos métodos de medición. La gráfica se denomina comúnmente gráfica de Bland-Altman. El método de Bland-Altman puede incluso incluir la estimación de intervalos de confianza para el sesgo y los límites del corredor de confianza.

La presentación de los límites de concordancia del 95 % es para el juicio visual de qué tan bien concuerdan dos métodos de medición: cuanto menor sea el rango entre estos dos límites, mejor será el acuerdo.

La pregunta de qué tan pequeño es pequeño depende del contexto del método analítico que estemos estudiando: ¿una diferencia entre los métodos de medición tan extrema como la descrita por los límites de concordancia del 95 % afectaría significativamente la interpretación de los resultados?

Vemos a continuación los gráficos Bland - Altman para los espectros crudos y los corregidos por línea de base junto con los coeficientes de correlación obtenidos.

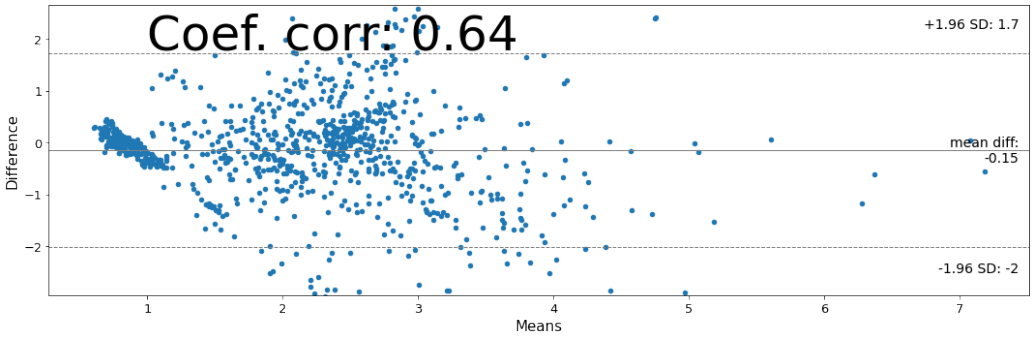

Figura 13: Evaluación del modelo: Bland-Altman - Espectros crudos

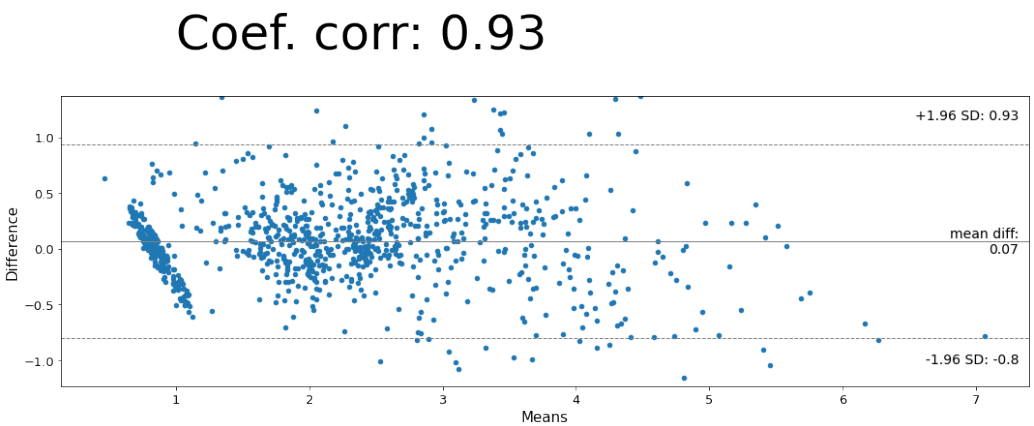

Figura 14: Evaluación del modelo: Bland-Altman - Espectros corregidos

Vemos una mejora importante en cuanto a la correlación obtenida para los espectros corregidos. Además mejoramos el "bias" y el ancho del corredor de confianza.

El segundo modelo construido en base a los espectros corregidos es mucho más usable que el primero desarrollado en base a los espectros crudos.

Estimación de Azúcar usando MLP Regressor

El punto de arranque es el mismo en cuanto al tratamiento de los espectros. Repasamos los gráficos de los espectros crudos y corregidos por línea de base:

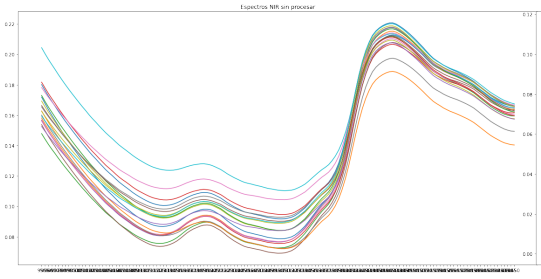

Figura 15: Espectros crudos

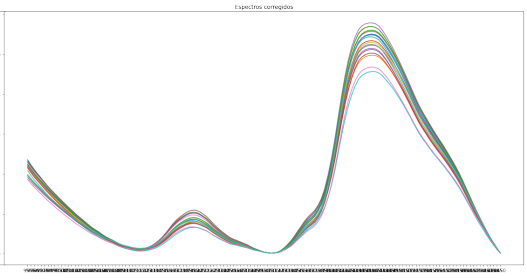

Figura 16: Espectros corregidos

Se optó por este tipo de redes para experimentar y ver su comportamiento y resultados en comparación con la red densa del problema anterior.

Recordemos que disponemos de un set de 3294 registros con sus respectivos resultados analíticos de referncia obtenidos en el laboratorio por métodos acreditados segúm ISO17025.

Se realizó el mismo tipo de tratamiento utilizando primero los espectros crudos y luego los espectros corregidos por línea de base.

La arquitectura del MPL Regressor es:

- MLPRegressor(hidden layer sizes=(141, 64, 32, 1)) con función de activación *relu* y *adam* para la optimización de los pesos.
- 200 iteraciones máximas.

Veamos el comportamientos de las métricas para cada uno de los sets de espectros:

Las respectivas funciones de pérdida:

ESPECTROS CRUDOS	ESPECTROS CORREGIDOS
MAE: 2.20	MAE: 1.40
MSE: 7.90	MSE: 3.51
RMSE: 2.81	RMSE: 1.90
R2: 0.68	R2: 0.86

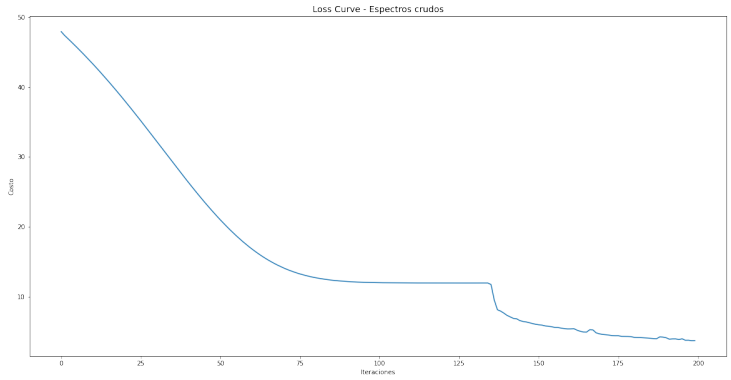

Figura 17: Función de pérdida para los espectros crudos

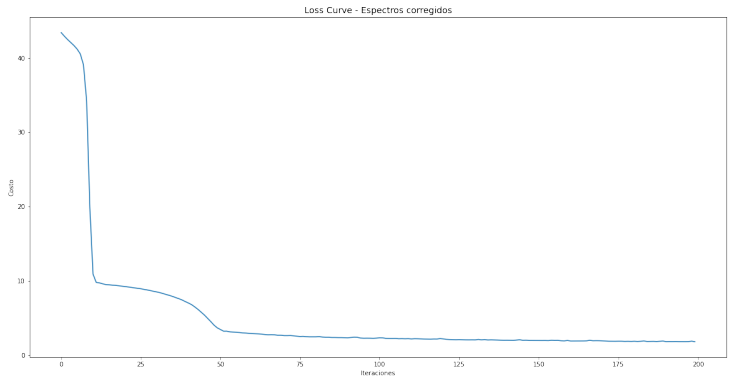

Figura 18: Función de pérdida para los espectros corregidos

Vemos también en este caso que la corrección de la línea de base mejora el comportamiento del modelo.

Clasificación de tabacos

El problema que se plantea es hacer una clasificación binaria que permita diferenciar aquellos tabacos que son del tipo Virginia de aquellos que no lo son. Recordemos que el set de datos a usar contiene 1140 registros de los

cuales 637 corresponden a tabacos tipo Virginia y 503 a tabacos tipo Burley, Oriental y venas (nervaduras de la hoja del tabaco).

El procedimiento es el mismo que para los modelos de estimación de y azúcar: generar, en este caso el modelo de clasificación, para el set de espectros crudos y para el set de espectros corregidos por línea de base y verificar si esto mejora la performance del modelo.

Repasamos los gráficos de los espectros crudos y corregidos por línea de base:

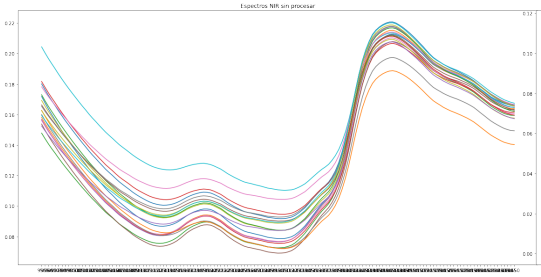

Figura 19: Espectros crudos

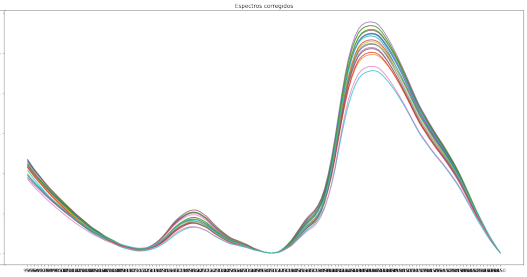

Figura 20: Espectros corregidos

La arquitectura del MPLRegressor para clasificación es la siguiente:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 16)                2272

 dense_1 (Dense)             (None, 18)                306

 dropout (Dropout)           (None, 18)                0

 dense_2 (Dense)             (None, 20)                380

 dense_3 (Dense)             (None, 24)                504

 dense_4 (Dense)             (None, 1)                 25

=================================================================
Total params: 3,487
Trainable params: 3,487
Non-trainable params: 0
```

Figura 21: Arquitectura de la red para clasificar tabacos Virginia

- Optimizador de los pesos: Adam
- Función de pérdida: Binary cross entropy (es una clasificación)
- Métrica: Accuracy
- epochs: 50

Veamos las matrices de confusión del test set y el full set con espectros crudos y corregidos por línea de base.

Para los espectros crudos:

	Test set		Full set	
	Recall: **0.83**		Recall: **0.84**	
0	162	17	568	68
1	27	136	80	423
	0	**1**	**0**	**1**

Para espectros corregidos:

	Test set		Full set	
	Recall: **0.95**		Recall: **0.94**	
0	165	14	603	33
1	8	155	28	475
	0	**1**	**0**	**1**

Vemos también mejoras para este caso en el cual desarrollamos un modelo de clasificación.

Conclusión

Es alentador ver que con solamente corregir la línea de base del set de espectros podemos mejorar la predicción de los analitos de interés y la clasificación de los tabacos.

Se necesita mucho más trabajo en acondicionar los sets de muestras (las distribuciones de las concentraciones de los analitos no son el todo adecuadas), el pre procesamiento de los espectros como se describió en la propuesta del trabajo (transformadas de Fourier, derivadas, PCA, algoritmos genéticos para la selección de las mejores longitudes de onda, etc) y en la arquitectura de los modelos.

Los pasos a seguir son los de profundizar en estos temas y ampliar el alcance para poder estudiar si es viable la construcción de modelos de predicción más complejos que incluyan, por ejemplo, las características organolépticas de los diferentes tipos de tabaco y así poder predecir el comportamiento en boca y nariz del fumador o la confección de mapas de similitud que apoyen a la compra de tabacos para la producción de nuesros productos.

Bibliografía

La bibliografía usada corresponde a la documentación de las diferentes librerías Python usadas:

- Keras
- Tensorflow
- Skitlearn
- Numpy
- Pandas
- Matplotlib
- StatsModels
- BaselineRemoval
- PyCompare
- Pingouin

Anexo

Vínculos a los notebooks en Colab de los tres problemas presentados:

- Estimación de Nicotina en tabaco
- Estimación de Azúcar en tabaco
- Clasificación de tabacos Virginia

Vínculo al repositorio GitHub con el material del trabajo final:

- Repositorio GitHub