BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

CARACTERIZACIÓN DE SENSORES BASADOS EN REJILLAS DE PERIODO LARGO PARA LA DETECCIÓN DE ACETONA COMO BIOMARCADOR DE LA DIABETES MELLITUS USANDO TÉCNICAS DE MACHINE LEARNING.

TESIS PRESENTADA PARA OBTENER EL GRADO DE

LICENCIATURA EN FÍSICA

POR

ADRIÁN ITURBE GARCÍA

DIRECTOR DE TESIS

DRA. GEORGINA BELTRÁN PÉREZ

Puebla, Pue. Junio 2025

Agradecimientos

Agradezco a mis padres Gloria Soledad García Quirino y Marcelo David Iturbe Castillo por el apoyo incondicional que me han brindado durante mi existencia. A mi hermana Nadia Itzel Iturbe García por acompañarme en el trayecto. A mi pareja Diana Karen Gonzales Rodríguez por estar a mi lado y darme ánimos a cada momento. Finalmente, al grupo de personas que me ayudó a mantener la cordura; Alan, Edgar, Ricardo, Benjamín y Alonso. Sin cada una de estas personas este trabajo no se hubiera completado.

Resumen

En este trabajo de tesis se presenta la aplicación de modelos de análisis supervisado para caracterizar el desempeño de sensores de fibra óptica basados en rejillas de periodo largo (LPFG) fabricados. Estos fueron implementados con el objetivo de detectar acetona, la cual es un biomarcador de la diabetes mellitus, una enfermedad que aqueja a millones de personas en el mundo y miles de las que lo padecen desconocen si la tienen. Los síntomas de la diabetes mellitus tardan en manifestarse dependiendo de la persona, al no detectarse a tiempo las consecuencias pueden ser costosas e incluso letales. De entre los métodos que se utilizan para la detección de esta enfermedad se encuentra la A1c, en la cual se extrae plasma de la persona para su análisis. Proceso que es costoso de realizar e invasivo. Este tipo de pruebas se realizan cuando síntomas severos de la enfermedad se presentan, siendo tarde para una posible prevención de esta. El método del análisis del aliento humano es una alternativa para la detección y monitorización de esta enfermedad, además de ser un método no invasivo. Las concentraciones de acetona presentes en el aliento son mayores a 1.8 ppm para las personas que tienen diabetes, mientras que las personas sanas exhalan concentraciones menores a 0.8 ppm. Actualmente para este método se utilizan equipos tales como: GC/MS (siglas de Gas Chromatography/Mass Spectroscopy) los cuales requieren de personal especializado y son costosos. Una alternativa accesible para el análisis del aliento son los sensores basados en fibra óptica debido a que son inmune a la interferencia electromagnética, portátiles, y de fácil acceso. En este trabajo se proponen sensores de rejilla de periodo largo (LPFG) usando diferentes periodos para la detección de acetona, utilizando polidimetilsiloxano (PDMS) como película sensible. Se utilizaron técnicas de machine learning para el procesamiento de las mediciones de los sensores, obteniendo el límite de detección (LOD) teórico a partir de regresiones tales como: de componentes principales, de estructuras latentes y de bosque aleatorio. El mejor sensor fue aquel con un periodo de 475 µm, obteniendo un límite de detección teórico de 4.65 ppm teórico regresión de bosque aleatorio.

Índice general

Lista de figuras.

Lista de tablas.

Capítulo 1: Introducción

1.1 Antecedentes

La diabetes mellitus es una enfermedad que aqueja a millones de personas alrededor del mundo. Esta enfermedad puede traer consecuencias fatales de no ser diagnosticada a tiempo. Los métodos usualmente utilizados para determinar si una persona tiene diabetes mellitus son costosos e invasivos, por lo que está la búsqueda de alternativas fiables de menor costo sin ser invasivos. Hay investigaciones que reportan que una persona enferma de diabetes mellitus exhala mayor concentración de acetona en el aliento que una persona sana [1], es decir, que la acetona es un biomarcador de la diabetes mellitus. Volviendo la detección de acetona un elemento importante en alternativas no invasivas para el diagnóstico de diabetes mellitus.

Existen varios instrumentos para la detección de acetona en el ambiente. Entre éstos se encuentran detectores de fotoionización [2], sensores de óxido metálico [3] y sensores basados en fibra óptica [4]. En el sector industrial se encuentran en mayor medida los detectores de fotoionización y los sensores de óxido metálico por su precisión y fiabilidad, pero tienen un alto costo para su implementación. Los sensores basados de fibra óptica son utilizados en distintos ámbitos por su versatilidad [5], al tener distintas opciones para su fabricación la elección de materiales y métodos repercute en su aplicación. Actualmente se han implementado opciones de bajo costo en fabricación de sensores de fibra óptica.

El grupo de investigación ha realizado distintas configuraciones de sensores de fibra óptica para la detección de acetona de bajo costo. Entre estos se desarrollaron sensores utilizando interferómetro de Mach-Zehnder por reflexión [6] y transmisión [7] utilizando distintos tipos de películas sensoras [8], además de rejillas de periodo largo por transmisión [9]. Pocos son los trabajos reportados fuera del grupo de investigación que utilizan sensores de rejillas de periodo largo para la detección de acetona [4]. En estos se obtienen límites de detección (LOD) de 2211 ppm utilizando rejillas con un periodo de 111 µm y una película sensora mesoporosa de nanopartículas de óxido de silicio [10], y 5.6 ppm utilizando rejillas con un periodo de 109 µm y una estructura zeolítica de imidazol como película sensora [11].

Estos resultados se obtienen a partir del método de demodulación, dónde se observa un punto del espectro en un pico de atenuación y se analiza su corrimiento espectral. Este método no considera información que se pueda ocultar en otros puntos del espectro, y machine learning proporciona herramientas que permiten aprovechar en su totalidad las mediciones realizadas.

Machine learning cuenta con herramientas que facilitan el procesamiento de datos complejos [12] Una de estas herramientas es el aprendizaje supervisado [13], que permite al investigador encontrar patrones o información oculta dentro de datos que cuentan con datos de salida. Este tipo de herramientas son populares para realizar análisis exploratorio y predictivo, de entre las técnicas para realizar predicciones se encuentran regresión de componentes principales (PCR) [14], regresión de estructuras latentes (PLSR) [15], y regresión de bosque aleatorio (RFR) [16]. En el grupo de investigación ya se ha utilizado PCR y PLSR para el análisis de las respuestas de sensores, revelando más información del comportamiento de los sensores, y obteniendo un límite de detección teórico (LODT) a partir de las mediciones realizadas con sensores de fibra óptica. De RFR se ha reportado su aplicación en sensores de fibra óptica para la medición de temperatura [16], por lo que esta técnica se propone para su aplicación en las mediciones realizadas con los sensores desarrollados en este trabajo. Realizando así una comparación entre estas técnicas de machine learning observando el rendimiento de los sensores y analizando cuál de estos aprovecha mejor las mediciones realizadas obteniendo más información de los datos.

1.2 Objetivos

1.2.1 Objetivo general

El objetivo del presente trabajo es mostrar que la aplicación de técnicas de machine learning permite obtener mejor información de las respuestas de los sensores basados en rejillas de periodo largo, en la detección de acetona como biomarcador de la diabetes mellitus.

1.2.2 Objetivos particulares

1. Estudio experimental de los sensores de rejillas de periodo largo con periodos de 475, 515 y 525 µm funcionalizados con polidimetilsiloxano para determinar su desempeño en la detección de acetona.

2. Utilizar técnicas de machine learning tales como: regresión de componentes principales (PCR), regresión de proyección de estructuras latentes (PLSR) y regresión de bosque aleatorio (RFR), para comparar la información obtenida de cada sensor con cada una de estas.

3. Obtener límites de detección teóricos que se encuentren dentro del rango exhalado por una persona enferma.

1.3 Justificación

La Diabetes Mellitus (DM) es una enfermedad crónica que se considera una pandemia a nivel global. De acuerdo con la Organización Mundial de la Salud (OMS) alrededor de 350 millones de personas en el mundo padecen de esta enfermedad [17]. Una persona puede desarrollar DM por distintos factores, algunos de los más comunes son: problemas genéticos, daño en el páncreas, hábitos alimenticios, y falta de actividad física. Esta enfermedad se divide en varios tipos; DM tipo 1 (DMT1), DM tipo 2 (DMT2) y DM gestacional (DMG) [18].

Una persona que padece de DMT1 tiene deficiencia de insulina [19]. Esto quiere decir que no procesan adecuadamente la glucosa que el cuerpo produce naturalmente, causando deshidratación, fatiga y, en casos extremos, cetoacidosis diabética. La cetoacidosis diabética es una condición grave que requiere tratamiento inmediato, de lo contrario, sus síntomas pueden ser letales. La DMT2 es el tipo de diabetes más común. Alrededor del 90% de la población total diagnosticada con DM [20] es la que padece esta variante de la enfermedad. DMT2 es causada por la condición se resistencia a la insulina, esta condición se da cuando partes del cuerpo no reaccionan adecuadamente a la insulina. Conforme avanza DMT2, al cuerpo se le va dificultando mantener la producción de insulina necesaria para estabilizar la glucosa en la sangre hasta que ya no le es posible, desarrollando hiperglucemia. Los síntomas de DMT2 pueden tardar años en manifestarse, por lo que hay miles de personas que desconocen que la padecen [20]. Un consistente estado de hiperglucemia descontrolada conduce a diversos escenarios; enfermedades cardiovasculares, ceguera, fallo del riñón, y coma diabético, por mencionar algunos de estos. Para evitar que una persona con DMT2 llegue a uno de los escenarios antes mencionados es importante detectar a tiempo la enfermedad.

Uno de los análisis más utilizados hoy en día para la detección de diabetes es la prueba A1c [21]. Esta prueba mide el promedio de azúcar en la sangre de los últimos 3 meses. Esta prueba requiere de extraer una prueba de sangre de la persona en un laboratorio o consultorio médico. Otro análisis utilizado es la prueba de tolerancia a la glucosa [21]. Dónde la persona tiene que estar en ayunas durante ocho horas antes de la prueba, y beber un líquido que contiene glucosa, dos horas después se extrae sangre de la persona para analizarla. Las desventajas más importantes de los análisis previamente mencionados es lo invasivos que son y el equipo especializado que se requiere para encontrar los biomarcadores en la sangre para diagnosticar fiablemente DM.

La alternativa para un diagnóstico de DM sin ser invasivo es a través del aliento. Esta alternativa es posible debido al proceso de cetosis, durante este proceso se produce cetona aumentando sus niveles en la sangre. El síntoma de interés de la cetosis es el aumento de concentración de acetona que una persona exhala en el aliento, este síntoma permite que la acetona sea un indicador de la DM. Por esto es importante el mejorar las capacidades de los

sensores ópticos para la detección de acetona, porque brindan una alternativa para que en un futuro de manera accesible y no invasiva diagnosticar DM [6-9].

Capítulo 2: Teoría de fibras ópticas

Introducción

La fibra óptica es una herramienta utilizada para guiar luz [22]. Su aplicación en los últimos años ha ido en aumento en sectores industriales e informáticos [23]. El uso principal de las fibras ópticas ha sido en el área de la comunicación debido a la eficiencia con la que estas envían y reciben información. Gracias a las propiedades físicas que posee, la fibra óptica tiene una amplia gama de uso para diversas áreas. Su composición también permite fabricar la fibra óptica acorde a las necesidades que se deben de cubrir, ya sea para transmitir imágenes o realizar mediciones de alta precisión. En este capítulo se introducirá al lector en la estructura de las fibras ópticas, junto con la teoría electromagnética que se maneja para el estudio de las fibras ópticas.

2.1 Fibra óptica

Las fibras ópticas tienen una geometría cilíndrica que se fabrica utilizando vidrio o plástico [23]. Estas son guías de onda que se aprovechan de un valor que poseen los materiales con los que son fabricadas denominado índice de refracción. El índice de refracción es el valor que expresa la razón entre la velocidad de la luz en el vacío, y la velocidad que tiene la luz al viajar en un medio, esto se expresa de la siguiente forma [24]

$$n_{med} = \frac{c_v}{c_{med}} \tag{2.1}$$

dónde c_v es la velocidad de la luz en el vacío, c_{med} es la velocidad de la luz en el medio j en el que esté viajando la luz, y n_{med} es el índice de refracción del medio. Las fibras ópticas están compuestas por un núcleo de índice de refracción n_1, una cubierta de índice de refracción n_2 que debe ser menor a n_1, y un revestimiento como se puede observar en Fig. 1. Desde el punto de vista de la óptica geométrica, la luz se representa como rayos que viajan por un medio, en este caso, el núcleo de la fibra [25]. Esta representación es utilizada para visualizar como viaja un haz de luz en los distintos tipos de fibra que existen y se pueden observar en la Fig. 2. Entre los tipos de fibra que se muestran está la fibra de salto de índice

monomodo, visualizada en Fig. 2a, que cuenta con un diámetro menor a los 10 µm en el núcleo, permitiendo un único modo óptico de propagación, siendo eficiente para el tránsito de información a grandes distancias. La fibra de salto de índice multimodo, representada en Fig. 2b, se caracteriza por tener un núcleo con índice de refracción uniforme y, a diferencia de la fibra monomodo, tienen un diámetro amplio de entre 50 µm y 200 µm, permitiendo el paso de múltiples modos ópticos. Este tipo de fibras ópticas se utiliza en distancias cortas para el envío de información. Finalmente, la fibra multimodo de índice gradiente Fig. 2c, cuenta con un núcleo no homogéneo, es decir, el índice de refracción de éste va cambiando gradualmente desde el centro hacia los extremos del núcleo. Este tipo de fibras ha tenido un amplio uso en infraestructura interurbana de alcance medio.

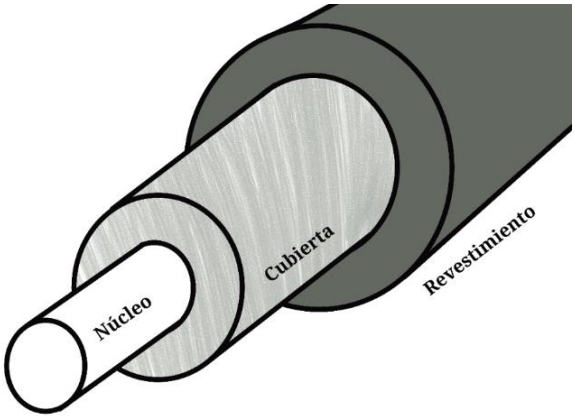

Figura 1. Representación de una fibra óptica.

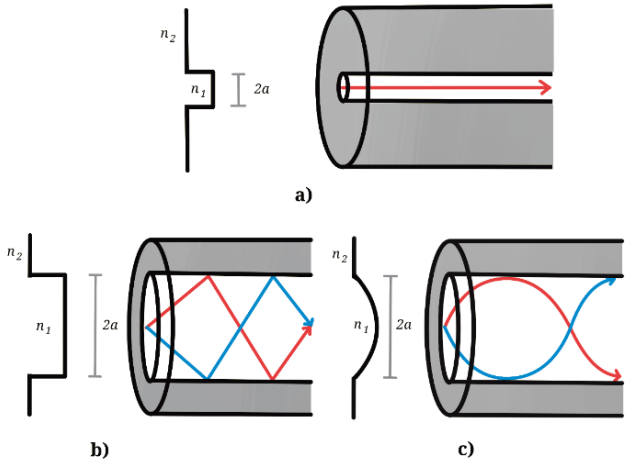

Figura 2. Perfil de una fibra a) monomodo de índice escalonado, b) multimodo de índice escalonado y c) multimodo de índice gradiente.

2.1.1 Reflexión total interna

El fenómeno físico que explica el funcionamiento de las fibras ópticas es la reflexión total interna [23-25]. Este fenómeno es explicado utilizando el acercamiento de la física geométrica que se vio con anterioridad junto con los fenómenos ópticos de reflexión y refracción. Cuando un haz de luz incide en una frontera dónde ocurre un cambio de medio, parte de éste se reflejará al medio dónde estaba viajando, mientras que el sobrante es refractado dentro del nuevo medio. Esto se expresa mediante la ley de Snell Ec. (2.2), dónde si se cumple que el índice de refracción del medio en el que viaja un haz de luz es mayor al del medio en el que incide ($n_1 > n_2$), ocurrirá la reflexión del haz incidente [24]

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{2.2}$$

Por esta condición las fibras ópticas poseen un núcleo con un índice de refracción mayor que el de la cubierta, para que la luz esté contenida dentro de este objeto al reflejarse cada vez que interactúe la luz con la cubierta. Otra consecuencia que conlleva la ley de Snell es que un haz de luz debe de incidir en el núcleo de la fibra sin sobrepasar un ángulo determinado, para que así, el haz de luz pueda quedarse confinado dentro de la fibra, en Fig. 3 se puede visualizar esto último. Si el ángulo de incidencia del haz está fuera del denominado ángulo de aceptación θ_{max}, éste no se propagará a lo largo de la fibra. El núcleo tiene un índice de refracción n_1 y el recubrimiento tiene un índice de refracción n_2, la diferencia entre ambos índices en la práctica es pequeño debido a los materiales utilizados para su fabricación. Arreglando los términos de la Ec (2.2) considerando el ángulo de aceptación se tiene la siguiente expresión [23]

$$n_i \sin\theta_{max} = (n_1^2 - n_2^2)^{\frac{1}{2}}, \tag{2.3}$$

dónde el término $n_i \sin\theta_{max}$ se define cómo la apertura numérica (NA) dónde n_i es el índice del medio que rodea a la fibra, en este caso aire, y se utiliza para determinar la recopilación de luz del sistema, además de que se obtiene el valor θ_{max}, también denominado el ángulo de aceptación del sistema.

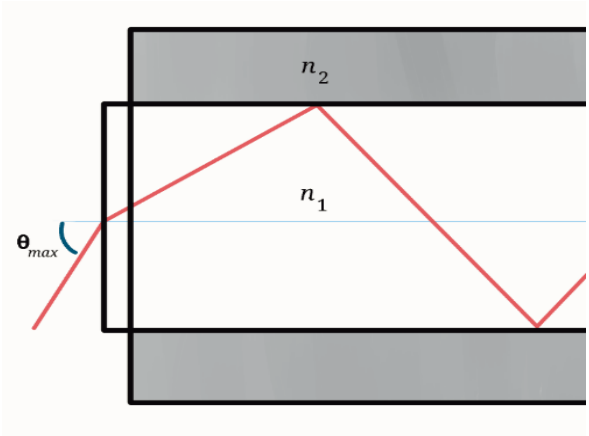

Figura 3. Representación de haz de luz que incide en una fibra

NA tiene relación con otro valor utilizado con normalidad en el área de fibras ópticas llamada diferencia normalizada del índice de refracción (Δ) [23]. Este valor expresa la diferencia entre los índices de refracción del núcleo y de la cubierta. Δ se expresa como [23]

$$\Delta = \frac{n_1^2 - n_2^2}{2n_1^2}, \tag{2.4}$$

este valor es utilizado para medir el tiempo que tarda una señal en recorrer una fibra óptica acorde a su longitud, y la relación entre NA y Δ es la siguiente [23]

$$\mathrm{NA} = n_1\sqrt{2\Delta}. \tag{2.5}$$

2.2 Ecuación de onda

El utilizar la óptica geométrica al analizar a las fibras ópticas ayuda a comprender su funcionamiento. Pero para explicar la propagación de la luz en estas es necesario recurrir a las características ondulatorias de la luz [22]. Para esto, se deben de obtener las ecuaciones de onda para las fibras de índice escalonado monomodo, que son de interés para este trabajo. Las ecuaciones de onda se derivan de las ecuaciones de Maxwell dadas las condiciones que se presentan en las fibras ópticas.

2.2.1 Ecuaciones de Maxwell

Las ecuaciones de Maxwell son las siguientes [23]

$$\nabla \cdot \boldsymbol{D} = \rho, \tag{2.6}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{2.7}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{2.8}$$

$$\nabla \times \boldsymbol{H} = J + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{2.9}$$

dónde $\boldsymbol{E}$ y $\boldsymbol{H}$ corresponden al campo eléctrico y magnético respectivamente, $\boldsymbol{D}$ es el desplazamiento dieléctrico tal que $\boldsymbol{D} = \varepsilon \boldsymbol{E}$ dónde ε la permitividad del medio, $\boldsymbol{B}$ es la inducción magnética tal que $\boldsymbol{B} = \mu \boldsymbol{H}$ dónde μ es la permeabilidad del medio, J es la densidad de corriente, y ρ es la densidad de carga. Keiser, G. [22] asume que las fibras ópticas son un medio lineal, dieléctrico e isotrópico. Por estas condiciones del medio se realizan las siguientes aproximaciones: $\rho = 0$ y $J = 0$. Causando que las ecuaciones (2.6) y (2.9) se simplifican de la siguiente manera [22]:

$$\nabla \cdot \boldsymbol{D} = 0, \tag{2.10}$$

$$\nabla \times \boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t}. \tag{2.11}$$

A partir de las ecuaciones (2.7) y (2.10) se tiene la siguiente relación entre $\vec{E}$ y $\vec{H}$ [22]

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\mu \frac{\partial}{\partial t} (\nabla \times \boldsymbol{H}) = -\varepsilon \mu \frac{\partial^2 E}{\partial t^2}, \tag{2.13}$$

de la cual se obtienen las ecuaciones de onda estándar aplicando el rotacional sobre la Ec (2.8), a lo cual tras aplicar la equivalencia de Ec. (2.3) se obtienen las siguientes equivalencias [22]:

$$\nabla^2 \boldsymbol{E} = \varepsilon \mu \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \tag{2.14}$$

$$\nabla^2 \boldsymbol{H} = \varepsilon \mu \frac{\partial^2 \boldsymbol{H}}{\partial t^2}. \tag{2.15}$$

2.2.2 Solución ecuación de Helmholtz en coordenadas cilíndricas

En la literatura es común observar que la solución general para la ecuación de onda en una fibra óptica es construida a partir de dos tipos de funciones Bessel [26]. Para obtenerlas primero se debe de resolver la ecuación de Helmholtz en coordenadas cilíndricas

$$\nabla^2 \boldsymbol{E}(\rho, \varphi, z) + k^2 \boldsymbol{E}(\rho, \varphi, z) = 0, \tag{2.16}$$

dónde k es el número de onda. Para resolver Ec. (2.16) se debe de realizar separación de variables tal que [26]

$$\boldsymbol{E}(\rho, \varphi, z) = R(\rho)\Phi(\phi)Z(z). \tag{2.17}$$

Las soluciones en forma compleja obtenidas de las ecuaciones diferenciales obtenidas para $Z(z)$ y $\Phi(\phi)$ son [26]

$$Z(z) = e^{\pm i\beta z}, \tag{2.18}$$

$$\Phi(\phi) = e^{\pm im\phi}, \tag{2.19}$$

dónde m es un valor que toma valores discretos positivos incluyendo el cero. Para $R(\rho)$ se obtiene la forma de una ecuación diferencial de Bessel [26]

$$\frac{d^2}{d\rho^2}R(\rho) + \frac{1}{\rho}\frac{d}{d\rho}R(\rho) + \left(q^2 - \frac{m^2}{\rho^2}\right)R(\rho) = 0, \tag{2.20}$$

dónde se define a $q^2 = k^2 - \beta^2$ tal que β es la constante de propagación que cumple con la condición $n_2 k < \beta < n_1 k$ y $k = \frac{2\pi}{\lambda}$. La Ec. (2.20) se resuelve considerando la dependencia temporal como una onda armónica $e^{-i\omega t}$ junto con el valor a, que es la distancia de la frontera entre núcleo y cubierta al eje de la fibra, dentro del núcleo, es decir cuando $\rho < a$, las soluciones toman forma de funciones de Bessel de primera clase [22]

$$E_z(\rho < a) = AJ_m(u\rho)e^{i(m\phi + \omega t - \beta z)}, \tag{2.21}$$

tal que A es una constante arbitraria y $u^2 = k_1^2 - \beta^2$ con $k_1 = \frac{2\pi n_1}{\lambda}$, para el campo magnético se encuentra una solución análoga. Fuera del núcleo, cuando $\rho > a$, se obtienen soluciones con la forma de función Bessel de segunda clase, denominadas K_m [22]

$$E_z(\rho > a) = CK_m(w\rho)e^{i(m\phi+\omega t-\beta z)}, \tag{2.22}$$

dónde C es otra constante arbitraria y $w^2 = \beta^2 - k_2^2$ con $k_2 = \frac{2\pi n_2}{\lambda}$, al igual que con Ec. (2.21) se encuentra una solución análoga para el campo magnético. Ante las Ecs. (2.21), (2.22) y sus análogas para el campo magnético se puede deducir que la constante de propagación β está condicionada en el rango de soluciones $k_2 < \beta < k_1$. Encontrando las soluciones de β ayuda a describir la propagación de la luz dentro de la guía de onda.

2.2.3 Propagación en una guía de onda cilíndrica

Las funciones Bessel requieren que el valor m tenga valores enteros incluyendo el cero [22]. De esta manera las funciones Bessel pueden expresar la propagación del haz de luz dentro de la fibra óptica. En este contexto, a m se le conoce como el orden del modo de propagación, el cual repercute en la distribución del campo eléctrico y magnético de la luz al viajar dentro de la guía de onda. En Fig. 4 se muestra las formas que toma la distribución del campo eléctrico cuando las funciones Bessel son de bajo orden, visualizando así el campo transversal eléctrico (TE) de orden de modo cero, (Fig. 4a), modo 1 (Fig. 4b), y modo 2 (Fig. 4c) [22]. El modo por el que se propague la luz en la fibra también afecta como ésta interactúa con componentes que se agreguen a la fibra, como lo pueden ser las rejillas de periodo largo (LPG, de sus siglas en inglés).

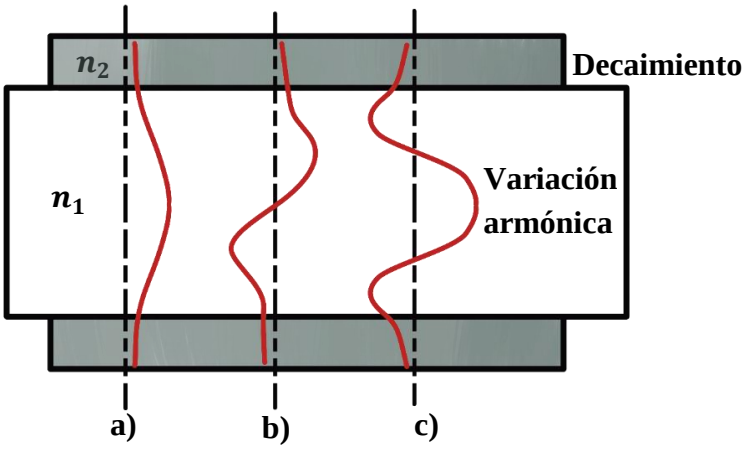

Figura 4. Distribución de campos eléctricos para modos guiados de bajo orden a) TE_0, b) TE_1, c) TE_2

Capítulo 3: Teoría de rejillas de periodo largo

Introducción

Las rejillas de periodo largo (LPG) son componentes que se fabrican en una sección sin revestimiento de la fibra óptica al causar una modulación periódica del índice de refracción (RI) [4]. Las LPG permiten el acoplamiento de los modos que se propagan en la cubierta de la fibra con el modo que se propaga en el núcleo de esta. Las LPG son utilizadas para detectar torsión, temperatura [27], tensión [28], e índice de refracción del medio circundante (SRI) [9]. Los sensores de LPG para SRI se han utilizado en áreas físicas, químicas y biológicas debido a las ventajas que las LPG proveen. Entre estas ventajas se encuentra; la no interferencia de campos electromagnéticos externos, inertidad química, y mejora de sensitividad del dispositivo. Las investigaciones que se han reportado de sensores de LPG en los últimos años han mostrado su potencial junto con una amplia área de desarrollo.

Las LPG se pueden fabricar mediante distintos métodos [4]. Actualmente los métodos para la fabricación más utilizados son por láser UV, láser de CO_2, o el método de descarga de arco eléctrico. La mayor diferencia entre las maneras de fabricar LPG es el equipo que se requiere, por ende, esto influye en la elección del método de grabado de rejilla. Otros factores relevantes para la fabricación de la rejilla es el periodo (Λ) de la rejilla, el tipo de fibra óptica, y el propósito del sensor. La elección de estos factores es crucial para la sensitividad del dispositivo.

En este capítulo se abordará la teoría de acoplamiento de las rejillas de periodo largo junto con el funcionamiento de sensores basados en LPG para SRI, dónde además de mencionarán trabajos recientes enfocados en la detección de componentes orgánicos volátiles (VOCs) dónde utilizan este tipo de dispositivos.

3.1 Acoplamiento en rejillas de periodo largo (LPG)

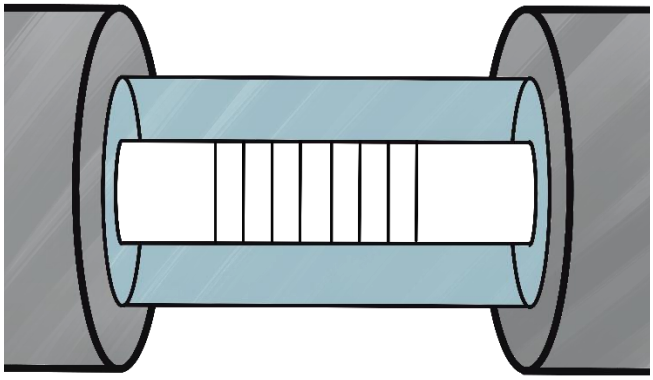

Figura 5. Representación de una LPG convencional.

Las LPG acoplan los modos ópticos del a partir de una alteración periódica del índice de refracción en el núcleo de la fibra [9]. En Fig. 5 se puede apreciar una representación gráfica de una LPG, en éstas el campo eléctrico de la luz que se acopla entre el modo guiado a través del núcleo y el de la cubierta, de acuerdo con Yu, L. se expresa de la siguiente manera [25]:

$$E(z) = \left[A_{01}^{co}(z)e^{-i\delta_m z}\right]e^{i\beta_{01}^{co}z} + \left[A_m^{cl}(z)e^{-i\delta_m z}\right]e^{i\beta_m^{cl}z}, \tag{3.1}$$

dónde $A_{01}^{co}(z)$ es la amplitud del primer modo en el núcleo, $A_m^{cl}(z)$ es el eme-ésimo modo que pasa a lo largo de la cubierta en la sección de la fibra dónde se encuentra la rejilla (0<=z<=L), L es la longitud de la LPG, β_{01}^{co} y β_m^{cl} son las constantes de propagación en el núcleo y la cubierta respectivamente. $\delta_m = \frac{1}{2}\left[\beta_{01}^{co} - \beta_m^{cl} - \frac{2\pi}{\Lambda}\right]$ es la desintonización de la longitud de onda resonante, y Λ es el periodo de la LPG [25].

Las ecuaciones de los modos acoplados de la LPG son [25]

$$\begin{cases} \frac{dA_{01}^{co}}{dz} = i[\delta_m A_{01}^{co} + \kappa_m A_m^{cl}] \\ \frac{dA_m^{cl}}{dz} = i[\delta_m A_m^{cl} + \kappa_m A_{01}^{co}] \end{cases}, \tag{3.2}$$

dónde κ_m es el coeficiente del modo linealmente polarizado m (LP_{0m}) de la cubierta. si se consideran las condiciones de contorno $A_{01}^{co}(0) = 1$, y $A_m^{cl}(0) = 0$. Cuando se tiene una LPG

convencional la amplitud de transmisión (t) y la proporción de acoplamiento de modos de LP_{0m} (r) se obtienen a partir de la siguiente ecuación [25]

$$\begin{cases} t = \cos(S_m L) + i\frac{\delta_m}{S_m}\sin(S_m L) \\ r = i\frac{\kappa_m}{S_m}\sin(S_m L) \end{cases}, \tag{3.3}$$

tal que $S_m = \sqrt{\kappa_m^2 + \delta_m^2}$. Finalmente, por conservación de la energía, en la LPG se calcula a partir de la siguiente expresión [25]

$$1 = |t|^2 + |r|^2, \tag{3.4}$$

tal que t es la amplitud de la transmisión y r es la proporción de acoplamiento. La expresión anterior ayuda a mostrar cómo ocurre la interferencia de la LPG. Los acoplamientos que ocurren del núcleo a la cubierta de la fibra que están centrados en la longitud de onda λ_i se ven descritas por [9]

$$\lambda_i = \left[n_{core} - n_{clad}^i\right]\Lambda, \tag{3.5}$$

dónde n_{core} es el índice de refracción efectivo del núcleo y n_{clad}^i es el índice de refracción efectivo del i-ésimo modo de la cubierta, tal que n_{core} es dependiente del índice de refracción del núcleo y de la cubierta, y n_{clad}^i es sensible a los índices de refracción del núcleo, cubierta, y del medio externo. Lo relevante de la teoría de LPG es que la longitud de onda de resonancia y la amplitud de las bandas de atenuación son dependientes a los factores externos a la fibra, lo cual permite ser utilizada para propósitos de medición como sensor.

3.2 Sensores de LPFG

Actualmente el uso de sensores ópticos está siendo un área de investigación profundizada para la detección de químicos y proteínas en investigaciones ambientales [29], químicas [30] y médicas [31]. Estas investigaciones utilizan un rango de longitudes de onda determinado por la fuente de luz que utilice para observar el espectro de transmisión. Las LPG causan que se observen picos de atenuación en el espectro de transmisión dónde se ve una disminución en la potencia que es medida por un analizador de espectros ópticos (OSA). Estos picos de atenuación cambian acorde el periodo de la LPG como se puede ver en Fig. 6 del trabajo de Esposito (2021) [4].

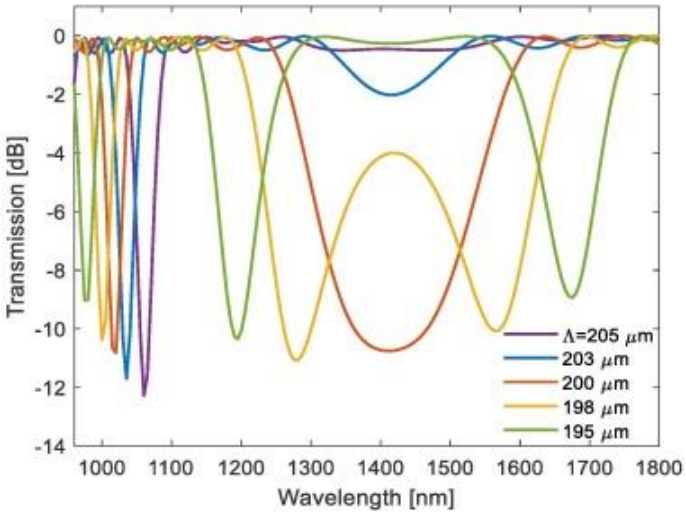

Figura 6. Espectro de transmitancia de una fibra óptica con LPG de distintos periodos [4].

Cuando se recubre a la LPG con un material con RI mayor que el de la cubierta de la fibra se induce el fenómeno de transición de modos [4]. Este fenómeno causa que se modifiquen los modos de la cubierta y genera un cambio en las bandas de atenuación. Este cambio dado por el material elegido permite al dispositivo adquirir una mayor sensitividad ante parámetros externos. La sensitividad depende de la condición de coincidencia de fases sobre el índice de refracción efectivo del modo de la cubierta [32], esta se puede utilizar para medir el IR de una solución que rodee al dispositivo, dando información de su concentración. Aquella solución o compuesto que es el objeto de medición se le denomina analito, y la elección del material debe de reaccionar ante la interacción del analito objetivo. El

comportamiento que se observa en los sensores de LPG cuando interactúan con una solución es un corrimiento de fase. La información del analito se suele obtener al observar el corrimiento en el pico de atenuación del espectro de transmisión, a esto se le conoce como el método de demodulación estándar. Este método es el que normalmente se utiliza para determinar el desempeño de un sensor. El método de demodulación estándar analiza la longitud de onda dónde se localiza el centro del pico de atenuación del espectro resultante a lo largo de las demás mediciones realizadas.

Se le denomina película sensora al material seleccionado para recubrir la LPG. La película sensora es vital para el buen desempeño del sensor, y un buen manejo de esta repercute positivamente en las mediciones que se realicen. Cada analito cuenta con múltiples películas sensoras viables para su detección, además, se ha ahondado en esta área probando el desempeño de distintos polímeros y nanomateriales. Por ejemplo; Marques (2016) utiliza nano partículas de oro con núcleo de silicio para la detección de estreptavidina, que es una proteína que se utiliza con fines de investigación biológica, obteniendo un límite de detección (LOD) de 19 picogramos sobre milímetro cuadrado (pg/mm^2) [33], Barnes (2010), utiliza un copolímero conformado por polidimetilsiloxano (PDMS) y polidimetil-octosiloxano (PMOS) para la detección de vapores de xileno para la identificación y distinción de este componente orgánico volátil con respecto a la gasolina, obteniendo un LOD de 134 ppm [34], Mientras que Rodríguez-Garciapiña (2021) utiliza PDMS para la detección de acetona con la finalidad de futuras aplicaciones médicas, con un LOD de 910 ppm [9].

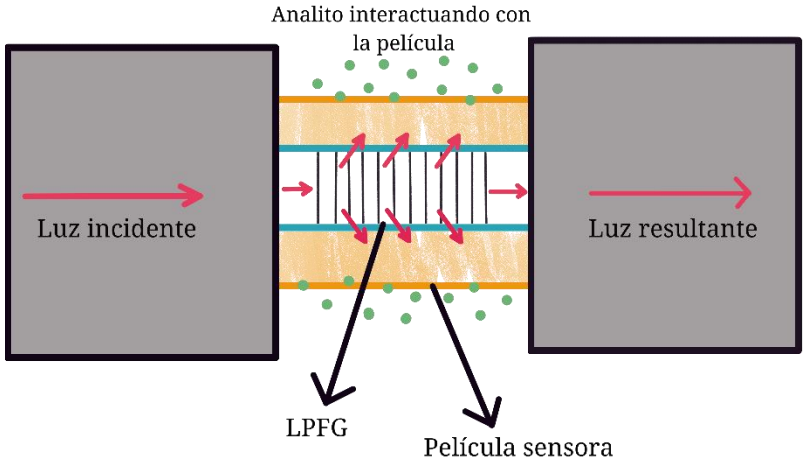

Figura 7. Representación de la interacción de la película sensora de un sensor de LPG con un analito.

3.2.1 Componentes orgánicos volátiles

Los componentes orgánicos volátiles (VOCs por sus siglas en inglés), son componentes químicos que tienen una alta presión en su estado gaseoso y una baja solubilidad en el agua [35]. Muchos de los VOCs que se encuentran en la actualidad son producidos para manufacturar pinturas, fármacos, refrigerantes, entre otros. Aquellos componentes que se encuentran entre los VOCs se encuentran los antes mencionados xileno y acetona, además se encuentran el etanol, metanol y cloroformo por mencionar algunos [35]. La detección de los VOCs es prioritaria en el sector industrial por el riesgo a la salud de los trabajadores. Como se mencionó en el capítulo 1, ya existen herramientas de alta precisión implementadas en este sector para este propósito, pero tienden a ser costosas. En el sector médico los VOCs tienen una estrecha relación con la salud de las personas [31], como se mencionó al inicio de este trabajo, algunos de estos, como la acetona, su detección presenta una opción para métodos de diagnóstico alternativos de la DM. Las opciones que brindan estos sensores aumentan su popularidad brindando resultados prometedores [5].

3.2.2 Mejora de sensitividad

Hay distintas maneras en las que un sensor de LPG tenga mejor sensitividad para su implementación [4]. Debido a criterios que se deben de cumplir en escenarios prácticos para que un dispositivo sea considerado confiable [5], se han desarrollado distintas técnicas para mejorar el desempeño de los sensores de LPG. De entre estas técnicas se encuentra; agregar componentes ópticos para amplificar la señal del sistema [34], utilizar distintos tipos de fibra, una de las opciones más comunes es la fibra codopada con boro/germanio [10], utilizar distintas técnicas para la funcionalización de la película sensora, un ejemplo es la funcionalización capa por capa de distintos materiales [32], el uso de nanopartículas en la película sensora, una de las frecuentemente utilizadas son las nanopartículas de oro [33], entre otras.

A las distintas permutaciones que se pueden realizar con los distintos parámetros posibles de estos sensores se les denomina configuración. Si se tienen dos sensores con configuraciones similares, si uno de los componentes es distinto, se tendrá un desempeño

distinto entre ambos. Corresponde al investigador determinar si la diferencia en el parámetro elegido mejora el desempeño del sensor a partir de la respuesta obtenidos de estos, y si es viable su fabricación con el equipo el que se tiene acceso.

Capítulo 4: Teoría de machine learning

Introducción

El desarrollo computacional brinda a la humanidad con herramientas para la toma de decisiones. Derivándose de la estadística, machine learning (ML) es un área que utiliza herramientas matemáticas y computacionales para ayudar a analizar información. A partir de los análisis realizados se obtiene información que permite tomar decisiones informadas. Debido al gran volumen de información que se genera diariamente, ML se ha vuelto esencial en el uso corporativo, administrativo y científico. Las herramientas que brinda ML para analizar la información son modelos matemáticos que transforman información compleja en datos que son fáciles de interpretar, encontrando relaciones ocultas en la información original y discriminando aquellos elementos relevantes dentro de los datos. Es importante tomar en cuenta que cada modelo es compatible con cierto tipo de datos, es decir, que no se puede procesar el mismo conjunto de información con todos los modelos que ofrece ML. Por ende, es necesario identificar y utilizar aquellos modelos que sean adecuados para la información que se desea procesar.

En este capítulo se explorarán las herramientas que ML provee para realizar análisis predictivo. Estas herramientas son parte de la categoría de ML denominada aprendizaje supervisado. Se explicarán los métodos de aprendizaje supervisado que se utilizarán en este trabajo de tesis, los cuales son viables para el análisis de los datos provenientes de sensores de LPFGs. Para el uso adecuado de estas herramientas es importante darles un tratamiento adecuado a los datos antes de procesarlos con los métodos de análisis. A este tratamiento previo se le denomina preprocesamiento de los datos, en el cual se verifica que no haya datos faltantes, fuera de lugar, y prepararlos acorde al método que se utilice. Además, hay variables que se pueden representar de distintas maneras para facilitar su procesamiento, en este caso se explicará la codificación de variables cíclicas. Finalmente, se hablará del método de validación cruzada. Este método es importante para determinar lo confiable que son los resultados obtenidos con los modelos y evita que se realice un sobreajuste de estos, es decir, que el modelo se acople tan bien a los datos que no sean confiables.

4.1 Codificar variables cíclicas

Una de las maneras para trabajar con variables cíclicas en el análisis de datos es codificándolas [36]. Esto es importante para facilitar la construcción del modelo matemático e incluso mejorar el desempeño de este. En el preprocesamiento de datos se debe de considerar las condiciones que deben de satisfacer los datos dependiendo del método a utilizar. En el caso que no se cumplan las condiciones del modelo afectan la credibilidad de los datos y los resultados obtenidos. De entre las técnicas más utilizadas de preprocesamiento está el centrado los datos, estandarización y normalización. Hay ocasiones donde una variable es necesario alterarla para un mejor procesamiento, en el caso de utilizar variables cíclicas, se requiere modificarlas para una mejor interpretación en análisis [37].

Si consideramos una variable cíclica como lo serían los días de la semana o los meses del año, en lugar de su respectivo nombre se suelen utilizar números para representarlos. En el caso de los días de la semana, que se utilizarán como ejemplo, usualmente se considera lunes como 1, martes como 2, hasta el domingo como 7 y agregarlos a lo largo de la variable día considerando este ejemplo, que se visualizaría como:

$$día = \begin{bmatrix} lunes \\ martes \\ miércoles \\ \vdots \\ domingo \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ 7 \end{bmatrix}.$$

Lo anterior se realizaría para poder procesar la variable día a lo largo de funciones matemáticas que son requeridas para los métodos de análisis, lo cual causa dos cuestiones relevantes al realizar el procesamiento de datos. El primero es que el modelo matemático puede tomar estos valores como si se le diera una mayor relevancia al domingo y una menor al lunes, además que al ponerlo de manera lineal no podría visualizarse de esta manera que el sábado está más cerca del lunes que el jueves, lo cual afecta al procesamiento de datos al darle un peso mayor a los últimos días de la semana al no poder expresar correctamente la naturaleza cíclica de la variable [37]. Una técnica reciente para expresar la naturaleza cíclica de este tipo de datos es colocándolos equidistantemente a lo largo de un círculo de radio 1, como se puede visualizar en Fig. 8. Para utilizar sus coordenadas polares expresadas en radianes, para identificar cada uno de los elementos conformando la variable cíclica.

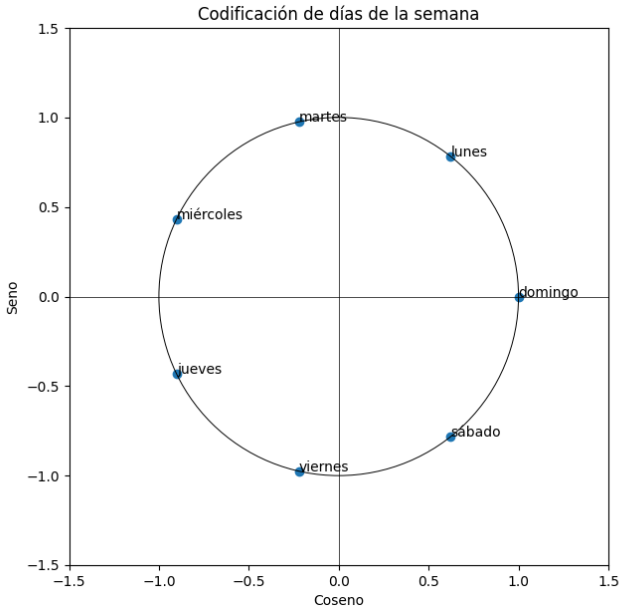

Figura 8. Días de la semana codificados.

De esta manera cada elemento de la variable día se volvería un par ordenado (cs_i, cc_i) que representarían las coordenadas del elemento a codificar, denotando la naturaleza cíclica de la variable obteniéndose de la siguiente forma [37]

$$cs_i = \sin\left(\frac{2\pi(d_i)}{n}\right), \tag{4.1}$$

$$cc_i = \cos\left(\frac{2\pi(d_i)}{n}\right), \tag{4.2}$$

dónde d_i es la representación numérica del elemento correspondiente de la variable a codificar, n es el número total de elementos que conforman a la variable original, cs_i es el elemento correspondiente codificado usando el seno del ángulo en radianes, y cc_i es con el coseno. Siguiendo con el ejemplo de los días de la semana, éstos se representarían de la siguiente manera [37]

$$día = \begin{bmatrix} lunes \\ martes \\ miércoles \\ jueves \\ viernes \\ sábado \\ domingo \end{bmatrix} = \begin{bmatrix} 0.781 & 0.623 \\ 0.974 & -0.222 \\ 0.433 & -0.9 \\ -0.433 & -0.9 \\ -0.974 & -0.222 \\ -0.781 & 0.623 \\ 0 & 1 \end{bmatrix},$$

dónde la columna izquierda corresponde a cs y los de la derecha con cc. Esto mismo se puede realizar si se utilizan meses, horas, minutos, o cualquier variable que se denote una naturaleza de este tipo para el experimento o suceso del que se esté recopilando la información, así permitiendo una mejora en el rendimiento del procesamiento de los datos.

4.2 Aprendizaje supervisado

Aquellos modelos que entran en la categoría de aprendizaje supervisado "aprenden" a partir de datos "etiquetados" [38]. Esto significa que los modelos se construyen a partir de datos que tengan alguna clase de clasificación arbitraria, esto también puede ejemplificarse de la siguiente manera; cuando a partir de una recopilación de información se tienen datos de entrada que corresponden a ciertos datos de salida Fig. 9.

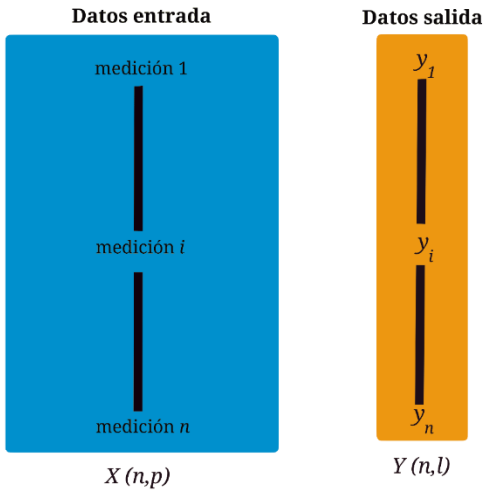

Figura 9. Representación de la composición de los datos de entrada y de salida.

Este tipo de aprendizaje es ideal para realizar tareas de predicción o de clasificación. Esto porque aquellos modelos dentro de esta categoría toman cierta cantidad de datos denominando a este grupo datos de entrenamiento A_n tal que [38]:

$$A_n = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n, \tag{4.3}$$

dónde $\mathcal{X}$ es el conjunto de datos entrada y $\mathcal{Y}$ el conjunto de datos de salida, y n es el número de elementos totales. A partir de A_n se construye una función *f tal* que relacione los datos de

entrada con los de salida acorde al método matemático utilizado para el análisis y se evalúa el desempeño de la función construida con datos no utilizados para su aprendizaje que se denominan datos de validación para determinar la certeza y confiabilidad del modelo para luego utilizar ese mismo modelo en datos nuevos que se vayan recopilando, acelerando el proceso de clasificación o predicción que se vaya a realizar.

Gradualmente la cantidad de información que se va recopilando requieren de una mayor complejidad para aumentar la capacidad de predicción para reducir los márgenes de error y obtener resultados más certeros. Gracias al desarrollo continuo de esta área permiten que modelos que cuentan con varias décadas desde su creación se adapten a las necesidades y criterio de análisis actuales. Aprendizaje supervisado cuenta con distintos métodos que tienen acercamientos distintos y podrían dar distintos resultados para un mismo A_n. Dependiendo también de la naturaleza de A_n también se considera si un método es adecuado o no para los datos recopilados. Por lo que en algunos casos es importante probar múltiples modelos para analizar el desempeño de éstos [38].

A continuación, se presentarán los modelos de análisis supervisados que son aptos para datos de sensores basados en fibra óptica, regresión de componentes principales, regresión de proyección de estructuras latentes, y regresión de bosque aleatorio.

4.2.1 Regresión de componentes principales

La regresión de componentes principales (PCR, por sus siglas en inglés) es un método de análisis supervisado es de los más utilizados por la facilidad de su implementación junto con la sencilla interpretación del conjunto de datos procesados. Este método consta de la unión de los métodos de; análisis de componentes principales y regresión lineal múltiple [39].

I. Análisis de componentes principales

El análisis de componentes principales (PCA, por sus siglas en inglés) es un método exploratorio que reduce la dimensión de los datos permitiendo visualizar algún comportamiento oculto entre los datos [13]. PCA consiste en obtener las componentes principales de la matriz de datos construida a partir de la información recopilada. Estas componentes son ejes perpendiculares dónde los datos se distribuyen con respecto a su

varianza, entre mayor varianza haya a lo largo de una componente principal se considera que contiene más información con respecto a los datos originales. Las componentes principales se obtienen a partir del método conocido como descomposición de valores singulares (SVD por sus siglas en inglés) [40], la cual se basa en que cualquier matriz se puede escribir como una operación aritmética de 3 matrices [40]

$$X = U\Sigma V^T, \tag{4.4}$$

dónde X es la matriz de datos, $U = XX^T$, $V = X^T X$ y Σ es la matriz de valores singulares de V. Esto quiere decir que se buscan los valores y vectores propios de V, pero a partir de la obtención de estos valores y vectores, se ordenan los valores propios en orden descendente acompañados por su respectivo vector propio. Teniendo los vectores propios ordenados éstos se volverán los pesos para obtener los valores rotados a lo largo de cada una de las componentes principales (PC) [14]

$$PC_j = X \cdot w_j, \tag{4.5}$$

dónde w_j es el j-ésimo vector propio para la obtención de PC_j que es la j-ésima componente principal [13]. Los valores propios nos dicen el porcentaje de varianza que contiene la componente principal, y usualmente después de realizar este procesamiento de los datos la mayoría de ésta se encuentra en las primeras componentes principales. Lo cual es lo que se busca con PCA. Al reducir la dimensionalidad de los datos originales reteniendo una cantidad concreta de componentes principales que mantienen la mayoría de la información de los datos, la cuestión entra en cuántas de estas componentes deben de retenerse para no tomar menos de las necesarias para tomar decisiones informadas, o un mayor de las que se requieren considerando información que puede ser irrelevante para la toma de decisión.

II. Regresión lineal múltiple

La regresión lineal múltiple (MLR, por sus siglas en inglés) es un método que busca una recta en el espacio de los datos que tenga la mínima distancia entre ésta, y todas las mediciones realizadas relacionando a los datos con una variable que se busca predecir [14]. Para esto MLR se basa en la siguiente ecuación [14]:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_n x_{in}, \tag{4.6}$$

dónde β_0 es la intercepción con el eje representando la variable a predecir *y*, los demás β_j son coeficientes de regresión representando la variable *j* que tiene la medición *i* a la cual se le relaciona con el valor $\hat{y}_i$ que serán denominados los valores predichos. Se va a encontrar una diferencia entre los valores predichos con los valores originales, esta diferencia puede tomar distintos valores estadísticos para evaluar el desempeño del método, uno de los más utilizados es el error cuadrático medio (MSE) que se calcula de la siguiente manera [6]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4.7}$$

tal que y_i es el valor original que se busca predecir con MLR. Combinando MLR y PCA se obtienen las predicciones con respecto a las componentes principales, es decir, PCR.

III. Regresión de componentes principales

El objetivo de la regresión de componentes principales (PCR, por sus siglas en inglés) es realizar predicciones de una variable objetivo a partir del conjunto de datos deseado [14]. PCR primero reduce la dimensión de los datos con PCA para retener la información importante del sistema, y luego de las PC seleccionadas y luego con los datos transformados se les aplica MLR para la predicción. Por lo que debería de realizarse lo siguiente:

$$\hat{y} = \beta_0 + \sum_{j=1}^{k} \beta_j \cdot PC_j \tag{4.8}$$

Aquí, *j* es el número de componentes principales que se desea retener. Para saber cuántas componentes principales retener hay varias maneras. Para un análisis rápido se utiliza la regla en la cual se deben de tener el mínimo de componentes que la suma de porcentajes de sus varianzas esté entre el 80% y el 90%, las demás se descartan. Otra regla que se suele utilizar

es retener aquellas componentes dónde la varianza sea mínima del 1% [41]. Este tipo de reglas arbitrarias llevan que se sobreestimen o subestimen el número de componentes a retener, para realizar una selección acertada se utiliza la validación cruzada.

En la validación cruzada se evalúa un parámetro arbitrario al procesar particiones de datos en el modelo de análisis utilizado. En este trabajo se evalúa la raíz del error cuadrático medio (RMSE, por sus siglas en inglés) obtenido al evaluar las particiones con diferente número de componentes principales. Una vez comparados los RMSE obtenidos se determina el número de componentes principales que proveen un modelo optimizado acorde a el propósito del trabajo y de los criterios establecidos por el investigador, esto se ahondará más adelante en la respectiva sección de validación cruzada.

4.2.2 Regresión de proyección de estructuras latentes

Este método de análisis supervisado busca componentes que sean buenos predictores tanto para los datos de entrada (X) como los de salida (Y) [15] Este método tiene una gran similitud a PCR [42], también realiza reducción de dimensión de los datos, y busca variables latentes para describirlos [43]. Pero, a diferencia de PCR, la proyección de estructuras latentes no solo busca aquellas que solamente describan X, sino que también describan Y y encuentren una fuerte relación entre éstos. Este método tiene una estrecha relación con el método de proyección de estructuras latentes [8].

I. Proyección de estructuras latentes

La proyección de estructuras latentes (PLS, por sus siglas en inglés) es un modelo exploratorio que tiene similitudes con PCA [43]. Realiza reducción de dimensionalidad sobre la matriz de datos y los representa en un nuevo eje coordenado reteniendo la mayor cantidad de información de estos. Los métodos PLS son una familia amplia, de la cual el más popular es PLSR, pero no es el único. A diferencia de PCA, que busca explicar variables latentes con respecto a la varianza entre los datos, PLS encuentra las variables que expliquen mejor la relación entre la matriz de datos X y la variable objetivo Y. Esto se realiza al extraer un único

dos grupos de puntuaciones T y U desde X y Y simultáneamente [42]. Cada una de estas puntuaciones se expresa de la siguiente manera [42]

$$t_i = X_i w_i, \tag{4.9}$$

$$u_i = Y_i c_i, \tag{4.10}$$

dónde w_i son las cargas de la i-ésima estructura latente (LS) de X y c_i son las cargas de la i-ésima LS de Y. Se busca por medio de PLS es que las puntuaciones de X y Y tengan máxima covarianza [42]

$$Cov(t_i, u_i) = \mathcal{E}\{(t_i - \bar{t}_i)(u_i - \bar{u}_i)\}, \tag{4.11}$$

dónde $\mathcal{E}\{z\}$ es la media poblacional tal que [42]

$$\mathcal{E}\{z\} = \frac{1}{N}\sum z. \tag{4.12}$$

dónde N es el número total de elementos, en el contexto de este trabajo, el número total de mediciones.

La interpretación geométrica de las estructuras latentes es que van orientándose en las direcciones de mayor covarianza descendentemente, es decir, que la primera estructura latente representa el eje dónde la covarianza entre los datos es mayor, la segunda estructura latente es la segunda mayor, y así consecutivamente. Por lo que usualmente el investigador se suele quedar con las primeras 3 estructuras latentes porque son las que retienen el mayor porcentaje de información y se pueden visualizar para observar el comportamiento de los datos.

II. Regresión de proyección de estructuras latentes

La regresión de proyección de estructuras latentes (PLSR, por sus siglas en inglés) difiere ligeramente a PLS en su algoritmo [43]. Ya que PLSR es un modelo predictivo, utiliza las estructuras latentes para utilizar datos que no sean parte de X y obtener un valor correspondiente a los datos de salida, para esto se requieren unos valores denominados como puntuaciones. Para esto, PLSR busca múltiples elementos por medio de un algoritmo iterativo. Este modelo dónde se encuentran los elementos requeridos se visualiza en Fig. 10.

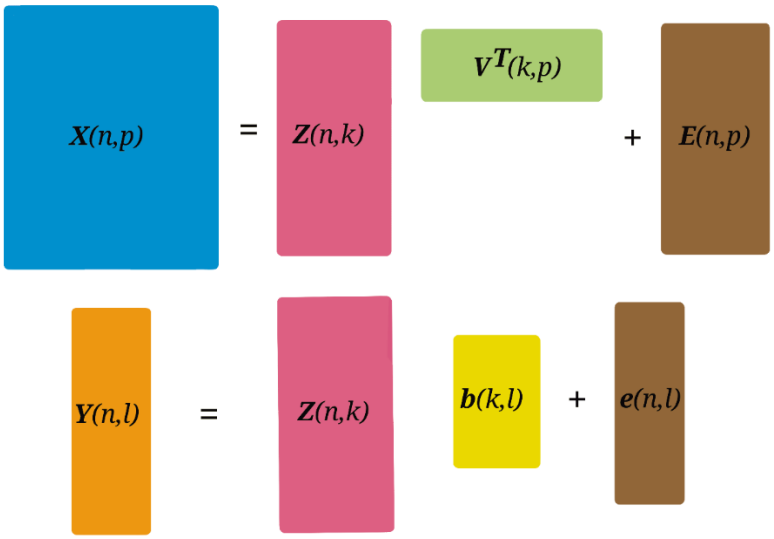

Figura 10. Estructura PLSR

Tal que se puede expresar como [43]:

$$X = ZV^T + E, \tag{4.13}$$

$$Y = Zb + e, \tag{4.14}$$

dónde Z es la matriz de puntuaciones de PLSR, V es la matriz de cargas, E es la matriz de residuales de X, b es el vector de coeficientes de regresión, y e es el vector de residuos de Y [43]. A partir de la matriz de puntuaciones Z se extraen las componentes, conocidas también como estructuras latentes, del modelo. Tal que $Z_k = [z_1, \cdots, z_k]$ dónde k es el número de componentes que se desea extraer, dónde cada una de las componentes se obtiene de la siguiente manera [43]

$$z_i = \frac{Xw_i}{w_i^T w_i}, \tag{4.15}$$

$$w_i = \frac{\widetilde{w}_i}{\|\widetilde{w}_i\|}, \tag{4.16}$$

$$\widetilde{w}_i = \frac{X_a^T Y}{Y^T Y}, \tag{4.17}$$

dónde X_a es la matriz de datos que pasa por un proceso que en inglés se conoce como "deflating the matrix". Este proceso se realiza cada que se obtiene una de las componentes,

por esto es por lo que se le dice que es un proceso iterativo, ya que, con cada obtención de una componente, la siguiente tendrá una X_a distinta para su construcción.

PLSR es un modelo iterativo que cuenta con principios similares a los de PCR, pero las distinciones en su implementación generan ventajas sobre PCR. La principal de las ventajas es que la correlación entre los datos de entrada y de salida no debe de ser baja entre éstos, permitiendo predicciones más precisas que se obtienen con los coeficientes calculados con la siguiente ecuación [43]

$$b_i = \frac{y^T z_i}{z_i^T z_i}. \tag{4.18}$$

A partir de los coeficientes obtenidos para cada componente dentro del rango a extraer, se obtienen las predicciones del modelo de la siguiente forma:

$$\hat{y} = Z\vec{b}, \tag{4.19}$$

dónde $\vec{b} = [b_1, \cdots, b_k]$ y $\hat{y}$ es el vector de predicciones, a partir de las predicciones realizadas se realiza la evaluación del modelo por medio de RMSE y R^2 [13].

4.2.3 Regresión de bosque aleatorio

El modelo de bosque aleatorio (RF, por sus siglas en inglés) es parte de los métodos de conjunto [45]. Los cuales consisten en métodos que utilizan múltiples modelos que se aglomeran para predicción o clasificación. RF se basa en utilizar múltiples modelos de árboles de decisión. Estos métodos se utilizan tanto para la predicción o clasificación del conjunto de datos deseado. RF es un método que se acerca mucho al término de caja negra dentro de ML ya que el mismo modelo crea múltiples condiciones que el investigador le tomaría mucho tiempo analizar detrás de los datos obtenidos.

I. Árboles de decisión

Los árboles de decisión (DT, por sus siglas en inglés) son un modelo que se utiliza para tareas de clasificación y de predicción [43]. Estos aplican un grupo de reglas que permiten al investigador asignar valores en los datos a una clase o agruparlos dentro de alguna condición. Estas reglas se acomodan en una estructura jerárquica, denominados nodos, que se visualiza como el diagrama de un árbol invertido en Fig. 11. El nodo inicial se le llama nodo raíz dónde se encuentran los datos que serán evaluados por medio del DT. Este se desglosa en "ramas" que imponen una condición a los datos y los asigna a nodos internos. Cuando un nodo interno se desglosa por varias ramas a más nodos, se le conoce como nodo padre, y a los demás, nodos hijos. Cuando un nodo no se desglosa en más nodos, éste se vuelve un nodo hoja o terminal. En los nodos terminales es donde se tiene una clasificación o predicción de los datos que, debido a las condiciones de los nodos internos, llegaron a ese nodo terminal. Para la construcción de un árbol de decisión se requiere de tres puntos importantes. Asignar el número de veces que un nodo se puede desglosar, definir un juicio para encontrar la mejor partición de un nodo, y establecer reglas para que un nodo se defina como uno interno o terminal [44].

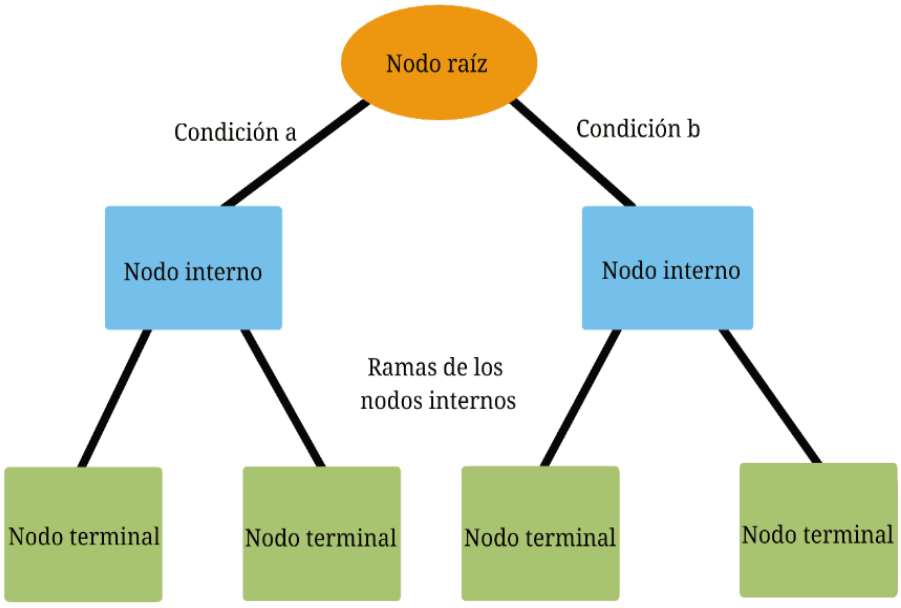

Figura 11. Ejemplo de un árbol de decisión binario [44].

El primer punto para la construcción de un DT acorde a Sanchez, G. (2024), es considerar la naturaleza de las variables que se vayan a utilizar [43]. Debido a que el número y el tipo de posibles particiones que se puedan realizar depende del tipo de variables que se utilicen. Las variables binarias solamente se pueden particionar en dos nodos hijos separando los respectivos valores de la variable binaria como se puede ver en Fig. 12.

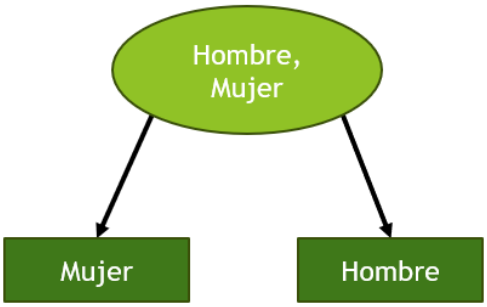

Figura 12. Ejemplo partición de variable binaria en DT binario.

De una variable nominal el número de diferentes particiones binarias que posee se obtiene de la siguiente expresión [43]

$$2^{q-1} - 1, \tag{4.20}$$

dónde q es el número de elementos que conforman a la variable nominal, como ejemplo si se tiene una variable nominal que cuenta con tres elementos, tendrá tres posibles maneras en las que se puede particionar. Las variables ordinales al tener que respetar un orden preestablecido, el número total de particiones binarias que poseen son [43]

$$q - 1 \tag{4.21}$$

Las variables continuas se particionan acorde a los valores que toma esa variable en los datos. Si estas son un número razonable acorde al investigador, se utiliza la expresión (4.21), considerando la variable como si fuera una nominal. En el caso que la variable tome un número extenso de valores, se toma un valor q^* menor que q y se sustituye en Ec. (4.21).

El segundo punto para la construcción de un DT es el de definir un criterio óptimo para particionar los datos [43]. Esto además requiere de otro criterio para optimizar los subgrupos en los que se particiona un nodo padre, tal que haya la mayor reducción de heterogeneidad entre los nodos hijos. La heterogeneidad denota las impurezas de los datos en el nodo, y por

norma, el nodo padre tiene mayor heterogeneidad que los nodos hijos y hay distintas maneras de medir la impureza de un nodo.

Una de las maneras de medir las impurezas de un nodo es a través de la entropía. La entropía en el contexto de ciencias computacionales, y específicamente en DT, se expresa de la siguiente forma [43]

$$H(nodo) = -\sum_{k=1}^{K} p_k \log_2(p_k) \tag{4.22}$$

dónde p_k es la probabilidad de seleccionar de manera aleatoria un objeto que provenga de una clase k. Es decir, un valor que pertenezca a un determinado grupo de datos. Un bajo valor de entropía significa que predominan los objetos de una clase y hay baja impureza en el nodo, en caso contrario se tiene un nodo con una alta impureza. Para determinar si una regla es adecuada para un grupo de datos utilizando partición basada en entropía, se debe de comparar la entropía de los nodos hijos con el del nodo padre, si la entropía de los nodos hijos es igual a la del nodo padre, el criterio de partición no es apropiado.

Para el tercer punto se debe definir una regla que determine cuando un nodo es terminal [43]. La repetición de los dos puntos anteriormente descritos es lo que conforma el árbol, pero para la realizar una clasificación de los datos se deben de determinar los nodos terminales. De entre las condiciones para detener la partición de un nodo es si este es puro, es decir que todos los objetos del nodo pertenecen a una sola clase, no hay más características en las que se pueda particionar, si el nodo es el último dentro de condiciones preestablecidas para la construcción del DT, o si el tamaño de los nodos resultantes sería demasiado pequeño. Para evitar un sobreajuste del DT se suele preestablecer un número de nodos máximos que se pueden crear en el DT.

II. Bosque aleatorio

Aunque el modelo de DT tiene un gran desempeño en realizar predicciones, solo las tiene para aquellos datos con los que fue construido [43]. Puede darse el caso en el que se tenga un error de cero al evaluar un DT con los datos utilizados para su fabricación, pero esto causa que su desempeño sea deficiente con otros grupos de datos. Hay una relación entre la

profundidad del DT y el porcentaje de error al predecir datos nuevos, entre mayor profundidad, mayor es el error. Pero cuando se tiene un DT con una baja profundidad, hay un riesgo de sub ajustar el modelo, por ende, el modelo tiene un bajo desempeño con los datos y no representa un resultado fiable. Para resolver esta problemática de los DT es necesario explicar la definición del término "bagging"

El bagging es un acrónimo del método "boostrap aggregating" publicado en 1990 [43]. Este método obtiene M grupos de datos a partir de uno $\mathcal{D}$, y de cada uno se obtiene un modelo h_m. Estos grupos de datos $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$, que contienen valores tomados al azar de $\mathcal{D}$, se vuelven datos de entrenamiento al ser evaluados por su respectivo modelo h_m, esto se puede visualizar en Fig. 13. El bagging ayuda a reducir la varianza a que si solamente se utilizara un único modelo sobre $\mathcal{D}$, lo cual significa que reduce la posibilidad de sobreajuste en los datos.

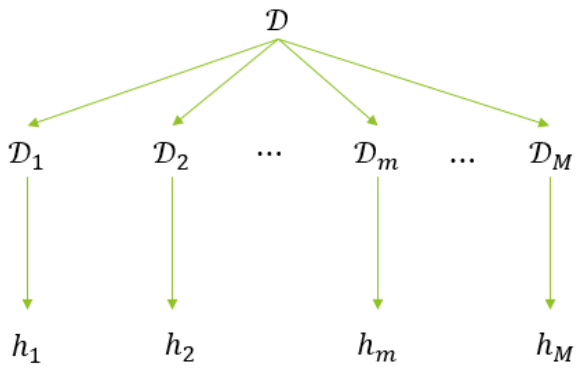

Figura 13. Diagrama del proceso de Boostrap aggregating (Bagging).

Un bosque aleatorio (RF, por sus siglas en inglés) es un grupo de M DTs construidos a partir de un conjunto de datos $\mathcal{D}$ que fue dividido en M grupos [45]. Para realizar tareas de predicción, es decir, tener un modelo de RFR, se realiza el promedio de las predicciones obtenidas en cada uno de los DT [43]

$$\hat{y}_0 = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_0^{(m)}. \tag{4.23}$$

El desempeño del RFR está ligado a varias de las condiciones para la construcción de DTs. Entre aquellos parámetros que afectan directamente al modelo RFR está el tamaño de

los DT, el número de características que debe de considerar cada DT, el número de elementos mínimos para conformar un nodo y el número M por mencionar algunos.

4.3 Validación cruzada

La validación cruzada (CV, por sus siglas en inglés) se utiliza para comprobar el desempeño de un modelo sobre los datos deseados [46]. Su uso principal es el determinar que un modelo de análisis supervisado no se sobre ajuste, es decir que llegue a un punto que no se desempeñe adecuadamente con nueva información [47]. Este proceso requiere de datos de entrenamiento y datos de validación, los datos de entrenamiento es la información con la que queremos que el modelo de análisis "aprenda". Mientras que los datos de validación se utilizan para comprobar si el modelo es adecuado para predecir o clasificar información nueva.

El método de CV más usual se le denomina CV de k-pliegues, o en inglés k-fold CV [47]. Este método particiona los datos de entrenamiento seleccionados, los revuelve aleatoriamente, y luego los particiona en k partes de tamaño similar. Luego una de las partes particionadas será designada como la partición de prueba, y las demás como particiones de entrenamiento. Se simulará el desempeño del modelo con datos nuevos partir de su construcción con las particiones designadas a entrenamiento, y evaluándolo con la designada a la prueba. Después se selecciona otra partición diferente para la de prueba obteniendo el valor de evaluación designado por el investigador, RMSE, error medio absoluto (MAE), entre otros. Finalmente se promedia este valor obtenido de cada iteración para su interpretación.

Hay parámetros en los modelos de análisis supervisado que afectan el desempeño de estos, por ejemplo; para PCR es el número de componentes principales (PCs, pro sus siglas en inglés), para PLSR las estructuras latentes (LSs, por sus siglas en inglés). Por lo que para la mejora del desempeño de éstos se utiliza CV y conocer cuál es el número óptimo de PCs y LSs para cada caso [42].

Capítulo 5: Materiales y desarrollo experimental

5.1 Materiales

Los materiales utilizados para la realización de este trabajo fueron; fibra SMF-28 con un núcleo de 8.2 micrómetros de diámetro y un revestimiento de 125 micrómetros de diámetro, polidimetilsiloxano (PMDS, CAS 9016-00-6) como reactivo químico, y acetona como ($(CH3)_2CO$, CAS 67-64-1) como analito. Un diodo superluminiscente con rango espectral entre 1400 - 1550 nm, analizador de espectros ópticos (OSA, Ando AQ6315 AA 9057), una cámara de teflón de 1L y una computadora personal para la captura de datos. Actuador lineal (Zaber, TLA 28 A) con precisión de 0.1 micras, empalmadora Furukawa, y una cortadora de fibra óptica monomodo (FiTel S326) para la fabricación de los sensores.

5.2 Desarrollo experimental

5.2.1 Fabricación de los sensores

Para la fabricación de cada sensor se utilizaron 80 cm de fibra SMF-28. A la mitad de cada fibra se designó una sección de 5 cm a la cual se le quitó el revestimiento para grabar la LPG. El arreglo experimental hecho para el grabado de la LPG por medio de arco eléctrico es el que se visualiza en Fig. 14.

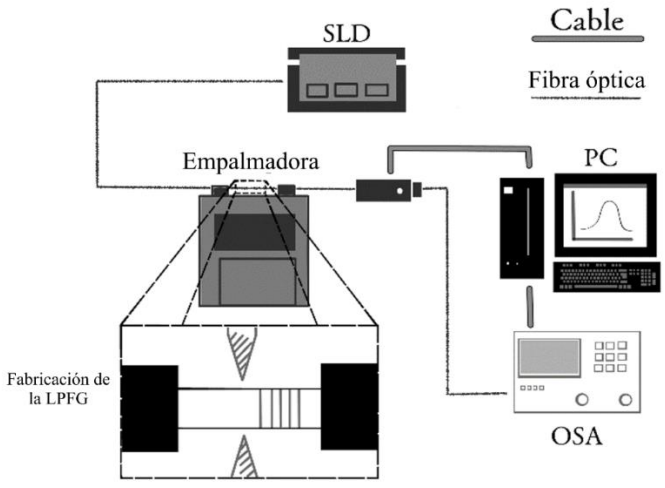

Figura 14. Arreglo experimental para el grabado de las LPFG.

Mediante el uso del actuador se realizaron las separaciones de periodo para cada rejilla. Además, se monitoreó el espectro de transmitancia a lo largo del grabado, con el fin de identificar anomalías que llegaran a ocurrir durante este proceso. Como podría ser una pérdida drástica en la potencia debido al derretimiento de la fibra óptica. Un total de 41 descargas se realizaron para la conformación de las rejillas en cada uno de los sensores. Una vez completado el proceso de grabado de la rejilla, se realizó la funcionalización del polímero PDMS. Para esto se distribuyó el PDMS a lo largo de la zona de la rejilla cubriéndola en su totalidad, luego se calentaron a 100 °C durante 30 minutos y se dejaron reposar durante dos días, este proceso se realizó como se visualiza en la Fig. 15. Finalizado el proceso de funcionalización, los sensores estarían preparados para la recolección de datos.

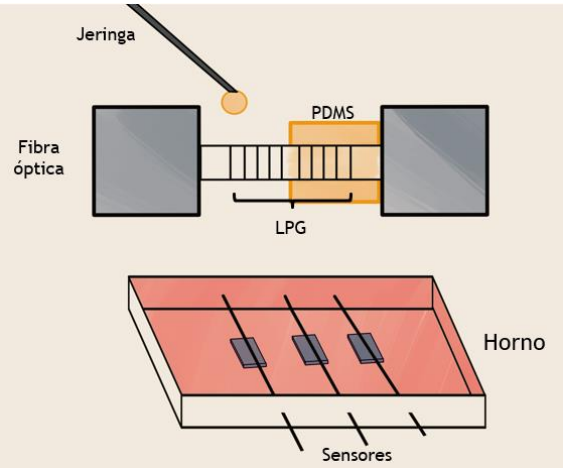

Figura 15. Funcionalizado de PDMS por método de inmersión.

Algo muy importante a considerar durante la fabricación de los sensores es la dificultad en su reproducibilidad. Esto por el método de arco eléctrico para el grabado de las rejillas. Durante la descarga eléctrica sobre la fibra se presentan variaciones no controladas que afectan al grabado de la LPG. Por lo cual, aunque se tengan dos LPG con el mismo número de puntos y grabadas con los mismos parámetros, se pueden presentar diferencias notables entre los espectros de salida de ambas LPG.

5.2.2 Obtención de los datos

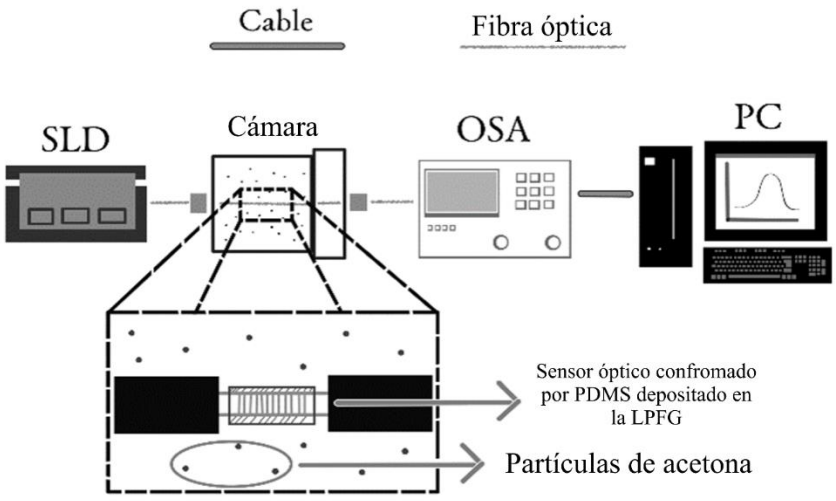

Figura 16. Arreglo experimental para la toma de datos.

Para la detección de datos se realizó el arreglo experimental mostrado en Fig.16. Dónde se colocó el sensor dentro de la cámara de teflón cuidando de no contaminarlo al ponerlo en contacto con alguna superficie. Después se acopló el sensor al sistema con el diodo superluminiscente y el OSA, el cual está conectado a la PC para la toma de datos del espectro de salida. Para analizar cómo el espectro de salida cambia según la concentración de acetona con la que interactúa el sensor, cada 10 minutos se insertó en la cámara 0.1 µL de acetona hasta llegar a 1 µL. Para realizar la conversión de unidades de µL a ppm se utilizó la siguiente ecuación [9]

$$C_{ppm} = \frac{22.4 T_a \rho_a V_l}{273 M_w V_c} \times 10^3 \qquad (44)$$

Dónde T_a es la temperatura a la que ocurre la interacción en grados Kelvin, ρ_a es la densidad de la acetona (g/mL), V_c es el volumen de gas de la cámara de teflón que es 1 L, V_l es el volumen insertado de acetona (µL), M_w es el peso molecular de la acetona (g/mol). Finalmente 22.4 L/mol es el volumen de 1 mol de gas a condiciones estándar (1 atm 273 K) [9]. Sustituyendo los valores correspondientes se calculó que 1 µL de acetona es equivalente

a 332 ppm. Durante el tiempo en el que se realizaron las deposiciones de 33.2 ppm de acetona, se automatizó la toma de datos con el OSA para observar los cambios del espectro a cada minuto.

5.2.3 Procesamiento de datos

Después de la captura de datos se realizó su procesamiento con herramientas de análisis supervisado. Ya que cada espectro cuenta con un valor de concentración al que está "etiquetado" la aplicación de estas herramientas no solamente es viable, también tienen la capacidad de mejorar el desempeño de los sensores sin tener que intervenir con las características físicas de los mismos. Los métodos que se utilizaron fueron PCR, PLSR y RFR con el objetivo de analizar el comportamiento de los sensores y verificar si son aptos para realizar predicciones obteniendo límites de detección (LOD) que entren en el rango deseado.

Para implementar estas herramientas primero se requirió recopilar los datos y aplicar su preprocesamiento correspondiente. En este tipo de casos dónde el experimento cuenta con una alta influencia por parte del tiempo, además se considera de manera cíclica durante la toma de mediciones, ya que se va repitiendo de minuto 1 al minuto 10 en concentraciones distintas. Por lo que es necesario tomarlo en cuenta durante el procesamiento de los datos, codificando los minutos para un mejor rendimiento. PCR y PLSR ambos son modelos que son susceptibles al centrado y la uniformidad entre los datos. Como se están considerando el tiempo codificado en cada medición, es necesario estandarizar los datos para ambos métodos, para RFR esto último es innecesario debido a que no le afecta negativamente datos no centrados o uniformes. Para determinar el número apropiado de PC y LS a retener en cada caso se utilizó CV. Se optó por considerar el número mínimo de componentes necesarias para explicar los datos, es decir, que ocurra una disminución considerable en el RMSE, pero, que posterior a ese número de componentes no ocurra un cambio significativo, de esta forma evitando el sobreajuste del modelo.

Para RFR, se realizó lo que en inglés se denomina como "hyperparameter tuning". Dado un grupo de parámetros seleccionados, buscar aquellos valores que optimicen el rendimiento del modelo y mejoren su desempeño. Los parámetros que se seleccionaron para su búsqueda

de valores fueron; el número de DT que se construyen durante la implementación del modelo, el máximo de profundidad que pueden alcanzar, el mínimo de elementos para que se desglose un nodo, y la cantidad máxima de características aleatorios de los datos a utilizar para la creación de cada DT. Como se puede apreciar en la descripción de cada uno de los parámetros anteriores, estos no solamente afectan a la precisión del modelo, también afectan los recursos computacionales que requieren. Finalmente se aplicaron los modelos utilizando los respectivos parámetros optimizados para obtener el LOD que alcanzó cada uno de éstos en cada sensor. Además, se realizó la comparación entre los LODs teóricos obtenidos por los métodos utilizando 1 CP y 1 LS para PCR y PLSR, mientras que valores estándar para los parámetros de RFR para obtener el porcentaje de mejora con respecto a sus valores optimizados.

Capítulo 6: Resultados y discusión

Los espectros de transmisión obtenidos durante las mediciones se pueden visualizar en Fig. 17. Se puede apreciar en cada uno de los sensores que hay un cambio en la amplitud y un corrimiento espectral conforme va aumentando la concentración de acetona. En la Fig. 17a correspondiente al espectro del sensor 1 se ve una clara diferencia entre mediciones a distintas concentraciones. Cuenta con un pico de atenuación centrada en 1495 nm, al realizar un acercamiento sobre este como se ve en Fig. 17b el comportamiento mencionado se resalta centrándose en la última medición en 1499 nm, mostrando un corrimiento de longitud de onda de 4 nm. Realizando un análisis visual y tomando en cuenta el método de demodulación tradicional reportado por la mayoría de los autores, se observa que este sensor tuvo buena respuesta ante la interacción con acetona. Para el sensor 2, que se visualiza su espectro de transmisión en Fig. 17c, se observan dos picos de atenuación, uno centrado en 1429.8 y otro centrado en 1506. El segundo pico de atenuación solamente presenta cambio de amplitud, pero no cuenta con corrimiento en su espectro, el primer pico de atenuación, que se observa su acercamiento en Fig. 17d, muestra un corrimiento en longitud de onda de 1.2 nm. En la última medición este pico de atenuación queda centrado en 1428.6 nm. En este sensor se ve un cambio de amplitud relevante cuando éste está interactuando con la acetona a bajas concentraciones, sin embargo, cuando el sensor interactúa a altas concentraciones de acetona se ve una superposición entre las amplitudes. A un primer vistazo, se observa que el desempeño de este sensor para la detección de acetona a altas concentraciones no fue adecuado. PLS dará más información de la respuesta de este sensor. Por otra parte, el sensor 3, del cual se ve su espectro de transmisión en Fig. 17e, muestra que el cambio en amplitud es pequeño a comparación de los otros sensores, ya que aquí no se ve un cambio significativo entre las mediciones. Éste tiene dos picos de atenuación, el primero centrado en 1402.4 nm en la primera medición y en la última en 1403.2 nm teniendo un corrimiento de 0.8 nm. El segundo pico de atenuación está centrado en 1516.2 nm en la primera medición, y en la última en 1516.8 nm con un corrimiento total de 0.6 nm. Al realizar un acercamiento a este último como se ve en Fig. 17f, se ve una clara distinción entre mediciones de diferentes concentraciones como en el sensor 1, con variaciones más pequeñas en comparación con el primer sensor. Lo cual podría llevar a la idea preliminar que el desempeño de este sensor fue

pobre y que la respuesta a la acetona no es significativa. Lo anterior es analizando el espectro de los sensores acorde a técnicas tradicionales considerando un punto del espectro. Pero el análisis resulta distinto cuando se utilizan herramientas de aprendizaje supervisado.

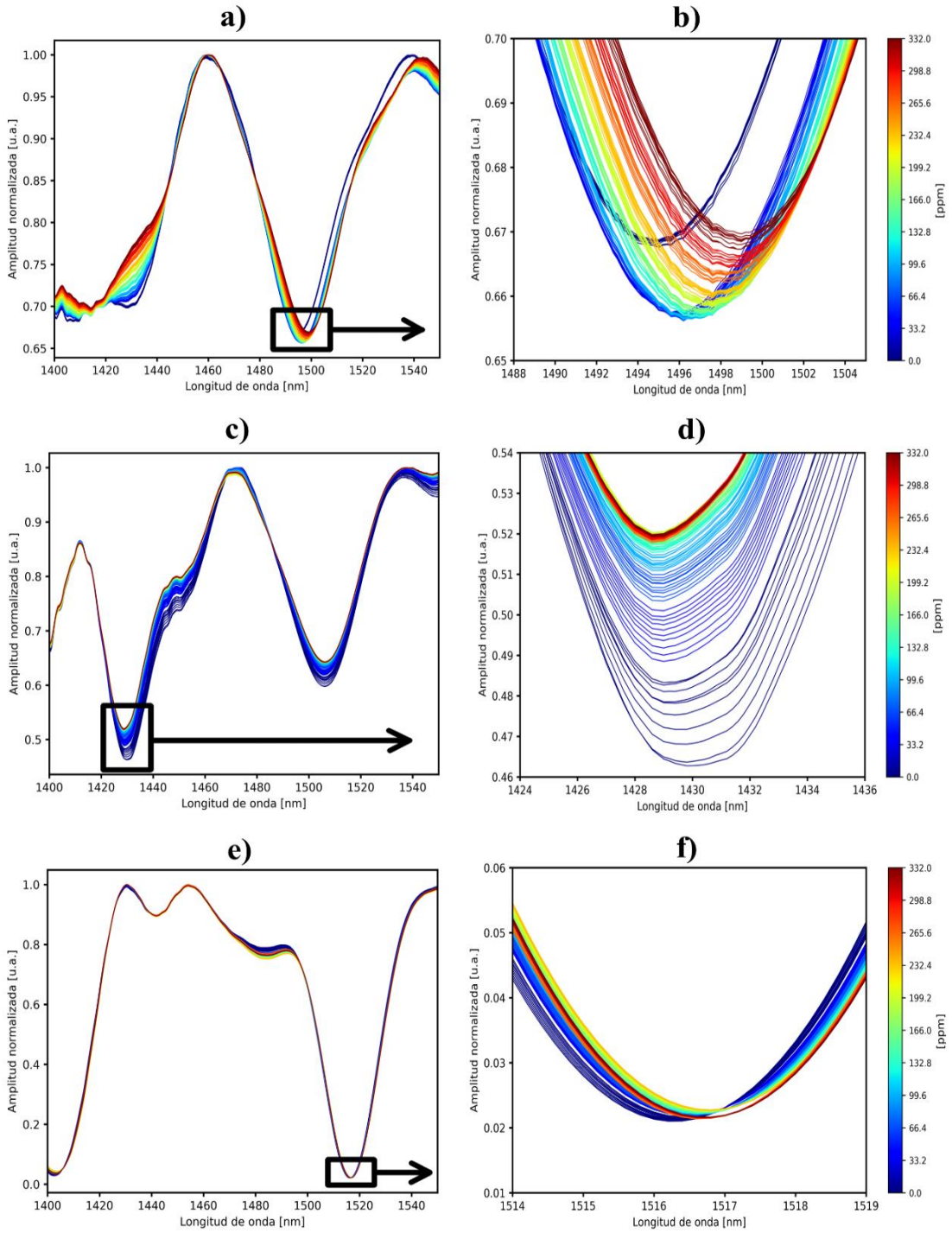

Figura 17. Espectros normalizados de las respuestas del a) sensor 1, c) sensor 2 y e) sensor 3 con sus respectivos acercamientos a sus picos de atenuación.

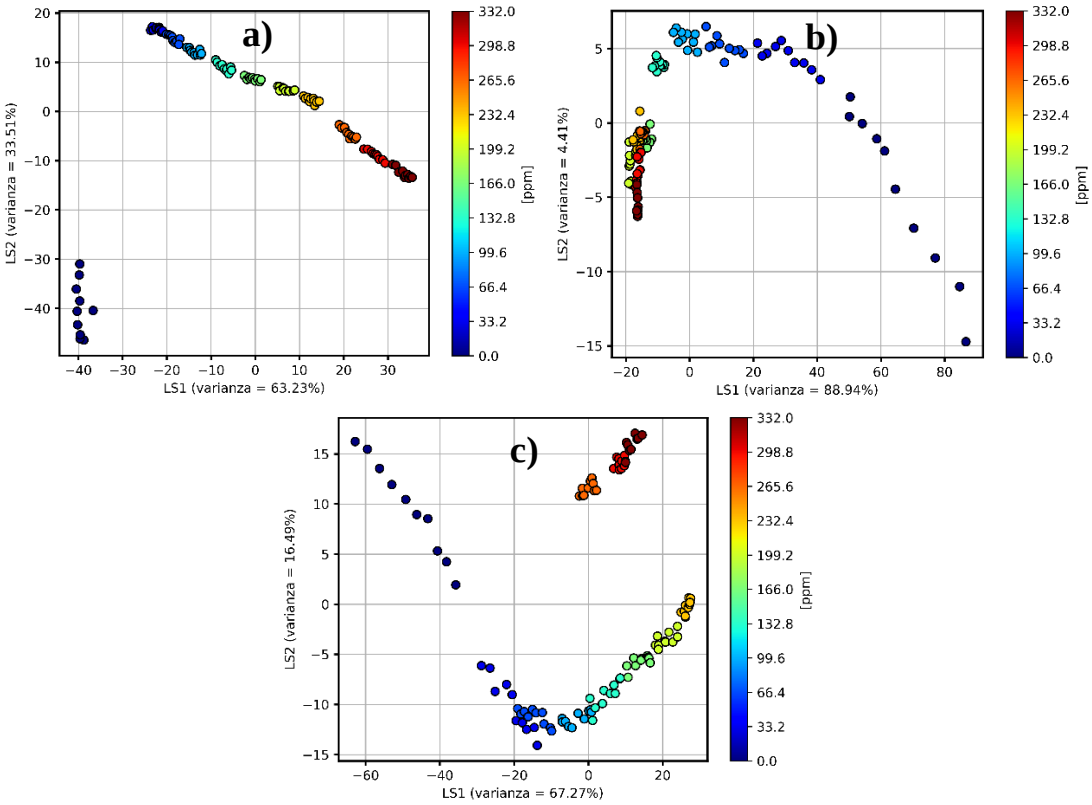

Figura 18. Comportamiento en PLS de el a) sensor 1, b) sensor 2 y c) sensor 3.

En Fig. 18 se muestran los análisis en PLS usando las dos primeras estructuras latentes (LSs) de los espectros mostrados en Fig. 17. Cada uno de los puntos que se visualizan representa a un espectro completo de transmisión utilizando todo el rango espectral. En el sensor 1 se corrobora su buen desempeño al haber una clara tendencia cuando el sensor responde a bajas y altas concentraciones de acetona como se puede observar en Fig. 18a. En el plano de LS se observa que a lo largo de LS1 se ve un comportamiento de valores negativos a valores positivos conforme aumenta la concentración de acetona, mientras que en LS2 tiene una tendencia de pasar de valores positivos a valores negativos, diferenciando perfectamente cada concentración medida. Mientras que, para el sensor 2, como se ve en Fig. 18b, el comportamiento es opuesto, para LS1 se ve una tendencia de la región positiva a la negativa, mientras que en LS2 se forma una curva dónde se produce una aglomeración en mediciones a más altas concentraciones de acetona, sin embargo, son perceptibles las mediciones, mientras que las de bajas concentraciones van disminuyendo gradualmente su dispersión. Pero se aprecia un perceptible cambio entre mediciones en altas concentraciones dentro de

la aglomeración formada. Finalmente, el sensor 3 que se visualiza en la Fig. 18c muestra una clara tendencia en LS1 de la región negativa a la positiva conforme aumenta la concentración, mientras que en LS2 sufre un comportamiento que genera una curva en la representación en los datos, a diferencia de lo que se percibe de Fig. 17e, con PLS demuestra una clara interacción entre el sensor y las concentraciones de acetona, esto no se muestra en los espectros mostrados en la Fig. 17c. Con este análisis usando el método de PLS en los sensores desarrollados, se tiene una mejor interpretación de sus desempeños. Cabe mencionar que también se realizó el análisis de componentes principales (PCA), sólo que no se muestran debido a que sus respuestas son muy similares a PLS. Para determinar la respuesta teórica de los sensores se aplicaron métodos de regresión lineales a los datos de PCA y PLS, y también se implementó el método de RFR. Para los métodos de PCR y PLSR es necesario establecer el número de componentes a utilizar, por tal motivo se realizó CV para evitar el sobreajuste. A continuación, se detalla el método de CV que se empleó.

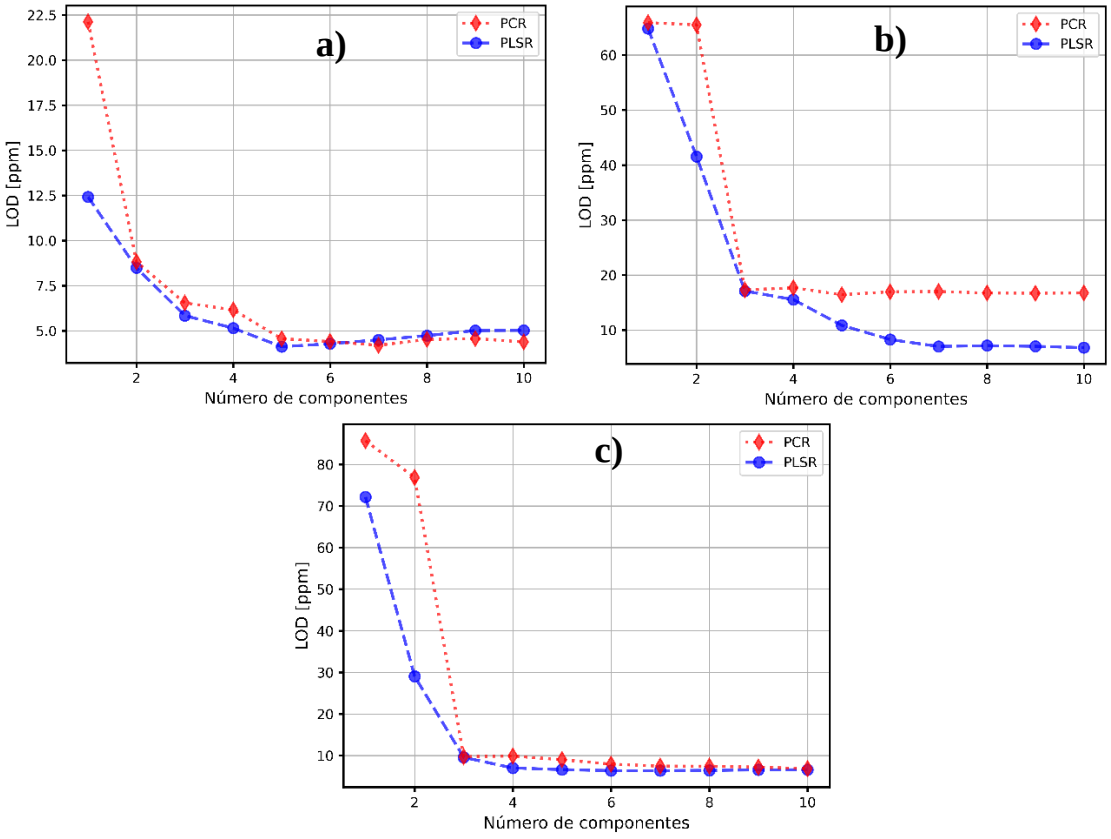

Figura 19. CV de a) sensor 1, b) sensor 2 y c) sensor 3.

Para la evaluación de la regresión, primero se realizó CV para determinar el número de componentes principales (PC) y estructuras latentes (LS) adecuadas para optimizar cada modelo. El valor decidido para evaluar el rendimiento de la CV es el RMSE, que es equivalente al LOD teórico. El cual se obtiene con la siguiente ecuación:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{5.1}$$

Dónde n es el número de observaciones realizadas, en este caso el número de mediciones, $\hat{y}_i$ es la predicción de la i-ésima medición, y y_i es el valor original de la i-ésima medición.

Los datos de los sensores se dividieron en 80% de entrenamiento y 20% de validación. Se utilizó el método de CV de k-particiones, siendo k=5, para los datos de entrenamiento. Como se estableció que el criterio de retención de componentes a utilizar sería por el mínimo de componentes que expliquen la mayor parte de los datos, se buscarían en las gráficas los puntos dónde decrezca abruptamente el límite de detección y no haya algún cambio relevante posteriormente. En Fig. 19a es notorio el bajo LOD que se obtiene desde la primera componente y que gradualmente baja hasta 5 PCs/LSs, pero se observa que los cambios en el LOD posterior a 3 PCs/LSs no es tanta en comparación a las primeras. Del sensor 2 y 3 que se pueden visualizar en Fig. 19b y Fig. 19c es evidente que a 3 PCs/LSs hay un cambio abrupto en el LOD, significando que son el mínimo de componentes requeridas para sus datos. Por lo que se estableció que el número de PCs y LSs para optimizar PCR y PLSR es 3.

Para RFR hay más parámetros a los cuales prestar atención para la optimización del modelo. Aquellos parámetros a los que se les prestó atención por su mayor influencia en la predicción de los resultados fueron; el número de DTs que se construyen durante el análisis, la máxima profundidad que pueden tener éstos, el máximo número de características a considerar de los datos para la construcción de cada DT, y el número mínimo de muestras que requiere un nodo para desglosarse. Para la búsqueda de los valores de estos parámetros, se utilizaron los datos de entrenamiento anteriormente separados de cada sensor y valores alrededor de los estándares como punto de partida. Los valores óptimos encontrados son aquellos mostrados en Tabla 1 en comparación con los valores estándar. Junto a éstos se

decidió no utilizar “bootstraping” durante la búsqueda, ya que esta técnica genera nuevas muestras tomando aleatoriamente valores de las mediciones existentes, lo cual no es requerido para la forma que están estructurados los datos obtenidos y esto perjudicaría. Con la obtención de los valores para la optimización de los parámetros elegidos, se utilizaron éstos para la construcción de RFR optimizado para los datos.

Tabla 1. Mejores valores para los parámetros seleccionados.

Hiperparámetros	**Sensor 1**	**Sensor 2**	**Sensor 3**	**Estándard**
Número de estimadores	80	160	40	100
Mínima muestra para partición	2	5	5	2
Máximas características	log2	log2	sqrt	0.1
Máxima profundidad	25	10	35	None

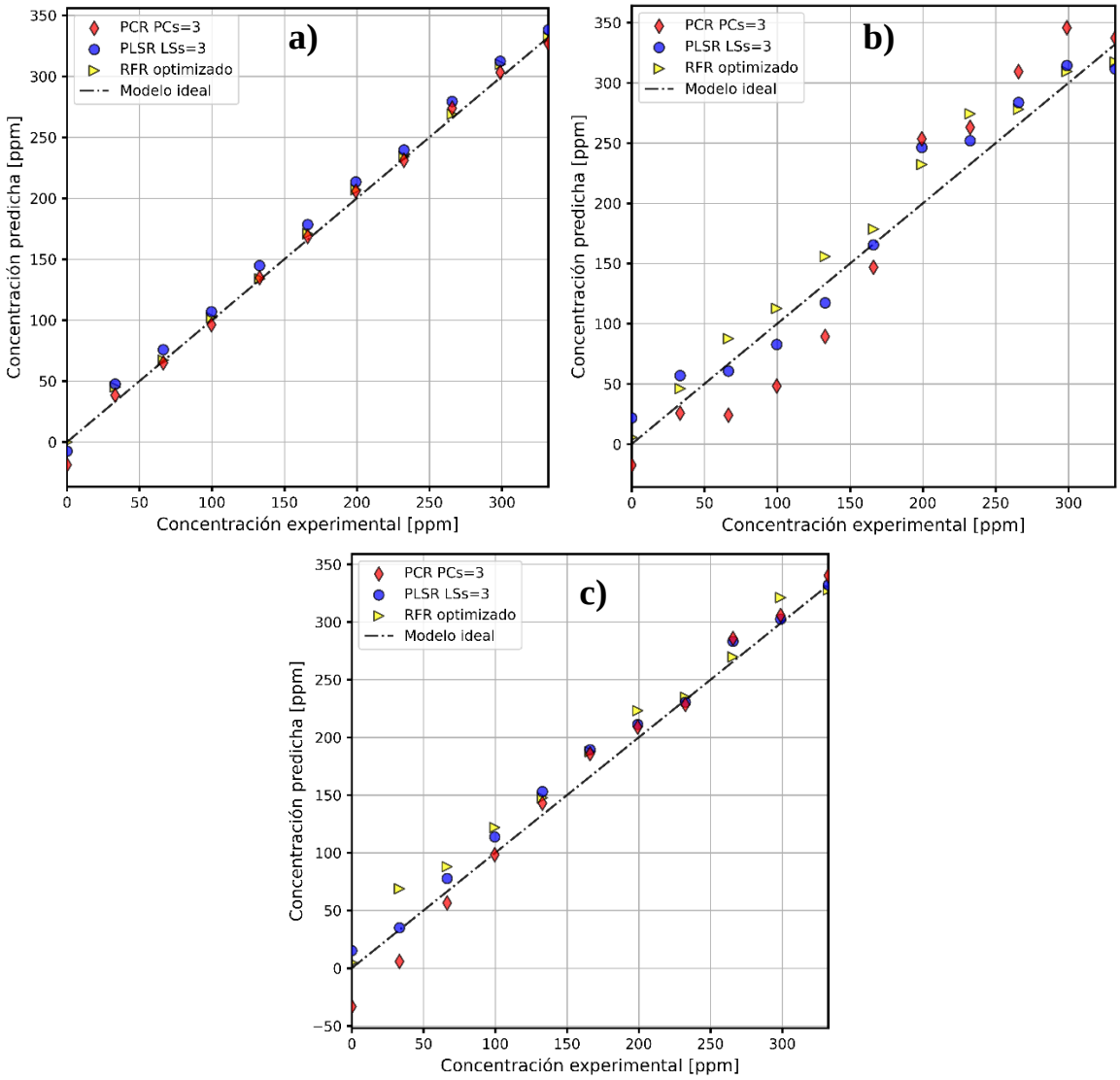

Figura 20. Valores predichos contra valores experimentales a minuto 10 de a) sensor 1, b) sensor 2 y c) sensor 3.

Al construir los modelos con los parámetros establecidos se evaluaron las predicciones con las mediciones al minuto 10 de cada una de las concentraciones. La diferencia entre el valor experimental y el modelo ideal de cada sensor se muestra en Fig. 20. Dónde el sensor 1, que se visualiza en Fig. 20a, muestra que sus predicciones son muy cercanas al modelo ideal en los 3 métodos de análisis, siendo PCR y RFR aquellos más acorde a este. El sensor 2, como se observa en Fig. 20b, tiene una separación evidente con saltos alrededor del modelo ideal. Mientras que en Fig. 20c, el sensor 3 también demuestra ir a lo largo del modelo ideal, teniendo mayores separaciones a las vistas por el sensor 1 con los tres métodos, pero viendo que PLSR se acopla más al modelo ideal. El valor de nuestro interés para determinar el

desempeño de los sensores es la desviación estándar que presentan las predicciones, lo cual equivale al LOD teórico que alcanza cada sensor.

Tabla 2. LODs teóricos y R^2 de los sensores comparando el número de PCs, LSs y los parámetros de RFR optimizados.

Sensor	Λ [µm]	LOD PCR [ppm]		R^2	LOD PLSR [ppm]		R^2	LOD RFR [ppm]		R^2
		1PC	3PC		1LS	3LS		Estandard	Optimizado	
1	475	23.54	7.13	0.995	16.22	11.18	0.988	5.36	4.65	0.998
2	515	63.28	37.07	0.875	66.49	21.75	0.957	32.33	26.23	0.937
3	525	88.49	16.74	0.974	75.17	13.47	0.983	18.38	17.04	0.973

En Tabla 2 se muestran los respectivos LODs teóricos de cada método sin optimizar y optimizado para cada sensor. Comprobando las nociones vistas desde la visualización de su espectro, el sensor 1 tiene el menor LOD teórico de entre los sensores, obteniendo el más bajo con RFR optimizado siendo 4.65 ppm. El siguiente mejor fue el sensor 3, que obtuvo con PLSR 13.47 ppm. Los LODs teóricos más altos con los métodos optimizados se obtuvieron con el sensor 2, el menor siendo de 21.75 ppm utilizando PLSR. A pesar de que PCR con 3 PCs obtiene un LOD teórico menor que PLSR para el sensor 1 y que RFR para el sensor 3, este método no destaca en otro aspecto para el análisis de este tipo de datos. Por otra parte, PLSR con 3 LSs obtiene el menor LOD teórico de entre los tres métodos para los sensores 2 y 3, demostrando su efectividad. RFR resalta de entre los otros métodos, analizando los valores obtenidos del modelo sin optimizar, todos aquellos con RFR son significativamente menores a aquellos de PCR y PLSR con una PC/LS, sin necesidad de optimización se ve un desempeño excepcional con RFR. A esto agregando que se obtuvo el LOD teórico más bajo de entre todos los sensores.

Tabla 3. Porcentaje de mejora de la caracterización con cada método.

Sensor	PCR	PLSR	RFR
1	69.71%	31.07%	13.24%
2	41.41%	67.28%	18.86%
3	81.08%	82.07%	7.29%

En Tabla 3 se ven los porcentajes de mejora que se ve en los LODs teóricos de cada método. Estos valores se obtuvieron restándole a 1 la división del valor del método optimizado contra el valor del método sin optimizar, y el resultado multiplicándolo por cien. Con los porcentajes calculados se ve que en todos los métodos hubo una mejora en el desempeño de los sensores. Aunque los porcentajes de RFR estuvieron por debajo del 20%, como se pudo ver en Tabla 2. LODs teóricos y R^2 de los sensores comparando el número de PCs, LSs y los parámetros de RFR optimizados., sin optimizar ya se tenían valores bajos para todos los sensores. Esto muestra la importancia de realizar la búsqueda de valores para los parámetros deseados. De PCR se ve una mejora mínima del 41% y para PLSR del 31% comprobando la efectividad de estas herramientas para el análisis de espectros sin alterar sus propiedades físicas.

Capítulo 7: Conclusiones

En el presente trabajo de tesis se presentaron tres sensores para la detección de acetona conformados por fibra óptica SMF-28, una LPG con distintos periodos para cada sensor, y PDMS como película sensora. Además, se utilizan tres herramientas de análisis supervisado para analizar las respuestas de los sensores presentados con el propósito de obtener límites de detección teóricos bajos.

Se observa en el espectro de transmisión obtenido de las mediciones realizadas que los sensores son aptos para detectar pequeños cambios de concentración de acetona en el ambiente, lo cual confirma el análisis exploratorio realizado con PLS. Mostrando que sensores fabricados con materiales de bajo costo y métodos sencillos de llevar a cabo presentan un buen desempeño.

Los resultados obtenidos de los modelos de análisis supervisados aplicados muestran las diferencias al darle distintos enfoques a un mismo grupo de datos. Los tres modelos determinan que el sensor con mejor desempeño es el sensor 1, siguiéndole el sensor 3, y finalmente el sensor 2. Lo que se resalta de estos resultados es la importancia de la optimización de los modelos, en cada caso se obtiene una mejora positiva para el objetivo de este trabajo, indicando que hay información oculta en los datos que puede aprovecharse a través de estos modelos. PCR y PLSR muestran porcentajes de mejora mayores al 30%, mientras que aquellos de RFR no superan el 20%. Observando detenidamente los resultados de cada modelo, al comparar los valores no optimizados de cada uno, la separación entre PCR y PLSR con RFR es significativa, siendo todos los LODs teóricos obtenidos con RFR sin optimizar menores. Concluyendo que los modelos de machine learning aprovechan mejor las respuestas de los sensores dando bajols LODs. Una propuesta para trabajos futuros sería el optimizar de mejor manera el modelo RFR al utilizar la validación cruzada anidada, que es un tipo de CV de mayor complejidad computacional, la cual realiza un diagnóstico más profundo del modelo a evaluar.

Los sensores cumplen con el objetivo propuesto de estar dentro del rango deseado de LOD teórico. Siendo las LPG con periodo de 475 y 525 µm aquellas que dan mejores

resultados para la detección de acetona, siendo el sensor con periodo de 475 µm aquel con los mejores LODs teóricos, 7.13 ppm con PCR, 11.18 ppm con PLSR y 4.65 ppm con RFR. En el caso del sensor 2 solamente son PLSR se obtiene un LOD teórico que entra dentro del rango antes mencionado. Esto muestra la efectividad para caracterizar sensores utilizando herramientas de machine learning, a lo cual se propondría el profundizar en estas herramientas para darle un trato adecuado a los datos y a los modelos utilizados para el procesamiento de las respuestas de los sensores.

Finalmente, se propone el refinar el método de fabricación de los sensores. Al concluir que con los materiales y métodos elegidos para este trabajo se obtuvo un buen desempeño y análisis por parte de los sensores fabricados. Este desempeño se vería beneficiado al utilizar métodos y materiales que mejoren la sensitividad del dispositivo, sin perder de vista el objetivo de mantenerlo accesible para futuras aplicaciones.

Bibliografía

[1] Das, S., Pal, S., Mitra, M. (2016). Significance of exhaled breath test in clinical diagnosis: a special focus on the detection of diabetes mellitus, J. Med. Biol. Eng. 36 (5) 605–624, https://doi.org/10.1007/s40846-016-0164-6

[2] Zhang, Z., Cang, H., Huang, W., Li, H., Li, H. (2025). Photoionization ion mobility analyzer for on-site measurement of exhaled acetone by coupling miniature thermoelectric cooling dehydration. Sensors and Actuators: B Chemical, 423(136743), 136743. https://doi.org/10.1016/j.snb.2024.136743

[3] Sun, Z., Sun, S., Hao, X., Wang, Y., Gong, C., Cheng, P. (2024) Gas sensor for efficient acetone detection and application basen on Au_modified ZnO Porous Nanofoam. Sensors, 24(24), 8100. https://doi.org/10.3390/s24248100

[4] Esposito, F. (2021). (INVITED) Chemical sensors based on long period gratings: A review. Results in Optics, 5. https://doi.org/10.1016/j.rio.2021.100196

[5] Cai, J., Liu, Y. and Shu, X. (2023). Long-Period Fiber Grating Sensors for Chemical and Biomedical Applications, Sensors, 23, 542. https://doi.org/10.3390/s23010542

[6] Meneses-Mijares J., Castillo-Mixcóatl, J., Muñoz-Aguirre, S. and Beltrán-Pérez, G. (2024). Application of principal component regression in Mach-Zehnder interferometer optical fiber sensors in reflection mode for acetone detection as biomarker of diabetes mellitus. Optics & Laser Technology, 177(111196),111196. https://doi.org/10.1016/j.optlastec.2024.111196

[7] Hernández-Guerrero, L.D., Castillo-Mixcóatl, J., Muñoz-Aguirre, S., Rodríguez-Torres, M., Ramírez-Sánchez, E., and Beltrán-Pérez, G. (2025). Projection to latent structures regression and its applications to Mach-Zehnder interferometer optical fiber sensors for acetone detection. Optics and Lasers in Engineering, 184(108689). 108689. https://doi.org/10.1016/j.optlaseng.2024.108689

[8] Ramírez-Sánchez, E., Muñoz-Aguirre, S., Castillo-Mixcóatl, J., González-León, K., Rodríguez-Torres, M., Hernández-Guerrero, L.D., Beltrán-Pérez, G. (2025). A comparative study between PCR and PLSR in tapered optical fiber sensor for acetone detection. Optics & Laser Technology, 181(111838), 111838. https://doi.org/10.1016/j.optlastec.2024.111838

[9] Rodríguez-Garciapiña, J.L., Beltrán-Pérez. G., Castillo-Mixcóatl, J., Muñoz Aguirre, S. (2021). Application of the principal components analysis technique to optical fiber sensors for acetone detection. Optics & Laser Technology, 143

[10] Hromdaka, J., Korposh, S., Partridge, M., James, S.W., Davis, F., Crump, D., Tatam, R.P. (2017) Volatile organic compounds sensing using optical long period grating with mesoporous nano-scale coating. Sensors, 17(2), 205. https://doi.org/10.3390/s17020205

[11] Hromdaka, J., Tokay, B., Correia, R., Morgan, S.P., Korposh, S. (2018). Highly sensitive volatile organic compounds vapour measurements using a long period grating optical fibre sensor coated with metal organic framework ZIF-8. Sensors and Actuators B: Chemical, 260, 685-692. https://doi.org/10.1016/j.snb.2018.01.015

[12] Chakraborty, D. Elzarka, H. (2018) Advanced machine learning techniques for building performance simulation: a comparative analysis. Journal of Building Performance Simulation, 14(4), 1-15. https://doi.org/10.1080/19401493.2018.1498538

[13] Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2010). Multivariate Data Analysis. 7th Edition, Pearson, New York.

[14] Jolliffe, I. T, Cadima, J. (2016). Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A. 374: 20150202. http://doi.org/10.1098/rsta.2015.0202

[15] Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression): PLS REGRESSION. Wiley Interdisciplinary Reviews, Computational Statistics, 2(1), 97-106. https://doi.org/10.1002/wics.51

[16] Paniagua-Median, J.J., Vargas-Rodriguez, E., Guzman-Chavez, A.D., Morales-Castro J.C., Correa-Jurado, R.J. (2025). Random forest regression for improving the measurement range of temperature interferometric sensor. IEEE Photonics Technology Letters, 37(2), 101-104. https://doi.org/10.1109/LPT.2024.3517424

[17] Das, S., Pal, S., Mitra, M. (2016). Significance of exhaled breath test in clinical diagnosis: a special focus on the detection of diabetes mellitus, J. Med. Biol. Eng. 36 (5) 605–624, https://doi.org/10.1007/s40846-016-0164-6

[18] National Institute of Diabetes and Digestive and Kidney Diseases. (2025, January) Symptoms & Causes of Diabetes – NIDDK. U.S. https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

[19] DiMeglio LA, Evans-Molina C, Oram RA. (2018). Type 1 diabetes. Lancet. Jun 16;391(10138):2449-2462. doi: 10.1016/S0140-6736(18)31320-5. PMID: 29916386; PMCID: PMC6661119

[20] Goyal R, Singhal M, Jialal I. Type 2 Diabetes. [Updated 2023 Jun 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513253/

[21] Centro para el control y la prevención de enfermedades. (2024,26,04) Pruebas de diabetes y prediabetes: A1c. https://www.cdc.gov/diabetes/es/diabetes-testing/pruebas-de-diabetes-y-prediabetes-a1c.html

[22] Keiser, G. (1991). Optical Fiber Communications. McGraw-Hill International Editions.

[23] Mitschke, F. (2016) Fiber Optics. Springer.

[24] Hetch, E. (2015). Óptica (5ta ed.). Pearson Education.

[25] Yu, L., John, W., Lin, Z., Ian, B. (1999) Phase shifted and cascaded long-period fiber gratings, Opt. Commun. 164 (1–3) 27–31, http://dx.doi.org/10.1016/s0030-4018(99)00191-1.

[26] Gómez-Correa, J.E., Balderas-Mata, S.E., Coello, V., Puente, N.P., Rogel-Salazar, J., Chávez-Cerda, S. (2017) On the physics of propagating Bessel modes in cylindrical waveguides. American Association of Physics Teachers. 85(5), 341-345. http://dx.doi.org/10.1119/1.4976698

[27] Subramanian, R., Zhu, C., Li, H. (2018). Torsion, strain and temperature sensor base don helical long-period fiber gratings. IEEE Photonics Technol, Lett. 30(4), 327-330. https://doi.org/10.1109/LPT.2017.2787157

[28] Del Villar, I., Fuentes, O., Chiavaioli, F., Corres, J.M., Matias, I.R. (2018) Optimized strain long-period fiber grating (LPFG) sensors operating at the dispersión turning point. J. Light. Technol. 36(11), 2240-2247, https://doi.org/10.1109/JLT.2018.2790434

[29] Barnes, J., Dreher, M., Plett, K., Brown, R.S., Crudden, C.M., Loock, H.-P. (2008). Chemical sensor based on a long-period fiber grating modified by functionalized polydimethylsiloxane coating. The Analyst, 133(11), 1541-1549. https://doi.org/10.1039/b806129g

[30] Wang, T., Yasukochi, W., Korposh, S., James, S.W., Tatam, R.P., Lee, S.W. (2016). A long period grating optical fiber sensor with nana-assembled porphyrin layers for detecting ammonia gas. Sensors and Actuators B: Chemical, 228, 573-580. https://doi.org/10.1016/j.snb.2016.01.058

[31] Trono, C. (2024) Long period fiber grating-based biosensing: Recent trends and future perspectives. TrAC Trends in Analytical Chemistry, 179(117875). https://doi.org/10.1016/j.trac.2024.117875

[32] Korposh, S. Lee, S.W., James, S. (2017) Long period grating based fibre optic chemical sensors. In: Matias, I., Ikezawa, S., Corres, J. (eds) Fibre Optic Sensors. Smart Sensors, Measurement and Instrumentation , vol 21. Springer, Cham. https://doi.org/10.1007/978-3-319-42625-9_12

[33] Marques, L., Hernandez, F.U., James, S.W., Morgan, S.P., Clark, M., Tatam, R.P., Korposh, S. (2016). Higly sensitive optical fibre long period grating biosensor anchored with silica core gold shell nanoparticles. Biosensors and Bioelectronics, 75 (2016), 222-231. https://doi.org/10.1016/j.bios.2015.08.046

[34] Barnes, J.A., Brown, R.S., Cheung, A.H., Dreher, M.A., Mackey, G., Loock, H.-P. (2010) Chemical sensing using a polymer coated long-period fiber grating interrogated by ring-down spectroscopy. Sensors and Actuators B: Chemical, 148 (2010), 221-226. https://doi.org/10.1016/j.snb.2010.04.007

[35] United States Environmental Protection Agency. What are organic volatile compounds? [en línea] EPA. 24 de febrero de 2025. https://www.epa.gov/indoor-air-quality-iaq/what-are-volatile-organic-compounds-vocs

[36] Adams, A., Vamplew, P. (1998) Encoding and decoding cylcic data. The South Pacific Journal of Natural Science. 16.

[37] P. Bescond, Cyclical features encoding, it’s about time! (2020), Towards Data Science

[38] Cord, M., Cunningham, P. (2008) Machine learning techniques for multimedia. Springer.

[39] Cadima, J., Cerdeira, J. O., Minhoto, M. (2004). Computational aspects of algorithms for variable selection in the context of principal components. Computational Statistics & Data Analysis. 47(2), 225-236. https://doi.org/10.1016/j.csda.2003.11.001

[40] Klema, V., Laub, A. (2019). The singular value decomposition: Its computation and some applications. IEEE Transactions on Automatic Control, 25 (2), 164-176. https://doi.org/10.1109/TAC.1980.1102314

[41] Cangelosi, R., Goriely, A. (2007). Component retention in principal component análisis with application to cDNA microarray data. Biology Direct, 2(2). https://doi.org/10.1186/1745-6150-2-2

[42] Dunn, K. (2024) Process improvement using data. Learnche.

[43] Sanchez, G., Marzban, E. (2024). All models are wrong: Concepts of statistical learning.

[44] Zhou, Z-H. (2012). Ensemble methods: foundations and algorithms. CRC Press.

[45] Paniagua-Median, J.J., Vargas-Rodriguez, E., Guzman-Chavez, A.D., Morales-Castro J.C., Correa-Jurado, R.J. (2025). Random forest regression for improving the measurement range of temperature interferometric sensor. IEEE Photonics Technology Letters, 37(2), 101-104. https://doi.org/10.1109/LPT.2024.3517424

[46] Refaeilzadeh, P., Tang, L., and Liu, H. (2009) Cross-Validation. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

[47] Berrar, D. (2018) Cross-validation. Reference Module in Life Science. https://doi.org/10.1016/B978-0-12-809633-8.20349-X