Proyecto Fin de Máster
Máster en Ingeniería Electrónica, Robótica y Automática

Determinación del estrés hídrico de una planta mediante sensores ZIM y técnicas de aprendizaje automático

Autor: Jaime Palomo Iranzo

Tutores: Teodoro Álamo Cantarero, David Muñoz de la Peña Sequedo

Dpto. Ingeniería de Sistemas y Automática
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2022

Proyecto Fin de Máster
Máster en Ingeniería Electrónica, Robótica y Automática

Determinación del estrés hídrico de una planta mediante sensores ZIM y técnicas de aprendizaje automático

Autor:
Jaime Palomo Iranzo

Tutor:
Teodoro Álamo Cantarero, David Muñoz de la Peña Sequedo
Catedráticos de Universidad

Dpto. Ingeniería de Sistemas y Automática
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2022

Proyecto Fin de Máster: Determinación del estrés hídrico de una planta mediante sensores ZIM y técnicas de aprendizaje automático

Autor: Jaime Palomo Iranzo
Tutor: Teodoro Álamo Cantarero, David Muñoz de la Peña Sequedo

El tribunal nombrado para juzgar el trabajo arriba indicado, compuesto por los siguientes profesores:

Presidente:

Vocal/es:

Secretario:

acuerdan otorgarle la calificación de:

El Secretario del Tribunal

Fecha:

Resumen

El estrés hídrico refleja el nivel de hidratación de una planta. Medirlo adecuadamente permite ajustar el agua de riego para maximizar el rendimiento de la planta, obteniendo la máxima producción con el menor consumo de agua, así como monitorizar su estado de salud.

En este sentido, se han propuesto numerosos métodos para realizar la medida. Este proyecto se centra en uno de ellos, enmarcado dentro de la investigación realizada por parte del IRNAS-CSIC, en el que se emplean sensores ZIM para medir el nivel de estrés hídrico a través de la turgencia de las hojas, centrándose por el momento en el olivo de la variedad arbequina, con el fin de aplicar técnicas de aprendizaje automático donde actualmente se necesita el análisis manual de un experto.

Índice

1 Introducción

El estrés hídrico representa una medida del nivel de hidratación de una planta en cada momento. Su correcta lectura e interpretación resulta de vital importancia especialmente en la agricultura, ya que un exceso o una carencia de agua puede provocar enfermedades e incluso la muerte del organismo. Además, el control del nivel óptimo de irrigación permite maximizar la producción a la vez que se minimiza el consumo de agua.

En este sentido, se han propuesto numerosas soluciones para la medida del nivel de estrés, incluyendo entre otros la observación y medida de la pigmentación de la hoja o su masa, así como medidas más especializadas como las obtenidas a través del uso de la cámara de Scholander [3] o sensores de actividad hídrica [5].

Recientemente se han realizado varios estudios desde el Instituto de Recursos Naturales y Agrobiología de Sevilla (IRNAS-CSIC) con el fin de emplear sensores de grosor del tronco (dendrómetros) y de turgencia de la hoja (sensores ZIM) [2] para la planificación del riego [6]. De las conclusiones de estos estudios se extrae que los sensores ZIM proporcionan suficiente información para determinar el nivel de estrés hídrico de la planta y con unos costes de instalación y mantenimiento menores que los de otros sensores no invasivos, además de permitir la recogida automatizada de datos. Estas características hacen que posean un gran potencial para su uso en cultivos comerciales, pero la interpretación de los datos recogidos resulta más compleja que empleando otras técnicas y requiere, por el momento, de un análisis manual por parte de un experto. Con el fin de divulgar los resultados de sus estudios, el IRNAS-CSIC ha publicado además un manual de riego [4] en el que se recogen estos avances.

1.1 Objetivos del proyecto

El objetivo principal de este trabajo es continuar la investigación realizada por el IRNAS-CSIC, empleando técnicas de aprendizaje automático para intentar determinar el nivel de estrés correspondiente a las medidas de los sensores ZIM, facilitando una interpretación automatizada que pueda ser aplicada a sistemas eficientes de riego automático o facilite la tarea de lectura de los sensores.

Como segundo objetivo se busca determinar si es necesario establecer un método de calibración del sensor o si el análisis es robusto ante desviaciones de la medida. Idealmente, el sistema automático debería ser capaz de identificar el nivel de estrés en años distintos, independientemente de si se cambia el sensor o si es instalado en un árbol distinto, siempre que sea de la misma especie.

1.2 Datos de partida del proyecto

Con el fin de realizar el proyecto, el IRNAS-CSIC ha proporcionado los datos recogidos durante sus estudios anteriores, concretamente los correspondientes a los artículos "Plant-based sensing to monitor water stress: Applicability to commercial orchards" [2], de J.E. Fernández, y "Scheduling regulated deficit irrigation in a hedgerow olive orchard from leaf turgor pressure related measurements" [6], de C.M. Padilla-Díaz, C.M.Rodriguez-Dominguez, V.Hernandez-Santana, A. Perez-Martina y J.E. Fernández, además de nuevos datos obtenidos a lo largo del año 2019 en la misma finca.

En las siguientes subsecciones se analizan los datos disponibles y su origen.

1.2.1 Origen de los datos

Los datos empleados en este trabajo fueron recogidos en la finca que emplearon los estudios del IRNAS-CSIC antes mencionados. La finca se encuentra a 25km al este de Sevilla, en las coordenadas GPS 37º 15Ñ, -5º48'W. Los árboles se encuentran plantados sobre caballones de 0.4m de altura, plantados en una rejilla de 4m x 1.5m (1667 árboles por hectárea), con las filas de árboles orientadas en dirección N-NE a S-SO. El suelo de la finca es arenoso, con baja capacidad de retención de agua y una profundidad útil máxima de 0.6m.

El clima en la zona es mediterráneo con inviernos suaves y húmedos y veranos calurosos y secos, con un período de lluvias entre septiembre y mayo, manteniéndose seco el resto del año. En el período de 2002 a 2014, la precipitación media fue de 540mm y la evapotranspiración media, 1528mm. En los meses más cálidos, julio y agosto, las temperaturas máximas del aire superaron los 40ºC y ocasionalmente superaron los 45ºC. En los más fríos, diciembre y enero, las temperaturas mínimas descendieron, de forma muy ocasional, por debajo de 0ºC y anecdóticamente fueron inferiores a -5ºC.

Los árboles son olivos de la variedad arbequina, monocónicos, con un único tronco y las ramas principales a 0.6m o 0.7m del suelo, sembrados en el año 2007.

1.2.2 Sensores ZIM

La turgencia es el fenómeno por el cual los fluidos dentro de una célula ejercen presión sobre sus paredes. También es aplicable a un conjunto de células. En el caso de este trabajo, la turgencia de una hoja es la cantidad de presión que el agua almacenada por las células de la hoja ejerce sobre el resto de células y el exterior de la hoja. Una hoja con una turgencia elevada tendrá un alto contenido en agua, mientras que una turgencia baja se relaciona con hojas en un estado de sequedad; al mismo tiempo, una hoja sana realiza un ciclo diario por el cual su turgencia varía de forma continua en función de diversos factores, tanto meteorológicos como relacionados con el estado del árbol. Para medir la turgencia de las hojas se pueden emplear diversos métodos, como la cámara Scholander o los sensores ZIM. Este trabajo se centra en estos últimos.

En la figura 1.1 se han representado las partes de un sensor ZIM. El sensor consta de dos imanes permanentes toroidales enfrentados entre sí que se atraen mutuamente. El imán 1 se sitúa en la cara superior de la hoja (haz) y es atravesado por la pieza 3, consistente en una placa circular soldada a una varilla roscada, que en este trabajo se ha denominado tornillo de calibración. El imán 2 se coloca en la cara inferior (envés) y dispone de un sensor de presión cubierto por una membrana de silicona (pieza 4).

El sensor de presión mide la presión ejercida por los imanes sobre la hoja, que deberá ser igual a la presión de turgencia (LTP) en equilibrio. La presión ejercida por los imanes dependerá del área de la superficie de contacto con la hoja y de la fuerza de atracción magnética (F_m). El tornillo de calibración permite añadir una cierta separación entre los imanes, ajustando el valor de la fuerza magnética máxima.

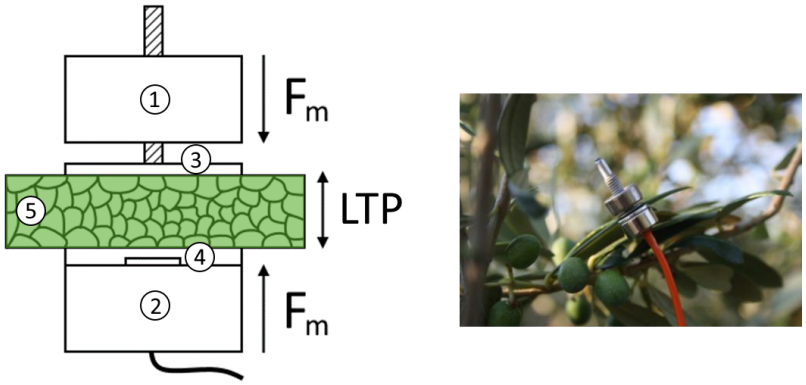

Figura 1.1 Izquierda: Componentes de un sensor ZIM. (1) y (2): Imanes permanentes toroidales. (3): Tornillo de calibración. (4): Sensor de presión y cubierta de silicona. (5) Hoja. Derecha: fotografía de un sensor ZIM instalado..

Debe tenerse en cuenta que la medida será mayor cuanto más cerca entre sí se encuentren los imanes y mayor sea la fuerza de atracción entre ellos, es decir, cuanto más se deforme la hoja antes de alcanzar el equilibrio y por tanto menor sea su turgencia.

También debe tenerse en cuenta que pueden aparecer ruidos en las medidas. Además, en las hojas más secas pueden formarse cámaras de aire que presentarán un comportamiento parecido al de las células turgentes, pudiendo dificultar la lectura de los datos fuera de contexto.

En este trabajo, la variable medida por el sensor se va a llamar LTP, por las siglas en inglés de presión de turgencia de la hoja.

1.2.3 Estados de estrés hídrico

Para facilitar la comprensión de la medida se han definido tres niveles o estados de estrés hídricos en los que puede encontrarse la planta:

- En el primer estado, la planta dispone de recursos hídricos en abundancia, pudiendo indicar incluso que el aporte de agua es mayor del necesario. Se considera que tiene un estrés hídrico leve.
- En el segundo estado, la planta dispone de suficiente agua para su desarrollo en la mayoría de las fases del ciclo anual de la planta. Se considera que tiene un estrés hídrico moderado.
- En el tercer estado, la planta carece de suficiente agua para un desarrollo adecuado, y pueden llegar a aparecer daños permanentes si se mantiene en este estado durante un tiempo prolongado. Se considera que tiene estrés hídrico severo.

Como se mencionó anteriormente, la presión de turgencia varía a lo largo del día. La forma de la curva que se obtiene al medir la presión de turgencia varía en función del estado en que se encuentre la planta, entre otros factores. La figura 1.2 representa la evolución de las curvas de turgencia según el estado en que se encuentra la planta adulta y sana en condiciones climáticas típicas.

Puede observarse que la gráfica se presenta sin valores numéricos. Según los estudios realizados anteriormente por el IRNAS-CSIC [2, 6], el valor numérico tiene una componente continua que depende del tamaño y la edad de la hoja y de la calibración del sensor y una amplitud que depende de las mismas variables, siendo únicamente la forma la que realmente aporta información sobre el estado hídrico y la turgencia, por lo que se han omitido los valores en la elaboración del ejemplo.

Estas formas poseen además un significado biológico. Debe recordarse que la medida del sensor se encuentra invertida respecto a la turgencia de la hoja, por lo que la máxima turgencia se encuentra en los mínimos de la gráfica y viceversa.

En el caso de la planta en el estado 1, la hoja dispone de agua suficiente para mantener su turgencia máxima durante la noche. Cuando amanece, la radiación solar y el aumento de la temperatura provocan un aumento de la evapotranspiración, por lo que la hoja comienza a perder agua, que recupera a través de la planta.

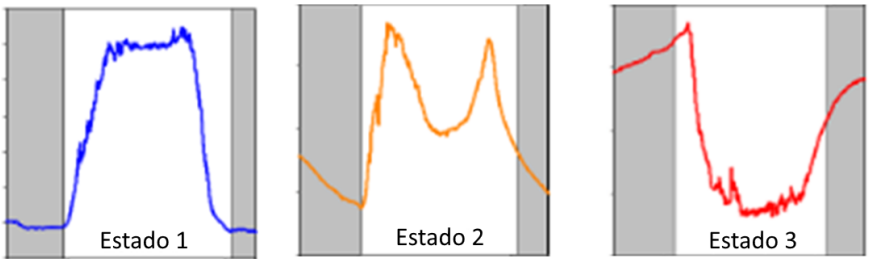

Figura 1.2 Turgencia de la hoja a lo largo de un día según su estado de estrés. La zona sombreada representa los periodos nocturnos..

Cuando la pérdida por evapotranspiración y el aporte de agua desde la planta se equilibran, se alcanza un valor de mínima turgencia, que se mantiene hasta el atardecer, cuando la temperatura y la radiación vuelven a disminuir, de forma que el aporte de agua desde la planta hace que la hoja recupere su turgencia máxima.

Cuando el estrés hídrico aumenta, la evapotranspiración puede superar un cierto límite que provoca que los poros de las hojas se cierren, minimizando su efecto. Al ocurrir esto, la eficiencia de las funciones vitales de las hojas disminuye, pero también lo hace su consumo de agua, recuperando la turgencia. Al volver a superar cierto límite, los poros vuelven a abrirse para recuperar su funcionamiento normal, volviendo a disminuir la turgencia y creando una figura con dos picos principales.

Cuando el estrés alcanza niveles severos, la hoja no es capaz de recuperar su turgencia durante la noche, por lo que los poros se cierran rápidamente al comenzar el día y permanecen cerrados la mayor parte del tiempo, causando una curva invertida respecto a la correspondiente a un estado de estrés leve.

Si se mantiene el nivel de estrés severo, la falta de agua puede ocasionar la aparición de burbujas de aire entre las células de la hoja, las cuales aplicarían presión desde su interior de forma parecida a la correspondiente a una célula turgente. Estas burbujas causan distorsiones en la forma de la curva, pudiendo aparecer formas que no se corresponden con el nivel de estrés real de la planta.

Debe tenerse en cuenta además que los factores ambientales pueden modificar la forma de las curvas. Por ejemplo, un día particularmente seco y cálido puede causar que la planta se vea forzada a cerrar los poros en un estado de hidratación que en otras circunstancias correspondería a un estado de estrés leve, en lugar de moderado.

1.2.4 Datos disponibles

Para la elaboración del trabajo se dispone de los datos recogidos en la finca durante los años 2011, 2014 y 2019. Estos datos incluyen las lecturas de los sensores ZIM y los datos meteorológicos recogidos en la finca durante aproximadamente seis meses en cada uno de esos años, comenzando en abril y terminando en octubre. Para cada año, los expertos del IRNAS-CSIC han seleccionado los 6 sensores más fiables de entre los instalados, con el fin de asegurar que la información proporcionada es correcta.

En cada uno de estos años se probaron diversas estrategias de riego deficitario, disponiendo de sensores ZIM en cada una de las parcelas en las que se dividió la finca, tanto en las que se aplicó riego deficitario como en las que formaban el grupo de control. Las medidas de los sensores ZIM se registraban con una frecuencia de 5 minutos, no sincronizadas entre sensores.

En cuanto a la información meteorológica, se dispone de la temperatura ambiente, la humedad relativa, la radiación global, parcial y neta y el déficit de presión de vapor. Esta información se registra cada 30 minutos, de forma sincronizada. En el caso del año 2011, el registro de radiación neta falló en la mayor parte de las lecturas, por lo que se desaconseja su uso.

También se dispone de la información diaria de la evapotranspiración, las precipitaciones y el riego asignado a cada parcela.

Por último, se dispone también del análisis por parte de expertos del IRNAS-CSIC de las lecturas de los sensores ZIM, que asigna a cada día un estado hídrico.

Con el fin de contar el número de datos que se pueden utilizar, se ha definido una muestra como los datos de lectura del sensor ZIM, meteorológicos y de estado hídrico correspondientes a un mismo día. La muestra será válida si se dispone de todos estos datos, mientras que si faltan datos de lectura del sensor o de estado hídrico, y es imposible su recuperación mediante interpolación en el caso de los datos del sensor, la muestra será descartada.

En el caso de que falten datos meteorológicos, la muestra será válida o descartada en función de los datos requeridos por el clasificador.

Debe tenerse en cuenta que no todos los días disponen de sus correspondientes análisis y medidas. Al tratarse de sensores instalados a la intemperie, no es inusual que ocasionalmente se desprendan de la hoja, perdiéndose la medida hasta que se reinstale el sensor. A causa de esto, si es necesario emplear medidas de varios días consecutivos, el número de muestras útiles será menor.

En total, se dispone de 762 muestras de 2011, ninguna de las cuales dispone de datos de radiación neta suficientes. Observando las muestras disponibles por sensor, se obtiene la siguiente lista:

- Del primer sensor de la primera parcela de control, denominado LTP Control_1_1, se han obtenido 42 muestras válidas consecutivas, 2 no válidas y otras 97 muestras válidas, con un total de 139 muestras.
- Del primer sensor de la segunda parcela de control, denominado LTP Control_2_1, se han obtenido un total de 120 muestras consecutivas.
- Del primer sensor de la primera parcela de riego deficitario al 60 %, denominado LTP 60RDI_1_1, se han obtenido un total de 141 muestras consecutivas.
- Del primer sensor de la segunda parcela de riego deficitario al 60 %, denominado LTP 60RDI_2_1, se han obtenido 64 muestras válidas consecutivas, 8 no válidas y otras 50 muestras válidas, con un total de 114 muestras.
- Del primer sensor de la primera parcela de riego deficitario al 30 %, denominado LTP 30RDI_1_1, se han obtenido un total de 122 muestras consecutivas.
- Del primer sensor de la segunda parcela de riego deficitario al 30 %, denominado LTP 30RDI_2_1, se han obtenido un total de 126 muestras consecutivas.

En cuanto al número de muestras en cada estado hídrico recogidas por cada sensor, se ha realizado la siguiente tabla:

Sensor	Estado 1	Estado 2	Estado 3
LTP Control_1_1	120	19	0
LTP Control_2_1	119	1	0
LTP 60RDI_1_1	65	32	44
LTP 60RDI_2_1	34	19	61
LTP 30RDI_1_1	2	24	96
LTP 30RDI_2_1	7	33	86

Del año 2014, se dispone de un total de 1160 muestras:

- Del primer sensor de la primera parcela de control, denominado LTP Control_1_1, se han obtenido un total de 204 muestras consecutivas.
- Del primer sensor de la segunda parcela de control, denominado LTP Control_2_1, se han obtenido un total de 172 muestras consecutivas.
- Del sensor de la primera parcela de riego deficitario al 45 % basado en las lecturas ZIM, denominado LTP 45RDI-PB_1, se han obtenido un total de 172 muestras consecutivas.
- Del sensor de la segunda parcela de riego deficitario al 45 % basado en las lecturas ZIM, denominado LTP 45RDI-PB_2, se han obtenido un total de 204 muestras consecutivas.
- Del sensor de la primera parcela de riego deficitario al 45 % basado en el coeficiente de cultivo, denominado LTP 45RDI-CC_1, se han obtenido un total de 204 muestras consecutivas.
- Del sensor de la segunda parcela de riego deficitario al 45 % basado en el coeficiente de cultivo, denominado LTP 45RDI-CC_2, se han obtenido un total de 204 muestras consecutivas.

En cuanto al número de muestras en cada estado hídrico recogidas por cada sensor, se ha realizado la siguiente tabla:

Sensor	Estado 1	Estado 2	Estado 3
LTP Control_1_1	171	29	4
LTP Control_2_1	163	9	0
LTP 45RDI-PB_1	76	51	45
LTP 45RDI-PB_2	134	50	20
LTP 45RDI-CC_1	127	42	35
LTP 45RDI-CC_2	94	72	38

Por último, del año 2019, se dispone de 1058 muestras:

- Del primer sensor de la primera parcela de control, denominado LTP Control_1_1, se han obtenido un total de 171 muestras consecutivas.
- Del primer sensor de la segunda parcela de control, denominado LTP Control_2_1, se han obtenido un total de 183 muestras consecutivas.
- Del segundo sensor de la primera parcela de riego deficitario al 45 %, denominado LTP RDI45_1_2, se han obtenido 21 muestras válidas consecutivas, 6 no válidas, otras 54 válidas consecutivas, 6 no válidas, 10 válidas, 1 no válida y 85 válidas, con un total de 170 muestras válidas.
- Del primer sensor de la segunda parcela de riego deficitario al 45 %, denominado LTP RDI45_2_1, se han obtenido un total de 182 muestras consecutivas.
- Del segundo sensor de la primera parcela de riego deficitario al 30 %, denominado LTP RDI30_2_1, se han obtenido un total de 176 muestras consecutivas.
- Del tercer sensor de la segunda parcela de riego deficitario al 30 %, denominado LTP RDI30_3_2, se han obtenido 9 muestras válidas consecutivas, 4 no válidas, otras 21 válidas consecutivas, 3 no válidas y 146 válidas, con un total de 176 muestras válidas.

En cuanto al número de muestras en cada estado hídrico recogidas por cada sensor, se ha realizado la siguiente tabla:

Sensor	Estado 1	Estado 2	Estado 3
LTP Control_1_1	168	3	0
LTP Control_2_1	179	4	0
LTP RDI45_1_2	126	36	8
LTP RDI45_2_1	141	15	26
LTP RDI30_2_1	86	32	58
LTP RDI30_3_2	93	36	47

Puede observarse que la mayoría de datos disponibles corresponden a plantas en un estado de estrés leve (1), con muy pocas en estado de estrés moderado (2). Los clasificadores a diseñar deberán tener en cuenta este desequilibrio o utilizar un conjunto de muestras inferior que contenga un número balanceado de muestras de cada tipo.

1.3 Resumen del trabajo realizado

Para el tratamiento de los datos se ha optado por emplear el lenguaje de programación Python, por su extenso uso en tareas de aprendizaje automático, y la librería scikit-learn[1], por ser una librería de código abierto y uso libre de alto nivel pero con un alto rendimiento [7], y por disponer de una extensa documentación sobre su uso y aplicaciones [8].

La memoria se ha organizado siguiendo una estructura en orden cronológico, comenzando por los primeros análisis realizados y terminando con los últimos clasificadores probados.

De esta forma, en el siguiente capítulo se hablará de los estimadores que se extrajeron de las curvas diarias de los sensores ZIM con la finalidad de obtener números indicativos de la forma de la curva sin necesidad de analizar la curva completa, como pueden ser el valor máximo, el mínimo, el área que encierra o la pendiente de recuperación. También se comprobará su correlación con el valor del nivel de estrés hídrico, con el fin de

[1] https://scikit-learn.org/stable/index.html

determinar cuál de ellos aporta mayor cantidad de información. Estas pruebas se realizaron sobre los datos de 2019, ya que fueron los primeros datos recibidos para el proyecto.

El capítulo 3 trata sobre un clasificador basado en una aproximación mediante mínimos cuadrados. Este clasificador intenta aproximar mediante una función continua los valores discretos del nivel de estrés, tomando como valores de entrada los estimadores explicados en el capítulo 2 que presentasen mejor correlación y ajustando sus parámetros mediante la fórmula de mínimos cuadrados. Una vez diseñada la función, se clasificarían los valores de salida según si superaban o no ciertos umbrales, obteniendo así los valores discretos del clasificador.

Para comprobar el efecto que podría tener seleccionar un estadístico u otro, se realizaron pruebas con los 12 mejores estadísticos propuestos, según su correlación.

En el capítulo 4 se proponen nuevos clasificadores que intentan mejorar esta primera clasificación, empleando para ello medidas más directas de la curva y clasificadores más complejos, como los proporcionados por la librería scikit-learn. Dado que se requería más volumen de datos para entrenar los clasificadores, se solicitaron los datos de 2011 y 2014, aunque finalmente sólo se emplearon los datos de 2014, por requerir el uso de la radiación neta para poder estimar con precisión las horas de amanecer y atardecer percibidas por la planta, con el fin de normalizar la escala temporal. Debe tenerse en cuenta que factores meteorológicos, como un exceso de nubosidad, pueden modificar esta percepción, haciendo que la planta perciba un amanecer más tardío o un atardecer más temprano, por lo que es más preciso emplear este dato que la hora solar.

Para estos clasificadores se ha normalizado la medida en base a la media y varianza de cada día, con el fin de eliminar los efectos de deriva lenta producidos por el crecimiento de la planta y por desajustes en la calibración, que no proporcionan información al clasificador.

Finalmente, en el capítulo 5 se exponen las conclusiones extraídas del trabajo y se hace una reflexión sobre posibles avances posteriores de cara a continuar con el proyecto.

2 Estimadores de estrés hídrico y características de la curva

Para interpretar los datos proporcionados por los sensores ZIM y diseñar los clasificadores se han propuesto una serie de estimadores del estado de estrés hídrico diario a partir de la curva medida por los mismos: Los valores máximos y mínimos de la curva, el área bajo ella[1] y la pendiente de recuperación vespertina, medida tanto en el instante en que se inicia como a través de un valor medio en la hora siguiente y hasta medianoche.

La figura 2.1 muestra una representación de los estimadores considerados sobre un ejemplo de curva LTP.

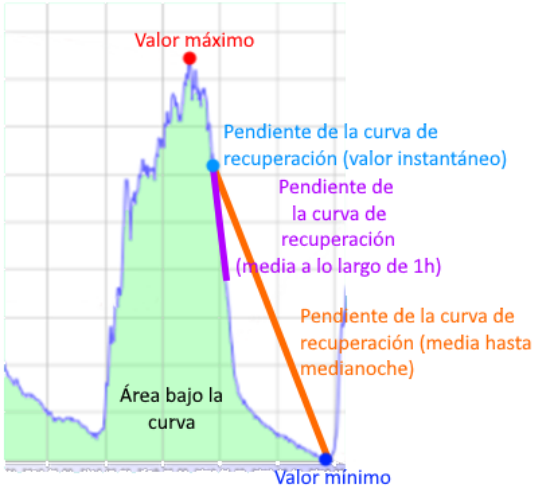

Figura 2.1 Estimadores de estrés hídrico a partir de sensores ZIM..

En las siguientes secciones se explica en detalle cada uno de estos estimadores, con un breve análisis a priori de su relación esperada con el estado hídrico. Además, se han considerado algunas variaciones y normalizaciones de estos parámetros que pueden añadir cierta robustez ante determinados cambios en las medidas del sensor, por ejemplo ante un cierto error en la calibración o una deriva temporal de la medida.

Para comprobar el grado de utilidad de cada estimador y sus variantes de cara al diseño de clasificadores, se ha realizado una tabla que recoge la covarianza entre los mismos y el nivel de estrés hídrico de cada uno de los árboles disponibles, tanto a nivel individual como en conjunto. Un fragmento de esta tabla con los datos más relevantes, junto a las conclusiones extraídas de ella, puede consultarse en la sección de análisis y resultados al final del capítulo.

[1] En este trabajo se ha llamado área bajo la curva a la integral de la curva LTP medida por los sensores ZIM. No confundir con el área bajo la curva ROC empleada en machine learning, que es la integral de la curva característica de funcionamiento del receptor.

Los datos empleados en estos cálculos han sido previamente remuestreados para obtener valores minutales sincronizados, en lugar de los valores cincominutales proporcionados. De esta forma, el volumen de datos es mucho mayor, pero los cálculos se vuelven más sencillos.

Para cada estimador, se proporciona una ecuación que resume el método de cálculo. En todas las ecuaciones, la señal remuestreada de los sensores se ha denominado $LTP(j,k)$, siendo j un índice que indica el día concreto y k el minuto dentro de ese día, con valores entre 0 y 1439 que se corresponden con las 00:00 y las 23:59, respectivamente.

2.1 Valor máximo

Consiste en el valor más alto obtenido cada día. Una curva correspondiente a un árbol con un estado 1 de estrés debería proporcionar un valor máximo más alto que una correspondiente al estado 2, que a su vez tendrá un valor mayor que la correspondiente al estado 3, como se puede observar en la figura 2.2.

$$LTPmax(j) = \max_k(LTP(j,k)). \tag{2.1}$$

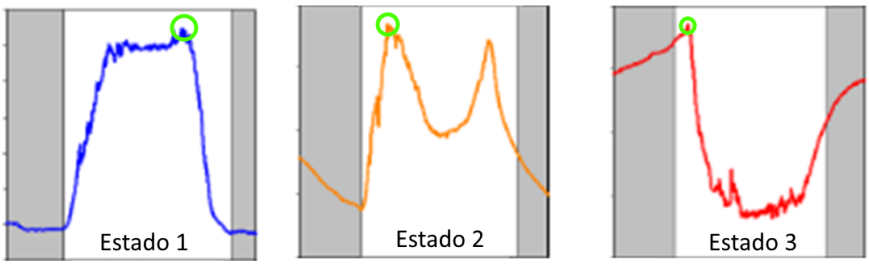

Figura 2.2 Máximo diario en cada estado de estrés..

En la práctica, este estimador puede verse muy afectado tanto por la deriva del sensor como por errores de calibración, además de por características del árbol como el grosor de la hoja elegida, por lo que no se pueden comparar datos de sensores con distinta calibración o en distintas hojas. Para corregirlo, a continuación se proponen varias formas de normalizar el estimador.

2.1.1 Diferencia entre el valor máximo y el valor inicial diario

Consiste en restar al valor máximo el valor obtenido al comienzo del día. Este valor inicial presenta una deriva debida a la desviación entre la hora solar y la hora oficial, además de un cierto ruido debido entre otros factores a las condiciones meteorológicas, pero a cambio permite eliminar el efecto de un valor continuo en el sensor.

$$LTPmax_0(j) = \max_k(LTP(j,k)) - LTP(j,0). \tag{2.2}$$

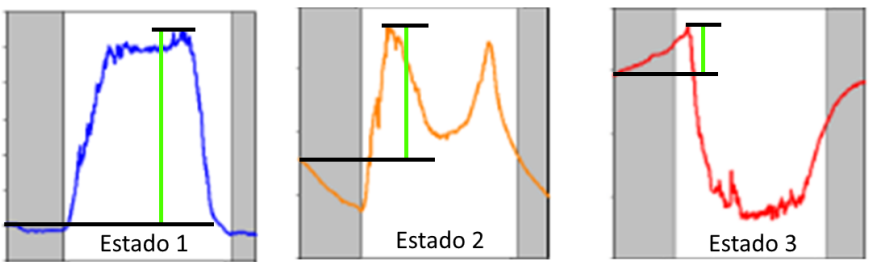

Figura 2.3 Diferencia entre el valor máximo y el inicial..

2.1.2 Diferencia entre el valor máximo y la media acumulada a 10 días

Se define la media dentro de un día como:

$$\overline{LTP}(j) = \frac{1}{N}\sum_{k=0}^{N-1} LTP(j,k), \tag{2.3}$$

donde N es el número de valores de LTP ese día (típicamente 1440). Se define la media acumulada a 10 días como:

$$\overline{LTP_{10D}}(j) = \frac{1}{10}\sum_{i=0}^{9} \overline{LTP}(j-i). \tag{2.4}$$

El estadístico a considerar es el resultado de restar al valor máximo del día el valor medio de los 10 últimos días:

$$LTPmax_{10D}(j) = \max_k(LTP(j,k)) - \overline{LTP_{10D}}(j). \tag{2.5}$$

Este valor no se ve afectado por la desviación entre la hora solar y la oficial, y es más robusto ante condiciones meteorológicas, pero requiere almacenar al menos el valor medio de 10 días para su cálculo.

2.1.3 Valor máximo normalizado a la media y varianza diarias

Retomando la definición de la media de la ecuación 2.3, se define la varianza dentro de un día como:

$$varLTP(j) = \frac{1}{N}\sum_{k=0}^{N-1} (LTP(j,k) - \overline{LTP}(j))^2, \tag{2.6}$$

donde N es el número de valores de LTP ese día (típicamente 1440). El estadístico a considerar es el resultado de restar al valor máximo el valor medio del día y dividir el resultado por la varianza en ese día:

$$LTPmax_{avg,var}(j) = \frac{\max_k(LTP(j,k)) - \overline{LTP}(j)}{varLTP(j)}. \tag{2.7}$$

Esto permite realizar una normalización de la curva cada día, eliminando errores de calibración y derivas, tanto de valor continuo como deformaciones lineales de la señal. No obstante, esto también puede introducir alteraciones en la medida del máximo que limiten la utilidad de la misma.

2.2 Valor mínimo

Consiste en el valor más bajo obtenido cada día. Una curva correspondiente a un árbol con un estado 3 de estrés debería proporcionar un valor mínimo más alto que una correspondiente al estado 2 o al estado 1, como se puede observar en la figura 2.4.

$$LTPmin(j) = \min_k(LTP(j,k)). \tag{2.8}$$

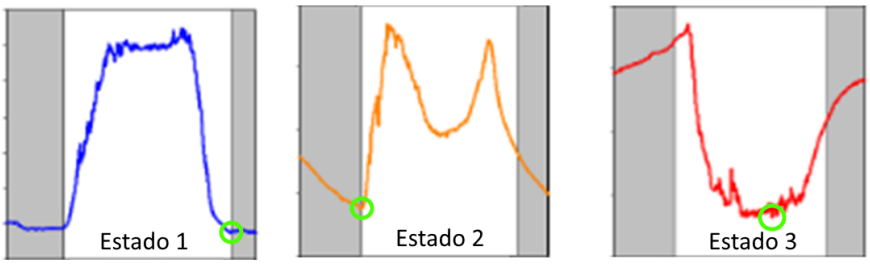

Figura 2.4 Mínimo diario en cada estado de estrés..

Al igual que ocurría con el valor máximo, este estimador puede verse muy afectado tanto por la deriva del sensor como por errores de calibración, además de por características del árbol como el grosor de la hoja elegida. Para corregirlo, a continuación se proponen varias formas de normalizar el estimador.

2.2.1 Diferencia entre el valor mínimo y el valor inicial diario

Consiste en restar al valor mínimo el valor obtenido al comienzo del día. Este índice siempre será negativo.

$$LTPmin_0(j) = \min_k(LTP(j,k)) - LTP(j,0). \tag{2.9}$$

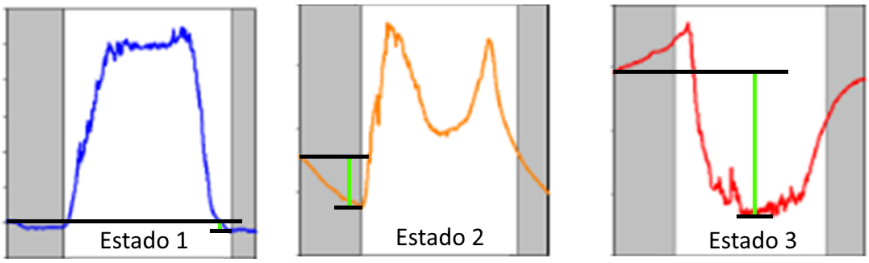

Figura 2.5 Distancia entre el valor mínimo y el inicial..

2.2.2 Diferencia entre el valor mínimo y la media acumulada a 10 días

Consiste en restar al valor mínimo el valor medio de los 10 últimos días. Este valor no se ve afectado por la desviación entre la hora solar y la oficial, y es más robusto ante condiciones meteorológicas, pero requiere almacenar al menos el valor medio de estos días.

Recuperando la definición de la media acumulada a 10 días de la ecuación 2.4,

$$LTPmin_{10D}(j) = \min_k(LTP(j,k)) - \overline{LTP_{10D}}(j). \tag{2.10}$$

2.2.3 Valor mínimo normalizado a la media y varianza diarias

Consiste en restar al valor mínimo el valor medio del día y dividir el resultado por la varianza en ese día. Esto permite realizar una normalización de la curva cada día, eliminando errores de calibración y derivas, tanto de valor continuo como deformaciones lineales de la señal. No obstante, como ocurre con el máximo, esto también puede introducir alteraciones en la medida del mínimo que limiten la utilidad de la misma.

Recuperando las definiciones de media y varianza de las ecuaciones 2.3 y 2.6,

$$LTPmin_{avg,var}(j) = \frac{\min_k(LTP(j,k)) - \overline{LTP}(j)}{varLTP(j)}. \tag{2.11}$$

2.3 Área bajo la curva (valor medio)

El área bajo una curva puede calcularse como la integral de la misma. Para variables discretas, una aproximación de la integral puede ser la suma de los valores disponibles, multiplicados por el tiempo entre los mismos; en el caso de variables minutales, considerando el minuto la unidad de medida de tiempo, puede emplearse la ecuación:

$$AreaLTP(j) = \sum_{k=0}^{N-1} LTP(j,k). \tag{2.12}$$

Para la medida del área bajo la curva se ha optado por emplear en su lugar el valor medio diario, definido anteriormente en la ecuación 2.3:

$$\overline{LTP}(j) = \frac{1}{N}\sum_{k=0}^{N-1} LTP(j,k) = \frac{AreaLTP(j)}{N}, \tag{2.13}$$

donde N es el número de valores de LTP ese día (típicamente 1440).

Esta medida es directamente proporcional al área, pero tiene la ventaja de emplear las mismas unidades que otros estadísticos, además de encontrarse en órdenes de magnitud similares.

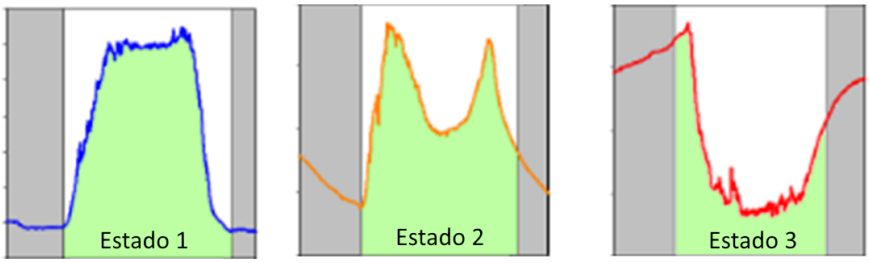

Figura 2.6 Área bajo la curva..

De nuevo, este estadístico puede verse afectado por la deriva del sensor y por errores de calibración, por lo que se han propuesto algunas modificaciones que mejoren su robustez.

2.3.1 Área sobre el valor inicial diario

Representa el área bajo la curva por encima del valor inicial, considerando negativa el área bajo el valor inicial que queda por encima de la curva. Dado que se está empleando el valor medio como una medida del área, puede restarse directamente el valor inicial al valor medio para obtener este estadístico.

$$\overline{LTP_0}(j) = \frac{1}{N}\sum_{k=0}^{N-1}(LTP(j,k) - LTP(j,0)) = \overline{LTP}(j) - LTP(j,0). \tag{2.14}$$

En la figura 2.7, se ha coloreado en verde el área positiva y en rojo el área negativa.

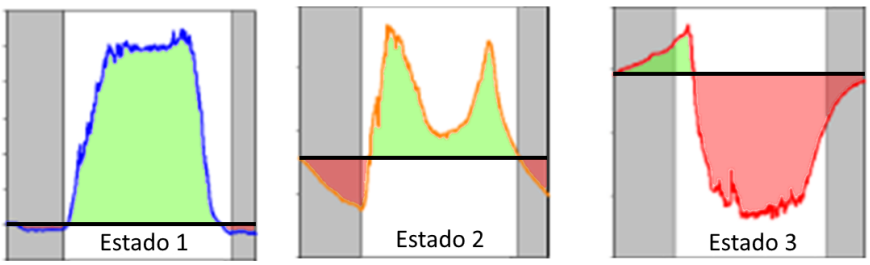

Figura 2.7 Área bajo la curva sobre el valor inicial..

Al igual que con los estadísticos anteriores, este cálculo permite minimizar el efecto de la deriva del sensor sobre la medida obtenida.

2.3.2 Área sobre la media acumulada a 10 días

Recuperando la definición de la media acumulada a 10 días de la ecuación 2.4 y siguiendo el mismo razonamiento que en la ecuación 2.14, se puede obtener el valor del área bajo la curva por encima del valor medio de los diez últimos días, considerando negativa el área bajo dicho valor que queda por encima de la curva:

$$\overline{LTP_{diff10D}}(j) = \overline{LTP}(j) - \overline{LTP_{10D}}(j). \tag{2.15}$$

2.3.3 Fracción del rectángulo que encierra a la gráfica cubierta por el área

Se ha considerado incluir como estadístico la cantidad de área encerrada bajo la curva en relación con el área de un rectangulo definido entre su valor máximo y su valor mínimo, como una medida de la forma de la curva. A la vista de la figura 2.8, no queda claro que este estadístico proporcione una medida distinguible para cada estado, pero debe comprobarse este hecho empleando la correlación entre el estimador y los estados.

$$area_{0,1}(j) = (\overline{LTP}(j) - LTPmin(j))/(LTPmax(j) - LTPmin(j)). \tag{2.16}$$

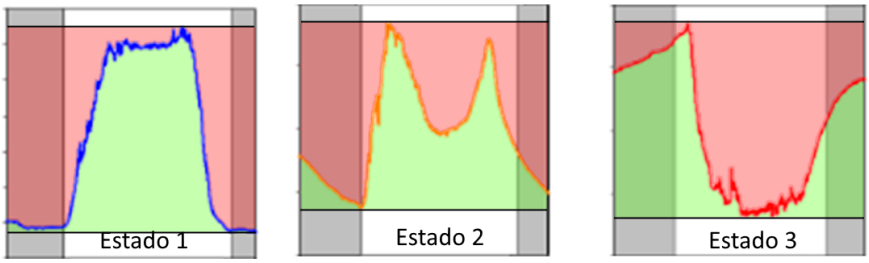

Figura 2.8 Área bajo la curva como fracción del rectángulo que la encierra..

2.4 Pendiente de la curva de recuperación

La curva de recuperación es el nombre que recibe la región de la lectura diaria durante la cual la hoja recupera su nivel de turgencia nocturno. Esta región suele comenzar con la puesta de sol, por lo que se ha optado por emplear la lectura de la radiación solar para determinar el momento en el que se produce, teniendo en cuenta así posibles factores atmosféricos o de elevación del terreno.

Para conseguir el instante en el que se produce la puesta de sol a partir de la radiación solar se puede buscar el último cambio de signo de la radiación neta media, ya que esta medida suele ser positiva durante el día y negativa durante la noche. En las ecuaciones de las siguientes subsecciones, se ha empleado $k_{CSR}(j)$ para indicar el instante en el que se produce el cambio de signo de la radiación el día j.

Se han considerado tres medidas de la pendiente: en un punto, a lo largo de una hora y hasta la medianoche.

2.4.1 Valor instantáneo

Mide el valor de la pendiente en el instante del cambio de signo como el incremento entre el valor en ese instante y el valor inmediatamente anterior. El resultado obtenido es muy susceptible a ruidos en la señal medida y puede verse afectado por derivas en la escala del sensor.

$$diff_i(j) = LTP(k_{CSR}(j), j) - LTP(k_{CSR}(j) - 1, j). \tag{2.17}$$

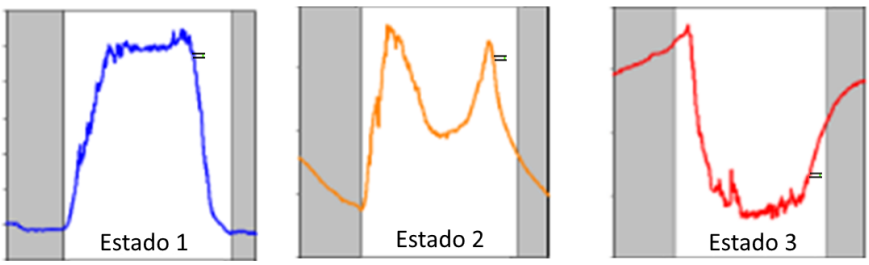

Figura 2.9 Incremento instantáneo en el cambio de signo de la radiación..

Para intentar hacer la medida más robusta ante cambios de escala, se ha probado también a dividir el resultado por la varianza del día, definida en la ecuación 2.6:

$$diff_{i,var}(j) = \frac{diff_i(j)}{varLTP(j)}. \tag{2.18}$$

2.4.2 Media a lo largo de una hora

Mide el valor de la pendiente en el instante del cambio de signo como el incremento entre el valor en ese instante y el valor una hora después.

$$diff_{1h}(j) = LTP(k_{CSR}(j) + 60, j) - LTP(k_{CSR}(j), j). \tag{2.19}$$

De nuevo, se ha considerado una versión robusta antre derivas del sensor dividiendo por la varianza del día:

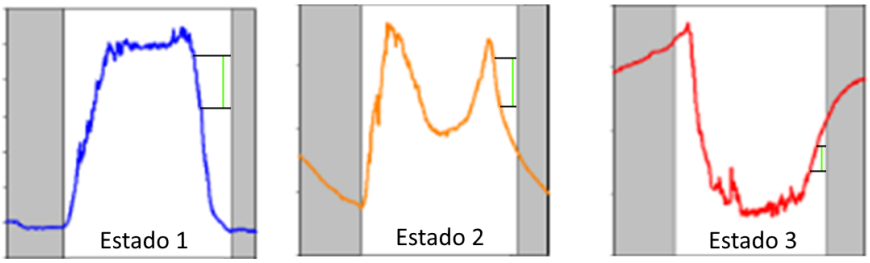

Figura 2.10 Incremento instantáneo en el cambio de signo de la radiación..

$$diff_{1h,var}(j) = \frac{diff_{1h}(j)}{varLTP(j)}. \tag{2.20}$$

2.4.3 Media hasta la medianoche

Mide el valor de la pendiente en el instante del cambio de signo como el incremento entre el valor en ese instante y el valor a medianoche. Este estadístico tiene el problema de verse afectado por la variación entre la hora solar y la hora oficial, lo que puede introducir ruidos debidos a que el árbol se encuentre en distintos instantes del ciclo cada día.

$$diff_0(j) = LTP(0, j+1) - LTP(k_{CSR}(j), j). \tag{2.21}$$

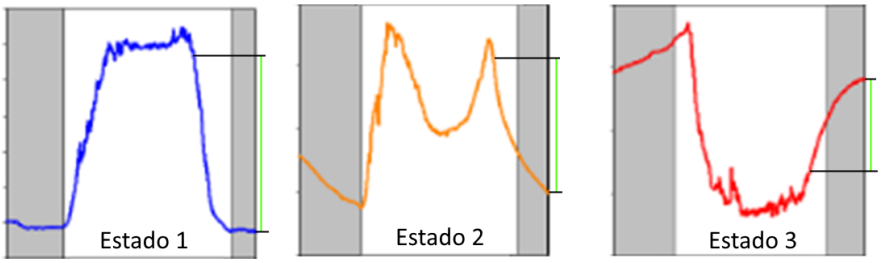

Figura 2.11 Incremento instantáneo en el cambio de signo de la radiación..

De nuevo, se ha considerado una versión robusta dividiendo por la varianza del día:

$$diff_{0,var}(j) = \frac{diff_0(j)}{varLTP(j)}. \tag{2.22}$$

2.5 Factor de olvido

Para tener en cuenta datos históricos sin añadir parámetros adicionales en los clasificadores, se ha aplicado un factor de olvido sobre una ventana finita de cada uno de los estadísticos anteriores. En todos los casos, la ecuación que permite extraer el valor del estadístico con factor de olvido para cada día es:

$$X_{f,w}(j) = \sum_{i=0}^{w} X(j-i) * f^i. \tag{2.23}$$

Siendo X el estadístico a considerar, j el día concreto cuyo valor se desea calcular, f el factor de olvido y w la ventana en que se aplica.

2.6 Análisis y resultados

Con el fin de comprobar el desempeño de los estadísticos con y sin factor de olvido, se ha calculado el coeficiente de correlación entre sus señales y la de los estados de estrés hídrico, agrupando los resultados en

una tabla. La ecuación del coeficiente de correlación entre dos señales X e Y con N muestras es:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \tag{2.24}$$

$$Var(X) = \frac{1}{N}\sum_{k=0}^{N-1}(X_k - \overline{X})^2. \tag{2.25}$$

$$Cov(X) = \frac{1}{N-1}\sum_{k=0}^{N-1}(X_k - \overline{X})(Y_k - \overline{Y}). \tag{2.26}$$

En el caso de los estimadores con factor de olvido, se ha calculado para tamaños de ventana de 1, 2, 3, 4 y 5 días con factores de olvido de 0.8, 0.9, 0.95 y 0.99.

Dado el tamaño de la tabla resultante, se han seleccionado las filas y columnas más representativas a modo de resumen. La siguiente tabla recoge los valores de correlación de los estimadores sin factor de olvido.

Estimador	Control_1_1	Control_2_1	RDI45_2_1	RDI30_3_2	Correlación media
$diff_{1h}$	0.111	0.160	0.82	0.60	0.459
$diff_i$	0.039	0.133	0.79	0.50	0.38
$\overline{LTP_0}$	-0.115	-0.186	-0.70	-0.59	-0.39
$LTPmin_0$	0.20	0.087	-0.70	-0.63	-0.34
$diff_{1h,var}$	-0.0530	-0.08	0.52	0.52	0.30
$LTPmax_0$	-0.14	-0.13	-0.42	-0.57	-0.288
$diff_{i,var}$	-0.080	0.055	0.346	0.30	0.21
$LTPMax$	-0.0185	-0.102	0.191	0.43	0.222
$\overline{LTP_{0,1}}$	-0.168	-0.252	0.18	0.429	0.141
$\overline{LTP}$	0.123	-0.062	0.147	0.66	0.27
$LTPmin$	0.22	0.103	0.079	0.60	0.28
$LTPmin_{avg,var}$	-0.229	-0.24	0.0054	-0.59	-0.200

Se observa que la correlación varía notablemente al modificar el sensor en el que se mide. Además, se observa que la correlación media de cada estimador es en general bastante baja, no llegando a alcanzar un valor de 0.5. Debe tenerse en cuenta que un valor negativo de la correlación indica parecido con una señal inversamente proporcional, y que las señales se parecerán más cuanto más se acerque la correlación a un valor absoluto de 1.

Los sensores se han ordenado en base a la correlación obtenida para el sensor RDI45_2_1. La elección de este sensor se debe a que es el sensor que proporciona simultáneamente más días con una medida fiable y una mayor variación en la señal de estrés real, por lo que se acerca más a las condiciones de trabajo hipotéticas para sistema una vez implementado.

En cuanto a los resultados obtenidos con factor de olvido, se muestran a continuación los mejores resultados obtenidos, omitiendo estadísticos repetidos con distintos factores de olvido por no aportar información nueva. Se observan resultados parecidos a los de la tabla anterior, aunque los resultados son en general algo mejores. Además, se observa que decrece notablemente la diferencia entre estimadores.

Se han omitido las columnas de sensores por motivos de espacio, pero los resultados son análogos a los anteriores en este sentido.

Estimador	Factor de olvido	Tamaño de ventana	Correlación media
$diff_{1h_{0.8,1}}$	0.8	1	0.473
$diff_{i_{0.8,5}}$	0.8	5	0.4539
$diff_{i,var_{0.99,5}}$	0.99	5	0.423
$\overline{LTP_{0_{0.8,2}}}$	0.8	2	-0.4197
$diff_{1h,var_{0.8,5}}$	0.8	5	0.4000
$LTPmin_{0_{0.8,1}}$	0.8	1	-0.3729
$LTPmin_{0.8,5}$	0.8	5	0.32

Si se representa, a modo de ejemplo, el estadístico con mejor correlación frente al estado hídrico de la planta para todos los sensores disponibles en el año 2019, se obtiene la gráfica 2.12, en la que se observa que

la señal continua del estimador tiene cierto parecido con la señal discreta del estado en algunos momentos, pero resulta complicado encontrar una relación entre ambas a simple vista.

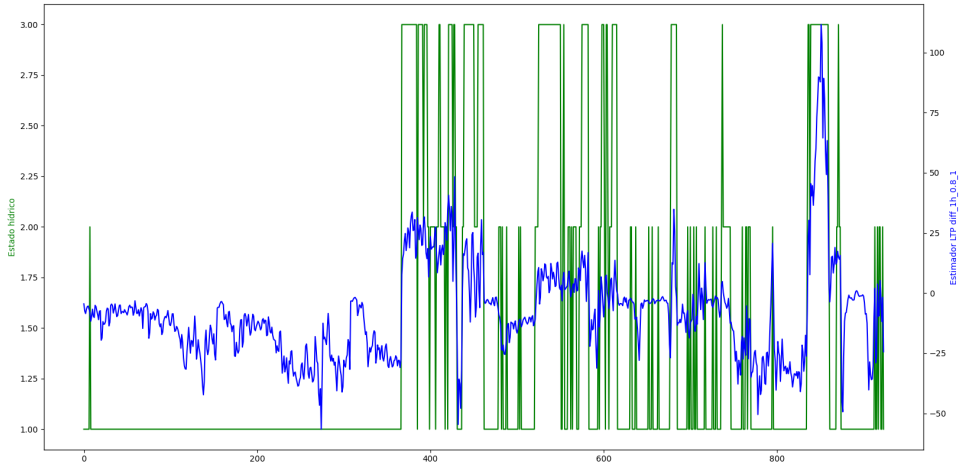

Figura 2.12 Estado de estrés hídrico y valor del estimador $diff_{1h_{0.8,1}}$..

A partir de estos resultados se puede concluir que existe una cierta relación entre los estadísticos y el estado de estrés hídrico, pero también existen otros factores que influyen en los valores obtenidos tanto en el estrés como en cada uno de los estadísticos.

3 Clasificador basado en mínimos cuadrados

El objetivo de esta sección es comprobar si es posible, a partir de los estadísticos con mejor correlación de la sección anterior y empleando el método de mínimos cuadrados, encontrar una función continua que obtenga valores más cercanos a los de estrés, para luego aproximar estos valores continuos a valores discretos mediante un criterio de selección basado en rango.

3.1 Cálculo de mínimos cuadrados

La ecuación general del modelo de mínimos cuadrados es:

$$\hat{y}(k) = \sum_{i=0}^{N} \theta_i \varphi_i(k). \tag{3.1}$$

Donde $\hat{y}(k)$ representa el valor predicho para el instante k, θ_i representa el iésimo parámetro de la ecuación que estima el sistema y $\varphi_i(k)$ representa el iésimo elemento del vector de estimadores en el instante k. Este vector de estimadores deberá contener, además de un elemento por cada estimador a considerar, un elemento de valor 1, que corresponderá a una constante libre en la ecuación resultante.

Para estimar el valor de θ_i se ha seleccionado un subconjunto de datos, al que se conoce como set o conjunto de entrenamiento. Este set debe incluir tanto los datos de las medidas del sensor, con sus respectivos estadísticos, como el nivel de estrés correspondiente. Si se expresan los estadísticos del set como elementos de una matriz:

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{bmatrix}; X_{ij} = \varphi_i(j); i \in [1,n]; j \in [1,m]. \tag{3.2}$$

Puede aplicarse la siguiente fórmula:

$$\hat{\theta} = (X'X)^{-1}X'Y. \tag{3.3}$$

Donde Y es el vector que almacena todos los valores de y(k):

$$Y = [y(1), y(2) \cdots y(m)]'. \tag{3.4}$$

Una vez calculado el valor de $\hat{\theta}$, debe comprobarse su desempeño empleando un nuevo conjunto de datos cuyo nivel de estrés correspondiente también sea conocido. Este nuevo conjunto de datos se denomina set de validación, y en este caso incluye el conjunto completo de datos disponibles.

En cuanto a la medida del desempeño, se obtendrá a través del error cuadrático ponderado, consistente en medir el error cuadrático medio en el set de validación para cada nivel de estrés y posteriormente calcular la media entre los tres errores.

Esta medida del desempeño, no obstante, queda sustituida por el porcentaje de acierto medio una vez implementado el clasificador.

3.2 Selección de estadísticos para el cálculo

El cálculo de mínimos cuadrados tiene el inconveniente de ser propenso al sobreajuste (también conocido como overfitting), consistente en que cuando se proporcionan demasiados parámetros del vector θ para un número pequeño de datos de entrenamiento, los parámetros adicionales empiezan a captar ruidos y valores atípicos como parte de la señal a estimar. Como consecuencia, a partir de un cierto número de parámetros, la señal estimada para conjuntos de datos que no pertenezcan al set de entrenamiento empieza ser muy parecida a la obtenida con menos parámetros o incluso puede dejar de parecerse a la señal real.

Por otro lado, probar todas las combinaciones de estadísticos posibles dentro de los planteados por este trabajo, incluyendo las variaciones en el cálculo del factor de olvido, supone un tiempo de cálculo muy extenso que proporciona un gran número de resultados similares entre sí.

En consecuencia, y partiendo del análisis realizado en el capítulo anterior, se ha optado por elegir los seis estimadores que mejor correlación proporcionan para el sensor RDI45_2_1, con el factor de olvido con el que lo hacen y sin factor de olvido, obteniendo un total de 12 estimadores a probar.

Para elegir el número de parámetros óptimo, se han entrenado modelos de mínimos cuadrados de 2 a 13 parámetros, contando los parámetros asociados a cada estadístico y al término independiente, con el fin de obtener su desempeño.

Para el entrenamiento se ha seleccionado un subconjunto de las muestras disponibles del año 2019. Los resultados se validarán con el conjunto completo de datos del año.

3.3 Resultados obtenidos mediante mínimos cuadrados

Con el fin de comparar los resultados obtenidos mediante esta implementación, se ha empleado la ecuación del error cuadrático medio:

$$E_{Y_{pred},Y} = \frac{\sum_{i=1}^{N}(Y_k - Y_{pred,k})^2}{N}, \tag{3.5}$$

donde Y es el conjunto de valores esperados de la variable e Y_{pred} es el conjunto de valores predichos.

Si se calculan por separado los valores de error cuadrático medio para las muestras en cada estado de estrés y luego se obtiene el valor medio de los errores cometidos, se obtiene el error cuadrático medio ponderado, que aproxima el valor del error cuadrático si el set de validación fuera un set balanceado, es decir, si hubiera el mismo número de muestras para cada estado de estrés.

Además, observar los errores cuadráticos por estado de estrés permite comprobar si la función tiende a acercarse más a un valor que a otro.

De esta forma se obtiene la tabla de la figura 3.1, en la que se muestran los mejores modelos obtenidos, es decir, aquellos cuyo error cuadrático es menor.

Se observa que el error es considerablemente mayor para el estado 3 en todos los modelos. Además, se puede comprobar que los errores cometidos por todos los modelos son similares en magnitud, por lo que resulta complicado asegurar que un resultado sea mejor que otro, pudiendo deberse el resultado a las características del set de validación empleado.

3.4 Cálculo de mínimos cuadrados ponderados

El cálculo de mínimos cuadrados descrito anteriormente asocia la misma importancia a cada muestra del set de entrenamiento y minimiza el error cuadrático medio de las muestras por igual. Este cálculo es ideal para sets balanceados, donde la distribución de los valores de las muestras es similar a la importancia de estimar bien dichos valores.

No obstante, en este caso, se ha considerado que estimar correctamente los tres posibles valores de estrés tiene aproximadamente la misma importancia, pero el set de entrenamiento contiene un número considerablemente mayor de muestras en el estado 1 que en el estado 3, mientras que apenas hay muestras en el estado 2, lo que ocasiona el desequilibrio observado anteriormente.

Estadísticos	Ventanas	Factores de olvido	E(y=1)	E(y=2)	E(y=3)	Error ponderado
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0 0.8 0 0	0.161771733	0.386171635	2.096121959	0.881355109
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 2 1 1 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0 0.8 0	0.168199987	0.36551478	2.110514201	0.881409656
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 2 1 1 1 1 1 1 1	0 0.8 0 0.8 0.8 0 0.8 0	0.166366922	0.367661879	2.113234851	0.882421217
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1 10 1	0 0.8 0 0.8 0 0.8 0 0.8 0.8 0	0.161323335	0.393339173	2.092838351	0.882500286
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1 1	0 0.8 0.8 0 0.8 0 0.8 0 0	0.16024547	0.389165583	2.098551887	0.882654313
LTP Area_n0 independiente	2 1	0.8 0	0.175382213	0.398499026	2.074500656	0.882793965
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1 1	0 0.8 0 0.8 0.8 0 0.8 0 0	0.160983184	0.391627908	2.096553445	0.883054846
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 2 1 1 1 1 1 1 1	0 0.8 0.8 0 0.8 0 0.8 0	0.166143188	0.367923328	2.117090371	0.883718962
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0.8 0 0	0.16148221	0.392430843	2.098316426	0.884076493
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 10 1	0 0.8 0.8 0 0.8 0 0.8 0.8 0	0.159944811	0.396688369	2.095692715	0.884108632
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 10 1	0 0.8 0 0.8 0.8 0 0.8 0.8 0	0.160712616	0.399525637	2.092878188	0.884372147
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 10 1	0 0.8 0 0.8 0 0.8 0.8 0 0	0.161217027	0.397994215	2.094114455	0.884441899
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1	0 0.8 0 0.8 0.8 0.8 0 0	0.161029924	0.395459573	2.098255959	0.884915152
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 2 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0	0.165712251	0.370151089	2.118901629	0.884921656
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 2 1 1 1 1 1 1	0 0.8 0 0 0.8 0 0.8 0	0.166436762	0.370694994	2.118633591	0.885255115
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0 0	0.159874587	0.396923787	2.099806238	0.885534871
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 10 1	0 0.8 0 0.8 0.8 0.8 0.8 0	0.160801056	0.402027914	2.094058378	0.885629116
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1	0 0.8 0.8 0.8 0 0.8 0 0	0.158842731	0.399808382	2.098329847	0.88566032
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0 0	0.159822745	0.390249426	2.106923771	0.885665314
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1 1	0 0.8 0 0 0.8 0 0.8 0 0	0.159965076	0.389769248	2.107343303	0.885692542
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 2 1 1 1 1 1	0 0.8 0.8 0.8 0 0.8 0	0.162946286	0.375375267	2.119081496	0.885801016
LTP Area_n0 LTP Area_n0 independiente	1 2 1	0 0.8 0	0.176397957	0.403808038	2.078041113	0.88608237
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 10 1	0 0.8 0.8 0 0.8 0.8 0.8 0	0.159787075	0.403339699	2.096533139	0.886553304
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 10 1	0 0.8 0 0 0.8 0 0.8 0.8 0	0.159507164	0.396825116	2.103564664	0.886632315
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 10 1	0 0.8 0 0.8 0 0.8 0.8 0	0.15940496	0.397524382	2.103184751	0.886704698
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h independiente	1 2 1 1 1 1 1 1	0 0.8 0 0.8 0 0.8 0.8 0	0.167650334	0.371229067	2.1229492	0.887276201
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 10 1	0 0.8 0.8 0.8 0 0.8 0.8 0	0.158792497	0.408296698	2.094835487	0.887308227
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1	0 0.8 0 0.8 0.8 0 0	0.159727938	0.393121331	2.109307516	0.887385595
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1 1	0 0.8 0 0 0.8 0 0.8 0	0.159759797	0.393027819	2.109494497	0.887427371
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 1	0 0.8 0.8 0.8 0 0 0	0.158994691	0.404242219	2.099934715	0.887723875
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 10 1	0 0.8 0 0.8 0.8 0.8 0	0.159406678	0.399167451	2.104784277	0.887786136
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg LTP diff_R_Neta_A	1 2 1 1 1 1 1 1 10 1	0 0.8 0 0.8 0 0.8 0 0.8 0 0.8 (	0.162044559	0.396312303	2.105027935	0.887794932
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1 10 1	0 0.8 0 0 0.8 0.8 0.8 0	0.159452912	0.398867687	2.105131778	0.887817459
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h independiente	1 2 1 1 1 1 1	0 0.8 0 0.8 0.8 0.8 0	0.166692013	0.371738225	2.125398661	0.887942967
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h independiente	1 2 1 1 1 1 1	0 0.8 0.8 0 0.8 0.8 0	0.165119565	0.374345504	2.126312102	0.88859239
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg independiente	1 2 1 1 1 10 1	0 0.8 0.8 0.8 0.8 0.8 0	0.158960698	0.411744484	2.095909403	0.888871528
LTP Area_n0 LTP diff_1h independiente	2 1 1	0.8 0 0	0.167575118	0.418496919	2.081685139	0.889252392

Figura 3.1 Tabla de resultados de mínimos cuadrados..

Para corregir esta carencia, una opción podría ser eliminar muestras de los estados 1 y 3 hasta que se tenga el mismo número de muestras de cada tipo, pero en ese caso se estaría disminuyendo el número de datos disponibles, facilitando que se produzca sobredimensionado.

En lugar de eso, se ha optado por proporcionar una ponderación que aumente o disminuya la importancia de cada muestra dentro de la ecuación de mínimos cuadrados. Si se calcula el peso de cada muestra en función de su tipo:

$$w(k) = \frac{\sum_{i=1}^{3} N_i}{N_k}, k \in 1,2,3. \tag{3.6}$$

Donde $w(k)$ es el peso de las muestras de tipo k y N_i el número de muestras de tipo i.

Puede aplicarse el peso a la ecuación de mínimos cuadrados modificando los valores de $y(k)$ y $\varphi(k)$ en función del peso correspondiente al estrés en el instante k:

$$y_w(k) = y(k)w(y(k)). \tag{3.7}$$

$$\varphi_{i,w}(k) = \varphi_i(k)w(y(k)). \tag{3.8}$$

Empleando los valores modificados con las ecuaciones de mínimos cuadrados explicadas anteriormente, se obtiene una nueva aproximación de mínimos cuadrados que captará mejor los estados más escasos, a costa de reducir su efectividad en los estados más frecuentes del set de entrenamiento.

3.5 Resultados obtenidos mediante mínimos cuadrados ponderados

Con el fin de comparar la efectividad de ambas versiones de mínimos cuadrados, se han calculado de nuevo los mismos modelos empleando mínimos cuadrados ponderados.

Se observa que, si bien el error cuadrático del estado 1 es mayor, el error cuadrático ponderado disminuye al disminuir el error de los estados 2 y 3, por lo que se puede considerar que esta aproximación es mejor que la anterior, pasando de valores de error cuadrático superiores a 0.88 a valores cercanos a 0.58.

No obstante, el problema de diferenciar entre sí los modelos obtenidos persiste, ya que el error ponderado varía muy poco de uno a otro.

3.6 Implementación de un clasificador

El resultado obtenido a partir de mínimos cuadrados es una función continua que intenta acercarse a la secuencia discreta obtenida experimentalmente. Los valores continuos obtenidos no tienen realmente un

Estadísticos	Ventanas	Factores de olvido	E(y=1)	E(y=2)	E(y=3)	Error ponderado
LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	2 1 1 1 1	0.8 0 0.8 0 0	0.781142721	0.052756791	0.908628193	0.580842569
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 10 1	0.8 0.8 0.8 0	0.829139219	0.071941698	0.842978247	0.581353055
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 10 1	0 0.8 0.8 0.8 0	0.824401522	0.072758179	0.847239934	0.581466545
LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 10 1	0.8 0 0.8 0.8 0	0.829828561	0.07170401	0.844658907	0.582063826
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1 10 1	0 0.8 0 0.8 0.8 0	0.828926896	0.072069385	0.84523263	0.582076303
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 2 1 1 1 1	0 0.8 0 0.8 0 0	0.769418928	0.051545078	0.925923066	0.582295691
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1 1	0 0.8 0.8 0 0	0.828370788	0.074035332	0.845133594	0.582513238
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1	0.8 0.8 0 0	0.833598895	0.073117524	0.840852197	0.582522872
LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	2 1 1 10 1	0.8 0 0.8 0.8 0	0.7767857	0.057249976	0.913975166	0.582670281
LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1 1	0.8 0 0.8 0 0	0.833870176	0.072139712	0.842673774	0.582894554
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1 1 1	0 0.8 0 0.8 0 0	0.83300397	0.072531582	0.843186085	0.582907212
LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 10 1	0 0 0.8 0.8 0	0.83311547	0.068316321	0.849967246	0.583799679
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 10 1	0 0.8 0 0.8 0	0.822820798	0.074305234	0.854454705	0.583860245
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 2 1 1 10 1	0 0.8 0 0.8 0.8 0	0.765971913	0.055800887	0.929947848	0.583906882
LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1 1	0 0 0.8 0 0	0.837875431	0.067709931	0.848145522	0.584576961
LTP Area_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	2 1 1 1	0.8 0.8 0 0	0.776062725	0.055106291	0.923629304	0.584932773
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1 1	0 0.8 0 0 0	0.827412199	0.075865223	0.851677341	0.584984921
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 10 1	0 0.8 0.8 0	0.822692777	0.073339847	0.859782053	0.585271559
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 10 1	0.8 0 0.8 0	0.835382177	0.07762207	0.843641051	0.585548433
LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h independiente	2 1 1 1 1	0.8 0 0.8 0 0	0.803462507	0.042398438	0.911504774	0.585788573
LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP diff_R_Neta_Avg independiente	2 1 1 1 1	0.8 0 0.8 0 0	0.853902352	0.077777402	0.826671905	0.58611722
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 2 1 1 1	0 0.8 0.8 0 0	0.762869989	0.054298565	0.942052749	0.586407101
LTP Area_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	2 1 10 1	0.8 0.8 0.8 0	0.771725562	0.059640612	0.928300098	0.586555424
LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP diff_1h LTP diff_1h independiente	2 1 1 1 1 1	0.8 0 0.8 0 0.8 0	0.906775229	0.059686923	0.793325548	0.5865959
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1	0 0.8 0 0	0.827943551	0.073951638	0.857954608	0.586616599
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1 1	0.8 0 0 0	0.841860026	0.079045269	0.839505094	0.586803463
LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP diff_1h independiente	2 1 1 1 1	0.8 0 0.8 0 0	0.906114502	0.060008084	0.79448512	0.586869235
LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	2 1 1 1 1 1	0.8 0 0.8 0 0.8 0	0.79595814	0.045241779	0.919884746	0.587028221
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 2 1 10 1	0 0.8 0.8 0.8 0	0.760035121	0.058142798	0.944281284	0.587486401
LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg independiente	1 1 1 1 10 1	0.8 0 0.8 0 0.8 0	0.817214834	0.080004014	0.865746832	0.587655227
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg independiente	1 1 1 1 1 10 1	0 0.8 0 0.8 0 0.8 0	0.816544474	0.080167516	0.866267275	0.587659755
LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg independiente	1 1 1 1 10 1	0 0.8 0.8 0 0.8 0	0.814205984	0.080655278	0.868941517	0.58793426
LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg independiente	1 1 1 10 1	0.8 0.8 0 0.8 0	0.816420656	0.080934844	0.86752105	0.588292183
LTP Area_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	2 1 1 1	0.8 0 0 0	0.791260757	0.062315273	0.911361881	0.588312637
LTP Area_n0 LTP diff_R_Neta_Avg independiente	2 1 1	0.8 0 0	0.822352942	0.071040021	0.871558907	0.58831729

Figura 3.2 Tabla de resultados de mínimos cuadrados ponderados..

significado físico, más allá de intentar acercarse al valor de estrés, por lo que la función resulta poco práctica por si misma.

Por ejemplo, según la tabla 3.2, el modelo con menor error ponderado, y por tanto el mejor de los modelos analizados, se basa en el área bajo la curva sobre el valor inicial diario con una ventana de dos días y un factor de olvido de 0.8, la diferencia entre el valor mínimo diario y el inicial sin memoria y con factor de olvido de 0.8 y ventana de un día, el valor instantáneo de la pendiente de recuperación sin memoria y un término independiente continuo.

Este modelo se ha representado junto a los datos de validación en la figura 3.3, con el fin de obtener una representación visual de la función. Se observa un mayor parecido entre los datos que en la figura 2.12, pero el significado físico del estimador se ha perdido en el proceso.

No obstante, la cercanía entre el valor obtenido y el valor real, suponiendo una efectividad suficiente del cálculo de mínimos cuadrados, debe permitir distinguir una serie de franjas de valores correspondientes a cada uno de los valores discretos originales.

En una situación ideal, estas franjas estarían suficientemente separadas como para considerar que todos los valores dentro de la franja corresponden exclusivamente a su valor discreto asociado, e incluso existiría un espacio entre las franjas en el que no aparecerían valores. En esta situación, el clasificador tendría un acierto del 100 %.

En una situación real, no obstante, suele resultar imposible obtener este nivel de fiabilidad, por lo que en ocasiones se obtendrá un valor predicho que no se corresponderá al valor real. Esto es especialmente cierto cuando se dispone de pocas muestras, cuando no se han tenido en cuenta los estadísticos adecuados o cuando la estrategia de mínimos cuadrados empleada sencillamente no es la más adecuada para resolver el problema.

3.7 Resultados obtenidos por el clasificador

Con el fin de visualizar la distribución de los resultados obtenidos de una forma más cómoda, se ha optado por representar el valor estimado de cada muestra en un eje horizontal, empleando el eje vertical para el porcentaje de muestras con el mismo nivel de estrés que se han estimado con un valor inferior al horizontal, y empleando un color distinto en función del valor real de la muestra. El resultado es una serie de gráficas con forma aproximadamente sigmoidal, cuya interpretación resulta mucho más intuitiva de cara al diseño de un clasificador, como se puede observar en el ejemplo de la figura 3.4, correspondiente al mismo modelo representado en la figura 3.3.

En este caso, se comprueba la cercanía entre las franjas de valores. Para escoger los valores del selector que minimizan el porcentaje de error, se ha optado por buscar el punto en el que la distancia vertical entre una gráfica y la siguiente es mayor, lo que implica que por debajo de ese punto se encuentra la mayor cantidad de valores de un estrés con el mínimo de valores del siguiente.

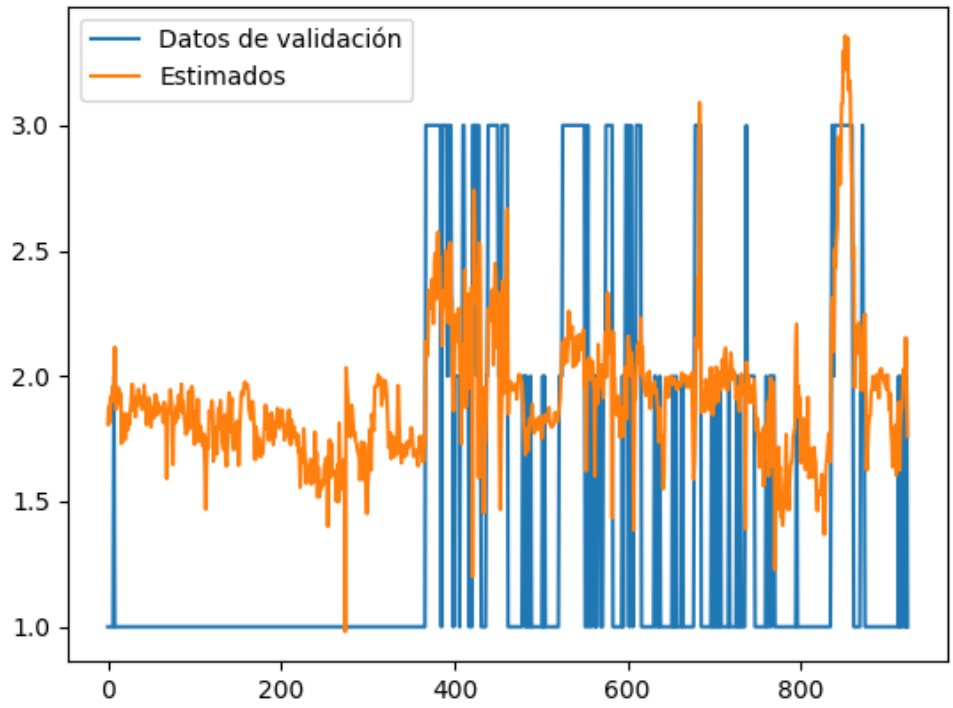

Figura 3.3 Estimación por mínimos cuadrados frente a valor real..

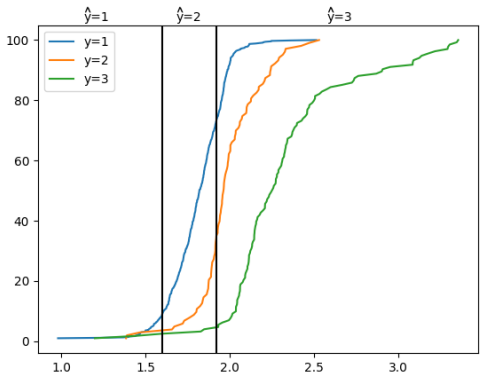

Figura 3.4 Visualización de datos clasificados (modelo de menor error cuadrático)..

Estos valores se calculan en el set de entrenamiento y posteriormente se aplican al de validación, por lo que es posible que el cambio de un set a otro altere las posiciones de las franjas. Este es de hecho el caso del ejemplo, por lo que a pesar de ser la mejor estimación para mínimos cuadrados, resulta en un mal clasificador.

No obstante, dado que la selección de los valores del clasificador es automática, puede aplicarse con facilidad sobre todo el conjunto de modelos calculados, y obtener una medida del acierto para escoger el mejor modelo de los planteados. En este caso, la medida empleada consiste en el porcentaje de acierto balanceado, consistente en calcular el porcentaje de acierto para cada uno de los estados y luego hacer la media entre los tres valores resultantes.

$$A_{balanceado} = \frac{\sum_{i=1}^{3} A_i}{3}. \tag{3.9}$$

$$A_i = \frac{N_{correctas,i}}{\sum_{j=1}^{3} N_j}. \tag{3.10}$$

siendo $N_{correctas,i}$ el número de muestras estimadas correctamente en el estado i y N_j el número de muestras de validación en el estado j.

De esta forma, se puede obtener la tabla 3.5, en la que se han ordenado los distintos modelos en función del acierto balanceado que proporcionan.

Estadísticos	Pesos	Acierto del selec	Acierto en estrés 1	Acierto en estrés 2	Acierto en estrés 3
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h independiente	1 1 1	0.503093369	0.551169591	0.215686275	0.742424242
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg independiente	1 1 1	0.500925144	0.558479532	0.254901961	0.689393939
LTP Area_n0 LTP Max_n0 LTP Max_n0 independiente	1 1 1	0.49789432	0.717836257	0.31372549	0.462121212
LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1	0.497764018	0.640350877	0.352941176	0.5
LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg_norm independiente	1 1 1	0.496554815	0.764619883	0.323529412	0.401515152
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg independiente	1 1 1	0.496406271	0.514619883	0.254901961	0.71969697
LTP Min_n0 LTP diff_R_Neta_Avg_norm independiente	1 1 1	0.496002335	0.766081871	0.343137255	0.378787879
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg independiente	1 1 1	0.495715671	0.578947368	0.362745098	0.545454545
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg_norm i	1.206106870229007	24.30769230769231	7.707317073170732	0.495230947	0.697368421	0.235294118	0.553030303
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg LTP i	1 1 1			0.49501204	0.627192982	0.274509804	0.583333333
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h independiente	1 1 1			0.495009434	0.507309942	0.235294118	0.742424242
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg LTP diff_R_Neta	1 1 1			0.493854958	0.565789474	0.196078431	0.71969697
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Ne	1 1 1			0.493805443	0.621345029	0.284313725	0.575757576
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Ne	1 1 1			0.493805443	0.621345029	0.284313725	0.575757576
LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg_norm independiente	1 1 1			0.493357205	0.668128655	0.37254902	0.439393939
LTP Min_n0 LTP diff_R_Neta_Avg_norm LTP diff_R_Neta_Avg_norm independiente	1 1 1			0.492911572	0.751461988	0.333333333	0.393939394
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Ne	1 1 1			0.492830785	0.618421053	0.284313725	0.575757576
LTP Min_n0 LTP Min_n0 LTP diff_R_Neta_Avg_norm LTP diff_R_Neta_Avg_norm independiente	1 1 1			0.492002064	0.745614035	0.31372549	0.416666667
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg LTP dif	1 1 1			0.491876974	0.574561404	0.264705882	0.636363636
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg LTP	1 1 1			0.491817035	0.567251462	0.196078431	0.712121212
LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h independiente	1 1 1			0.490680802	0.464912281	0.264705882	0.742424242
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Ne	1 1 1			0.490594803	0.592105263	0.303921569	0.575757576
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Ne	1 1 1			0.490594803	0.592105263	0.303921569	0.575757576
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_1h i	1.206106870229007	24.30769230769231	7.707317073170732	0.490443652	0.472222222	0.362745098	0.636363636
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 independiente	1 1 1			0.490383713	0.714912281	0.294117647	0.462121212
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Ne	1 1 1			0.490250805	0.600877193	0.294117647	0.575757576
LTP Min_n0 LTP diff_R_Neta_Avg_norm LTP diff_R_Neta_Avg_norm independiente	1.206106870229007	24.30769230769231	7.707317073170732	0.489685295	0.752923977	0.284313725	0.431818182
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg LTP i	1 1 1			0.489539356	0.622807018	0.254901961	0.590909091
LTP Min_n0 LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg_norm independiente	1 1 1			0.489388206	0.669590643	0.31372549	0.484848485
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg LTP diff_R_Neta	1 1 1			0.489364752	0.590643275	0.294117647	0.583333333
LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg_norm independiente	1.206106870229007	24.30769230769231	7.707317073170732	0.48935954	0.684210526	0.215686275	0.568181818
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP diff_1h LTP diff_1h LTP diff_R_Neta_Avg_norr	1 1 1			0.489052027	0.621345029	0.254901961	0.590909091
LTP Area_n0 LTP Area_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg LTP diff_R_Neta_Avg_norm i	1.206106870229007	24.30769230769231	7.707317073170732	0.488960816	0.675438596	0.215686275	0.575757576
LTP Area_n0 LTP Area_n0 LTP Max_n0 LTP Min_n0 LTP Min_n0 LTP diff_1h LTP diff_R_Neta_Avg inde	1 1 1			0.488447426	0.475146199	0.323529412	0.666666667

Figura 3.5 Tabla de resultados del selector..

En base a la tabla, puede escogerse el mejor modelo para clasificación, con un acierto balanceado del 50 %, representado en la figura 3.6. En este caso, el modelo utiliza el área bajo la curva sobre el valor inicial diario sin memoria y con una ventana de 2 días y factor de olvido de 0.8, un término independiente continuo, el máximo y el mínimo sobre el valor inicial diario y la pendiente de recuperación media a lo largo de una hora, empleando una ventana de un día y factor de olvido de 0.8 en estos tres últimos estadísticos.

Resulta llamativo, además, que es uno de los modelos calculados inicialmente empleando mínimos cuadrados sin ponderación, con un error cuadrático de 0.89, lo que se asocia a una mayor distancia entre el valor continuo estimado y el valor discreto de estrés hídrico correspondiente.

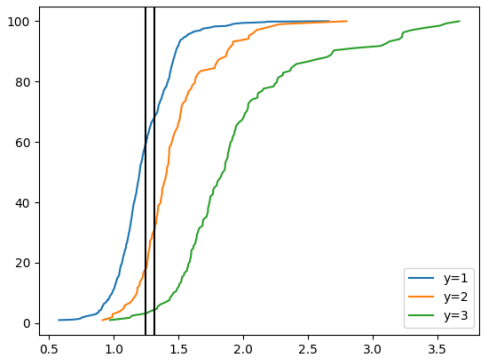

Figura 3.6 Visualización de datos clasificados (modelo con mayor acierto balanceado)..

Este modelo capta peor el estado de estrés 2, pero capta mejor el resto de estados. No obstante, ninguno de los modelos evaluados logra superar el 50.03 % de acierto balanceado, como se observa en la siguiente tabla, en la que la columna pesos indica los pesos que se han aplicado para balancear el cálculo de mínimos cuadrados.

De nuevo, se observa también que el acierto balanceado varía poco entre los modelos probados, por lo que resulta difícil discernir si la variación se debe a cambios en el modelo o al conjunto de datos empleados en validación.

Este tipo de resultados puede deberse a que el algoritmo empleado no sea el más adecuado para la clasificación que se debe realizar, por lo que en las siguientes secciones se probarán otros algoritmos para resolver el problema.

4 Clasificadores basados en la curva diaria

Dado el resultado obtenido por el clasificador anterior, se ha optado por probar el diseño de clasificadores basados directamente en la forma de la curva diaria, en lugar de emplear los estimadores propuestos inicialmente. Este tipo de clasificadores, al emplear un mayor número de variables, requieren habitualmente un mayor número de datos tanto de entrenamiento como de validación Para ello, primero es necesario realizar un preprocesado de los datos, que será común a todos los clasificadores empleados.

Para la evaluación de los resultados se van a emplear dos herramientas simultáneamente. Por un lado, el porcentaje de acierto balanceado, que ya se ha introducido anteriormente, para evaluar el comportamiento global del clasificador. Por otro, la matriz de confusión, que consiste en una matriz que contiene una columna por cada clase que puede predecir un clasificador y una fila por cada clase existente en el conjunto de clasificación, y en la que se anota el número de elementos según la clase a la que pertenecen y la clase predicha por el clasificador. En condiciones ideales, la matriz de confusión es una matriz diagonal cuadrada, lo que indica que el clasificador es capaz de predecir todas las clases disponibles sin cometer errores.

Una versión más fácil de interpretar de la matriz de confusión consiste en normalizarla dividiendo cada fila por el número total de muestras de la clase correspondiente. Puede comprobarse que la media de los valores en la diagonal de la matriz de confusión normalizada coinciden con el valor del porcentaje de acierto balanceado expresado en tanto por uno, aunque aporta más información sobre el tipo de errores cometidos, al mostrar en qué categoría se han clasificado las muestras erróneas.

4.1 Normalización de datos

Para evaluar la forma de la curva primero se realiza una normalización consistente en restar la media de cada día a los valores de ese período y luego dividirlos entre la varianza, obteniendo una curva de 24 horas con media nula y varianza unidad.

Es decir, dado un instante i de un día j y retomando las definiciones de media y varianza de las ecuaciones 2.3 y 2.6:

$$LTP_{valor\ normalizado}(i,j) = \frac{LTP(i,j) - \overline{LTP}(j)}{varLTP(j)}. \tag{4.1}$$

Posteriormente se realiza un ajuste en el tiempo. De manera orientativa y para que sea más sencillo marcar instantes equivalentes de forma independiente al número de muestras tomadas, se ha empleado el concepto de horas normalizadas como medida de tiempo.

Estas horas normalizadas se obtienen para cada día fijando el instante correspondiente al amanecer a las 6:00 y el correspondiente al anochecer a las 18:00 de cada día, y escalando los tramos intermedios mediante interpolación. Para determinar los instantes en los que amanece o anochece cada día se emplean los puntos del cambio de signo de la radiación neta.

Es decir, sea k el valor de tiempo en horas de una muestra tras la normalización, i el valor de la misma muestra antes de normalizar, i_{am} el valor de i en el primer cambio de signo de negativo a positivo de la radiación neta, o amanecer, y i_{pm} el valor de i en el último cambio de signo de positivo a negativo de la radiación neta, o anochecer:

$$k = \begin{cases} 6\frac{i}{i_{am}} & \text{si } 0 \leqslant i \leqslant i_{am} \\ 6+12\frac{i-i_{am}}{i_{pm}-i_{am}} & \text{si } i_{am} < i \leqslant i_{pm} \\ 18+6\frac{i-i_{pm}}{24-i_{pm}} & \text{si } i_{pm} < i \leqslant 24 \end{cases}. \tag{4.2}$$

$$LTP_{normalizado}(k,j) = LTP_{valor\ normalizado}(i,j). \tag{4.3}$$

A la hora de introducirlas en los clasificadores, las señales se han recortado al periodo diurno, ya que se ha observado que el periodo nocturno presenta grandes variaciones dentro de un mismo estado de estrés hídrico, lo cual puede reducir la efectividad del entrenamiento. No obstante, se ha conservado la descripción del periodo nocturno en la definición anterior, ya que sí se ha conservado dicho período al representar las señales normalizadas para observación.

Por último, con el fin de igualar el número de muestras de cada señal y de obtener muestras equidistantes entre sí, se ha modificado la frecuencia de muestreo a un período de muestreo de 0.1h, obteniendo 120 muestras por día. Algunos de los clasificadores utilizados realizan este muestreo a una frecuencia distinta, en función de sus necesidades. Con carácter general, se va a llamar M al número de muestras por día.

La misma normalización en el tiempo, incluyendo el recorte y remuestreo, deberá aplicarse a todas las señales empleadas por los clasificadores, para mantener la sincronización de los datos. Debe tenerse en cuenta que la normalización se realiza por días, de forma que si se emplea memoria de días anteriores, cada día almacenado deberá tratarse por separado, y no emplear el valor medio total.

Representando todas las curvas normalizadas de cada día para cada uno de los años, así como su primera y segunda derivadas, se pueden obtener gráficas como las de las figuras 4.1 a 4.3, que proporcionan información sobre la dispersión de los datos.

Dado que no se disponía de datos de la radiación neta para el año 2011, se ha realizado en su lugar una estimación de la hora de amanecer y atardecer de cada día basada en la radiación global y parcial.

Esta aproximación no es tan exacta como el sensor disponible en los otros años, por lo que se ha optado por no emplear estos datos por el momento en el diseño de clasificadores.

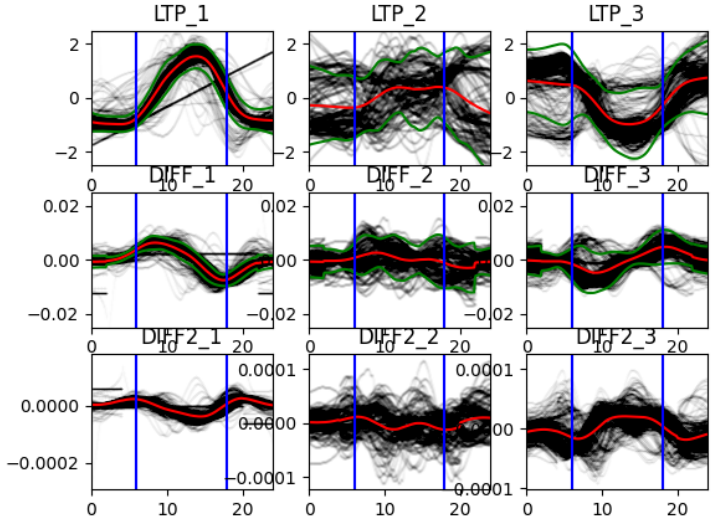

Figura 4.1 Curvas diarias del año 2011..

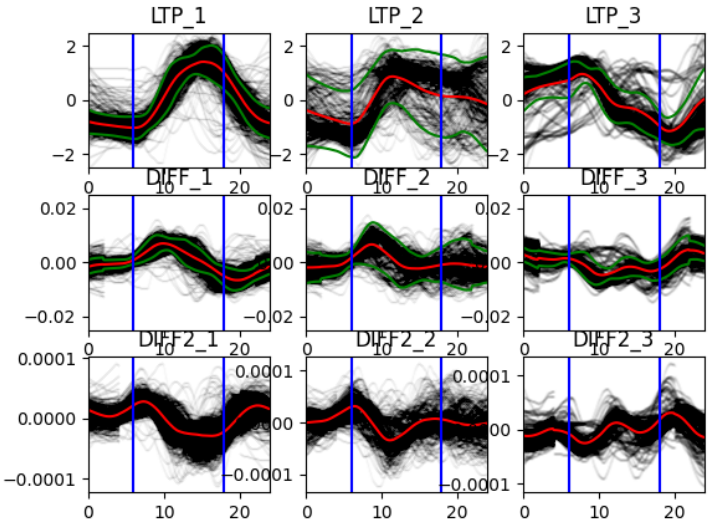

Figura 4.2 Curvas diarias del año 2014..

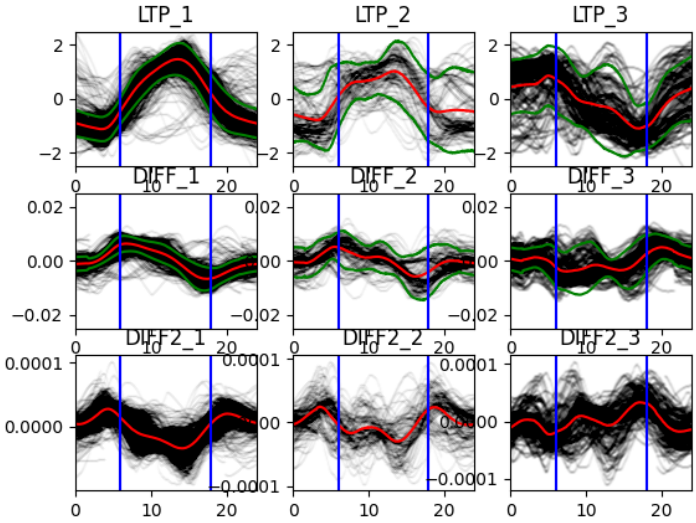

Figura 4.3 Curvas diarias del año 2019..

En estas gráficas, las líneas azules representan el amanecer y el atardecer. Las líneas negras, que poseen un cierto grado de transparencia para que una mayor acumulación de líneas se muestre en un tono más oscuro, representan las curvas individuales de cada día. Las curvas rojas representan los valores medios, y las verdes representan la varianza en torno al valor medio.

En general, se puede ver que la varianza en los tipos 2 y 3 es mucho más elevada que en el tipo 1. Esto se debe en parte a la menor disponibilidad de datos de estos tipos, que implica que las medidas espúreas tienen mayor influencia sobre los resultados, además de a la naturaleza física de estos estados.

Respecto a las medidas espúreas, puede observarse como algunas curvas correspondientes a un estado hídrico tienen una forma más parecida a un tipo de dato diferente que al que realmente le corresponde; por ejemplo, las curvas de tipo 1 en la figura 4.3 tienden a encontrarse entre las líneas de varianza, pero pueden observarse algunas curvas parecidas a las de tipo 3 en las que el pico se encuentra invertido. Esto sugiere que, si bien la forma de la curva suele corresponderse a un estado hídrico, en ocasiones aparece una curva de una forma distinta en su lugar.

Esto puede deberse a varios motivos, entre ellos factores meteorológicos o que afecten a la hoja sobre la que se encuentra el sensor, por lo que puede ser necesario una segunda capa de clasificación que tenga en cuenta estos factores además de la forma de la curva y corrija la primera clasificación. Este tema se tratará más adelante.

Para los siguientes clasificadores se han empleado los datos del año 2014 para entrenamiento y los del año 2019 para validación. De esta forma, se puede comprobar si el clasificador sigue siendo válido pasado un tiempo, y se maximiza el número de datos disponibles para entrenamiento y validación, sin que se empleen datos repetidos.

4.2 Clasificador basado en el error cuadrático

Una aproximación inicial a un clasificador sencillo basado en la forma de la curva puede consistir en el cálculo del error cuadrático entre una curva tipo y la curva diaria que se desea clasificar. Si se realiza este cálculo para una curva tipo correspondiente a cada uno de los estados, el menor valor de error cuadrático se corresponderá con un mayor parecido, por lo que se puede considerar que la curva diaria pertenece a la misma clase que la curva tipo.

Una forma de entrenar este clasificador consiste en basar las curvas tipo en los valores medios de los datos en un intervalo de tiempo, para luego validar los resultados empleando los datos de un intervalo distinto. Siendo N el número de días de los que se disponen muestras y M el número de datos por día:

$$LTP_{ref,estado\ i}(k) = \frac{\sum_{j=1}^{N} LTP_{entrenamiento,estado\ i}(j,k)}{N}. \tag{4.4}$$

Una vez obtenidas las tres curvas, se calcula el error cuadrático para cada estado y día como:

$$E_{estado\ i}(j) = \frac{(\sum_{k=1}^{M} LTP(k,j) - LTP_{ref,estado\ i}(k))^2}{M}. \tag{4.5}$$

El estado estimado cada día será el que corresponda a un menor valor del error cuadrático.

En este caso se han seleccionado como datos de entrenamiento los del año 2014 para estimar los estados del año 2019, proporcionando un porcentaje de acierto balanceado del 57.94 %.

La matriz de confusión de este clasificador, entrenado con los datos de 2014 y validado con los de 2019, puede observarse en la ecuación 4.6:

$$C_E = \begin{bmatrix} 247 & 417 & 23 \\ 14 & 59 & 30 \\ 0 & 26 & 108 \end{bmatrix}. \tag{4.6}$$

Es decir, de todas las muestras que pertenecen al estado 1, se han clasificado correctamente 247, mientras que 417 se han clasificado como estado 2 y 23 como estado 3; de las pertenecientes al estado 2, 14 se han clasificado como estado 1, 59 como estado 2 y 30 como estado 3; y de las pertenecientes al estado 3, 26 se han clasificado como estado 2 y 108 como estado 3. Si se dividen estas muestras por el número total de muestras del tipo correspondiente, se obtiene la matriz de confusión normalizada:

$$C_{E,n} = \begin{bmatrix} 0.35953421 & 0.6069869 & 0.03347889 \\ 0.13592233 & 0.57281553 & 0.29126214 \\ 0 & 0.19402985 & 0.80597015 \end{bmatrix}. \tag{4.7}$$

Se comprueba que el clasificador consigue distinguir bastante bien las curvas de tipo 1 y las de tipo 3, pero comete algunos errores al intentar distinguir las de tipo 2 de las de tipo 3 y es prácticamente incapaz de distinguir las de tipo 1 de las de tipo 2, habiendo clasificado la mayoría en el segundo tipo.

Como ya se mencionó anteriormente, en algunos casos la forma de la curva no coincide con el estado correspondiente, por lo que resulta de interés comprobar las curvas mal clasificadas para observar los posibles errores. En este caso, se puede observar que muchos de esos casos, como el de la figura 4.4, se deben a un desplazamiento en la curva.

Aún así, puede observarse que se ha conseguido un aumento apreciable del porcentaje de acierto balanceado sobre el alcanzado por el clasificador basado en mínimos cuadrados.

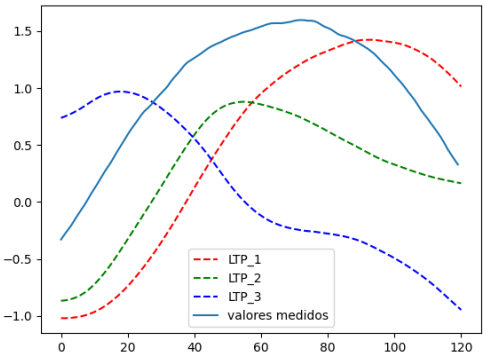

Figura 4.4 Curva de tipo 1 clasificada como tipo 2 por el clasificador por error cuadrático..

4.3 Clasificador basado en la convolución

Una forma de tener en cuenta el desplazamiento e intentar corregir el error del clasificador anterior es emplear un clasificador basado en la integral de convolución, también llamada simplemente convolución. La convolución es una operación matemática, muy empleada en análisis de señales, cuyo cálculo se expresa en la ecuación 4.8.

$$(f*g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau. \tag{4.8}$$

Este cálculo equivale a hallar el área bajo una curva resultante de multiplicar ambas señales cuando una de ellas se desplaza un valor τ, que va recorriendo todos los posibles valores para generar una nueva señal. Habitualmente, dado que las señales f y g suelen ser limitadas en el tiempo, se omiten todos los puntos en los que las señales tienen valor nulo, por lo que la convolución tiene una longitud igual a la suma de las longitudes de las señales originales. El valor de la convolución será máximo para las posiciones en las que los picos estén más cerca entre sí.

Sea j un día concreto y $LTP_{ref,estado\ i}(k)$ la curva de referencia para el estado i definida en la ecuación 4.4:

$$ConvLTP_{estado\ i}(j) = (LTP(j,k) * LTP_{ref,estado\ i}(k))(k). \tag{4.9}$$

En este caso, se extraen dos valores de interés: la posición del máximo de la convolución y su valor en ese punto.

Representando los valores obtenidos para cada curva en una gráfica y coloreándolos en rojo, verde y azul según la clase a la que pertenecen se obtiene la figura 4.5, en la que se observa que este método tampoco separa bien las clases 1 y 2, que tienen formas muy parecidas.

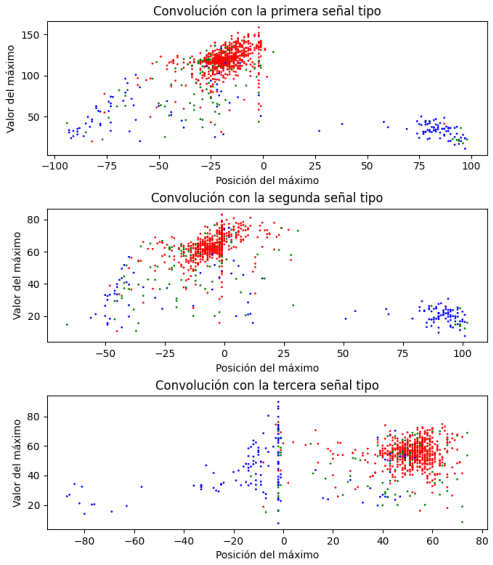

Figura 4.5 Distribución de valores de posición del máximo y valor del máximo de convolución para cada estado de estrés (rojo, verde y azul respectivamente, para los estados 1, 2 y 3) y señal tipo..

Se han obtenido dos clasificadores, uno por mínimo desplazamiento de la convolución y otro por mayor valor de convolución. El primero, con un porcentaje de acierto balanceado del 51.29 % devuelve la siguiente matriz de confusión:

$$C_{C,Desp} = \begin{bmatrix} 38 & 625 & 24 \\ 7 & 69 & 27 \\ 4 & 21 & 109 \end{bmatrix}, \tag{4.10}$$

en la que se observa que clasifica mal las curvas de tipo 1.

Y el clasificador por valor de convolución, con un acierto del 50.89 %:

$$C_{C,Max} = \begin{bmatrix} 132 & 521 & 34 \\ 11 & 56 & 36 \\ 2 & 26 & 106 \end{bmatrix}. \tag{4.11}$$

Por tanto, no se ha conseguido mejorar el resultado obtenido con el clasificador basado en el error cuadrático

4.4 Clasificador por aproximación a una parábola

Se consideró la posibilidad de usar un clasificador basado en la aproximación de la curva diaria por una parábola, mediante un ajuste por mínimos cuadrados, con el fin de emplear los parámetros de la parábola como una medida de la curvatura de la secuencia, disminuyendo notablemente la dimensionalidad del problema.

Para ello, se parte de la ecuación de la parábola, que será distinta para cada día j:

$$P(j,k) = A(j)k^2 + B(j)k + C(j). \tag{4.12}$$

Al aplicar la ecuación de mínimos cuadrados que se presentó en el capítulo 3, tomando $P(j,k)$ como una aproximación de la curva de LTP del día, se obtienen los valores de $A(j)$, $B(j)$ y $C(j)$, también llamados término cuadrático, lineal e independiente, respectivamente.

Con el fin de obtener una idea previa del desempeño de este clasificador, se realizaron las gráficas 4.6 y 4.7, en las que se representan los tres parámetros de la parábola para cada curva de 2014 y 2019, respectivamente, coloreando según la clase a la que pertenecen (rojo, azul y verde para los estados hídricos 1, 2 y 3 respectivamente).

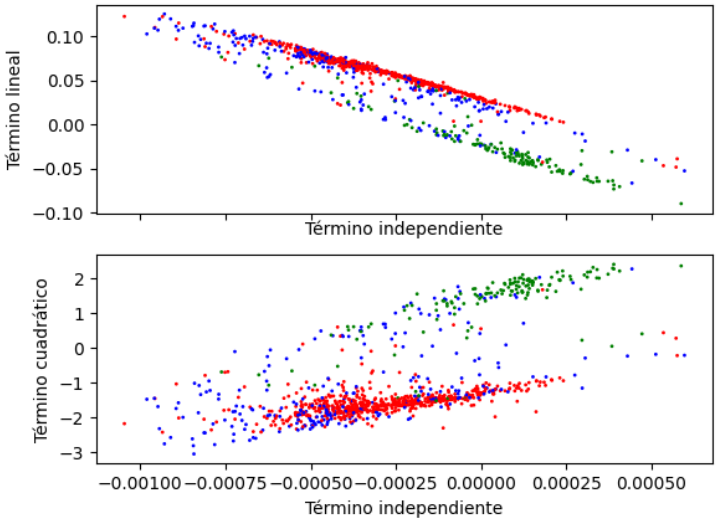

Figura 4.6 Parámetros de parábola para el año 2014..

Se observa como de un año a otro la distribución de los parámetros cuadráticos e independientes ha variado notablemente, por lo que no se espera un comportamiento fiable empleando este método con un selector. No obstante, sí se observa un cierto parecido en los valores del término lineal entre ambos años.

Si se utilizan los datos de 2014 para entrenar un selector basado en el término cuadrático y validarlo en los datos de 2019, se obtiene un acierto balanceado del 48.54 %, por debajo de lo obtenido anteriormente.

$$C_{Parab,Q} = \begin{bmatrix} 183 & 493 & 11 \\ 14 & 68 & 21 \\ 2 & 61 & 71 \end{bmatrix}. \tag{4.13}$$

Para el término lineal se consigue un acierto balanceado del 60 %, lo que mejora los resultados obtenidos hasta ahora. No obstante, en la matriz de confusión puede apreciarse que el estado 2 sigue sin clasificarse correctamente.

$$C_{Parab,L} = \begin{bmatrix} 662 & 6 & 19 \\ 67 & 12 & 24 \\ 5 & 32 & 97 \end{bmatrix}. \tag{4.14}$$

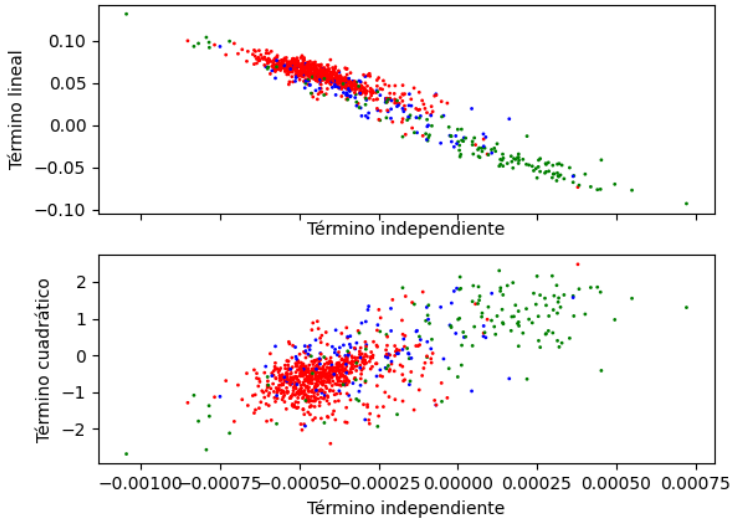

Figura 4.7 Parámetros de parábola para el año 2019..

4.5 Clasificador LDA

El análisis discriminante lineal (LDA, por sus siglas en inglés) se basa en encontrar una combinación lineal de una serie de características que permita distinguir dos o más clases. Si bien esta definición asemeja el LDA al método basado en mínimos cuadrados empleado anteriormente, la principal diferencia se encuentra en la forma de hallar dicha combinación lineal; en este caso, en lugar de intentar hacer que los valores se acerquen a los de un número que identifique cada clase, se busca hallar la probabilidad de que cada conjunto de características pertenezca a una clase u otra en función de la distancia de Mahalanobis, que tiene en cuenta la covarianza de las características.

También puede emplearse LDA para reducir la dimensionalidad del problema, obteniendo proyecciones del problema con menos características que el problema original pero minimizando la pérdida de covarianza entre clases.

Para más información acerca de los fundamentos teóricos del LDA, puede consultarse el apéndice A.

4.5.1 Scikit-learn

Para la implementación de este clasificador se ha utilizado la librería scikit-learn, que puede instalarse en un entorno de python mediante el comando `pip install -U scikit-learn`. En esta librería, cada clasificador se crea como un objeto, que debe ser declarado y entrenado, en ese orden, antes de poder utilizarse.

A continuación hay un fragmento de código de python a modo de ejemplo. Las variables Xtr e Ytr contienen los datos de entrenamiento, mientras que Xv e Yv contienen datos de validación. Todas estas variables son matrices de numpy creadas previamente.

```
import sklearn.discriminant_analysis as sklda
# crea el modelo
clf = sklda.LinearDiscriminantAnalysis(solver='svd')#,shrinkage='auto')
# entrena el modelo
clf.fit(Xtr,Ytr)
# predice los valores de Yv
Ypred=clf.predict(Xv)
```

De esta forma se obtienen los valores predichos para Yv, que en este caso será el estado de estrés hídrico, a partir de los datos almacenados en Xv. Debe tenerse en cuenta que si Xv es una matriz de la forma NxM, donde N es el número de días y M el número de datos para estimar el estado hídrico de cada día, Ypred será un vector con el conjunto completo de días.

Otra opción es emplear `Yprob=clf.predict_proba(Xv)` para obtener la probabilidad de cada muestra de pertenecer a cada clase.

Para estos análisis se ha optado por realizar la predicción directamente, ya que se facilita el análisis de un gran volumen de datos.

4.5.2 Implementación directa sobre los datos diarios del sensor ZIM

Una forma de implementar LDA sobre una curva dada consiste en muestrear la curva y emplear cada punto muestreado como una característica. A la hora de ajustar la frecuencia de muestreo, debe tenerse en cuenta que un número excesivo de muestras requiere más datos de entrenamiento para devolver resultados fiables, mientras que un número demasiado reducido aporta muy poca información sobre la curva.

La matriz X, tanto para entrenamiento como para predicción, se construye:

$$X = \begin{bmatrix} LPT(0,0) & \cdots & LPT(0,M-1) \\ \vdots & \ddots & \vdots \\ LPT(N-1,0) & \cdots & LPT(N-1,M-1) \end{bmatrix}, \tag{4.15}$$

siendo N el número de días y M el número de datos por día.

En cuanto al vector Y:

$$Y = \begin{bmatrix} Estado(0) & \cdots & Estado(N-1) \end{bmatrix}. \tag{4.16}$$

De esta forma, tomando de nuevo como datos de entrenamiento los correspondientes a 2014 y validando con los de 2019, se ha comprobado que empleando las curvas normalizadas a media diaria nula y varianza diaria unidad, y tomando 120 muestras entre el amanecer y el anochecer, definidos por el cambio de signo de la radiación neta, se obtiene un porcentaje de acierto balanceado sobre el estado de estrés hídrico del 56.5 %, algo inferior que el obtenido mediante el clasificador basado en el parámetro lineal de la parábola de mejor ajuste.

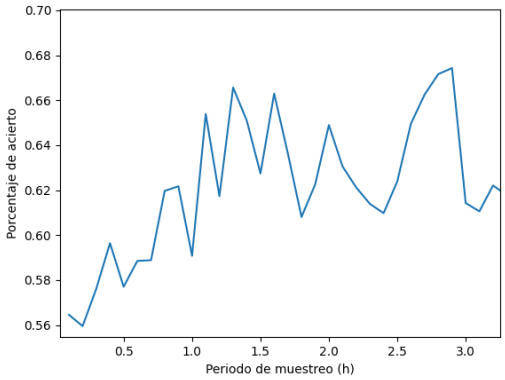

Figura 4.8 Acierto balanceado en función del periodo de muestreo..

Sin embargo, si se disminuye el número de muestras, se observa un incremento en el porcentaje de acierto balanceado, como se aprecia en la figura 4.8, calculada variando el periodo de muestreo de la curva LTP.

El pico más alto, con un porcentaje de acierto balanceado del 67 %, se observa al tomar un periodo de muestreo de 2.6h, lo que corresponde a 4 muestras por día, con la última muestra a las 16:24, según la hora normalizada. Su matriz de confusión es:

$$C_{LDA} = \begin{bmatrix} 525 & 156 & 6 \\ 34 & 46 & 23 \\ 2 & 24 & 108 \end{bmatrix}. \tag{4.17}$$

No obstante, dada la forma de la gráfica 4.8, es posible que este pico se deba a características particulares de los años empleados, y el acierto se encuentre entre el 62 % y el 66 % con un modelo totalmente entrenado.

Este incremento del porcentaje de acierto respecto al valor obtenido inicialmente puede deberse a que, al incrementar el periodo de muestreo, disminuye el número de muestras, y por tanto también el número de

características. Con un periodo de muestreo más reducido, la cantidad de información disponible es mayor, pero se hace necesario emplear un mayor número de datos para el entrenamiento.

Si se aumenta demasiado el periodo de muestreo, el porcentaje de acierto vuelve a disminuir, al carecer de información suficiente.

4.5.3 Implementación con memoria de días anteriores

De forma análoga al procedimiento anterior, pueden añadirse como características los datos de días anteriores, extendiendo las matrices Xtr y Xv. Debe recordarse que las señales se normalizan con carácter diario, por lo que la normalización de los días incluidos en memoria es independiente del resto de días.

La matriz X, por tanto, queda:

$$X = \begin{bmatrix} LPT(0,0) & \cdots & LPT(0,M-1) & LPT(1,0) & \cdots & LPT(n,M-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ LPT(N-(1+n),0) & \cdots & LPT(N-(1+n),M-1) & LPT(N-n,0) & \cdots & LPT(N-1,M-1) \end{bmatrix}, \quad (4.18)$$

donde n es el número de días almacenados en memoria.

Y el vector Y:

$$Y = \begin{bmatrix} Estado(n) & \cdots & Estado(N-1) \end{bmatrix}. \quad (4.19)$$

Debe tenerse en cuenta que algunos días pueden no tener suficientes lecturas previas para el cálculo. En este caso, los sensores LTP se mantuvieron en funcionamiento en todo momento y con la antelación suficiente para realizar el cálculo. En caso de que hubiera faltado información, una solución habría sido propagar la información más reciente hacia atrás, sustituyendo el día faltante por el siguiente día completo disponible.

Tomando por ejemplo dos días en memoria ($n = 2$), tres si se incluye el día actual, se obtiene la gráfica 4.9.

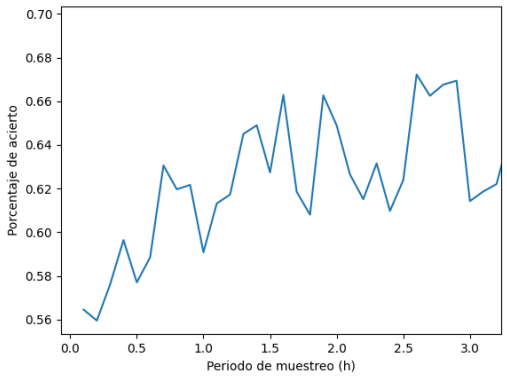

Figura 4.9 Acierto balanceado en función del periodo de muestreo con memoria de 2 días..

Debe tenerse en cuenta que los datos contienen en ocasiones días no consecutivos, lo que no permitiría que el clasificador funcionase correctamente hasta que se hubiera acumulado un número suficiente de datos.

Se puede observar que la gráfica es similar en magnitud a la obtenida anteriormente, lo que indica que la memoria no ha afectado significativamente al clasificador.

Incrementar el número de días en memoria sigue sin modificar el resultado. La gráfica 4.10, idéntica a la gráfica 4.9, se obtiene usando una memoria de 5 días.

En ambos casos, el mejor resultado se ha obtenido con un período de muestreo de 2.6h, presentando un acierto balanceado del 67 % con la matriz de confusión:

$$C_{LDA,mem} = \begin{bmatrix} 525 & 156 & 6 \\ 34 & 46 & 23 \\ 2 & 24 & 108 \end{bmatrix}. \quad (4.20)$$

En consecuencia, y dado que aunque hay variaciones en los resultados, el mejor resultado obtenido es idéntico al obtenido sin memoria, se ha considerado que el efecto de añadir memoria al clasificador no justifica el incremento de coste computacional.

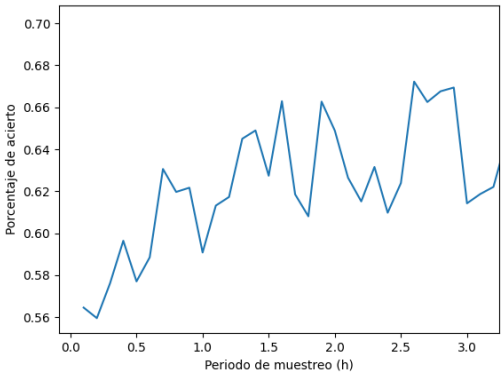

Figura 4.10 Acierto balanceado en función del periodo de muestreo con memoria de 5 días..

4.5.4 Implementación con preprocesado mediante PCA

El análisis de componentes principales busca reducir el número de características de un conjunto manteniendo la mayor varianza posible entre sus integrantes, para lo cual busca las direcciones ortogonales de mayor varianza. Se ha incluido una explicación de sus fundamentos teóricos en el apéndice B.

Como se ha observado, uno de los principales problemas de aplicar LDA con un elevado número de características es el número de datos de entrenamiento necesarios. En los casos anteriores se ha solucionado el problema reduciendo el número de características mediante el muestreado, que implica una cierta pérdida de información. En este caso, puede comprobarse que al haber una menor pérdida de información, el comportamiento mejora, aunque debe tenerse en cuenta que dado que el PCA no distingue las clases, el número de componentes debe escogerse de forma que no opaque características distintivas; si una característica presenta una gran varianza, pero sus valores son independientes de la clase a la que pertenezca la muestra, el PCA la integrará en una de las componentes principales, mientras que el LDA tenderá a descartarla.

Por este mismo motivo, descartar características poco relevantes a la clasificación de clases ayuda a mejorar el acierto del clasificador. En este caso se está empleando la curva muestreada, por lo que de nuevo se podría variar el periodo entre muestras para obtener resultados distintos, pero la elección de características concretas es más compleja.

En este trabajo se ha utilizado la librería scikit-learn para implementar el análisis. Su implementación es similar a la de LDA, empleando un objeto para almacenar la información del entrenamiento y realizar las acciones de entrenamiento y predicción:

```
import sklearn.discriminant_analysis as sklda
import sklearn.decomposition as skdecomp

#aplica PCA
pca = skdecomp.PCA(n_components=comp+1)
pca.fit(XtrBase)
Xtr = pca.transform(XtrBase)
Xv = pca.transform(XvBase)

# aplica LDA
clf = sklda.LinearDiscriminantAnalysis(solver='svd')
# entrena el modelo
clf.fit(Xtr,Ytr)
# predice los valores de Yv
Ypred=clf.predict(Xv)
```

XtrBase y XvBase son los datos de partida antes de aplicar PCA, análogos a las matrices Xtr y Xv de los apartados anteriores. Las matrices Xtr y Xv devueltas por el método `pca.transform(X)` son matrices de numpy que contienen los datos de dimensionalidad reducida.

La matriz XBase, tanto para entrenamiento como para predicción, se construye de forma análoga a la de la implementación directa:

$$X = \begin{bmatrix} LPT(0,0) & \cdots & LPT(0,M-1) \\ \vdots & \ddots & \vdots \\ LPT(N-1,0) & \cdots & LPT(N-1,M-1) \end{bmatrix}, \tag{4.21}$$

siendo N el número de días y M el número de datos por día.

En cuanto al vector Y:

$$Y = \begin{bmatrix} Estado(0) & \cdots & Estado(N-1) \end{bmatrix}. \tag{4.22}$$

El número de componentes escogidas en el análisis de componentes principales es una variable de diseño, y debe ser menor o igual que el número de componentes de partida. En este caso, si el número de componentes de partida es 120, que es el número obtenido al muestrear con un periodo de muestreo de 0.1h, puede hacerse un barrido entre 1 y 120 componentes para buscar el mejor resultado, obteniendo así la imagen 4.11

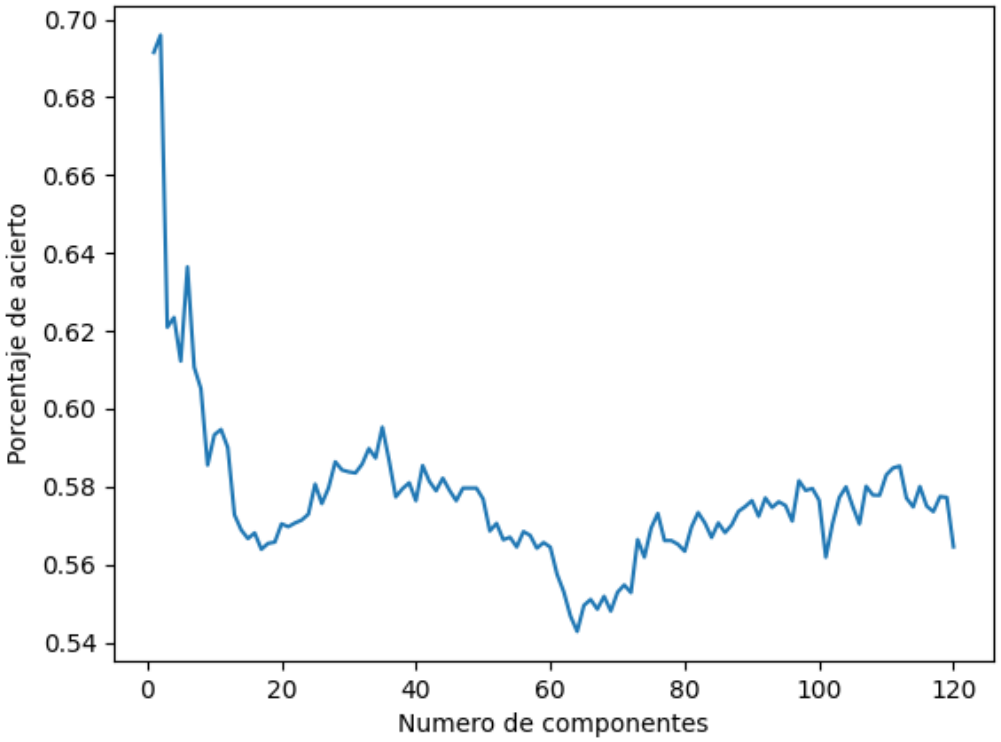

Figura 4.11 Acierto balanceado en función del número de componentes del análisis PCA..

Se observa que en este caso un número bajo de componentes principales proporciona mejor resultado. En concreto, con dos componentes se obtiene un porcentaje de acierto del 69.59 %. La matriz de confusión asociada es:

$$C_{LDA,PCA} = \begin{bmatrix} 649 & 24 & 14 \\ 57 & 20 & 26 \\ 5 & 32 & 97 \end{bmatrix}. \tag{4.23}$$

Estos resultados son, hasta este punto del trabajo, los mejores obtenidos, aunque aún se clasifican mal la mayoría de elementos en el estado de estrés hídrico 2.

4.5.5 Implementación con acceso a datos meteorológicos

La meteorología puede afectar considerablemente al comportamiento de las plantas. Determinadas condiciones climáticas pueden causar que las curvas de turgencia se parezcan más a las de un determinado tipo,

cuando en realidad la planta se encuentra en un estado distinto de estrés.

Con el fin de compensar este efecto, se dispone de los datos de temperatura ambiente, humedad relativa, radiación global, parcial y neta y déficit de presión de vapor. Se ha comprobado el efecto que tiene sobre la clasificación repitiendo la clasificación LDA de un solo día, añadiendo las curvas meteorológicas muestreadas como un conjunto de características adicional. Las curvas meteorológicas deben normalizarse en el tiempo, pero la medida se empleará con su valor medido.

De esta manera, la matriz X y el vector Y tendrán la forma:

$$X = \begin{bmatrix} LPT(0,0) & \cdots & LPT(0,M-1) & METEO(0,0,0) & \cdots & METEO(0,M-1,P) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ LPT(N-1,0) & \cdots & LPT(N-1,M-1) & METEO(N-1,0,0) & \cdots & METEO(N-1,M-1,P) \end{bmatrix}. \quad (4.24)$$

$$Y = \begin{bmatrix} Estado(0) & \cdots & Estado(N-1) \end{bmatrix}. \quad (4.25)$$

Donde $METEO(n,m,p)$ representa el dato del día n, en el instante m, para el sensor meteorológico p.

En este caso, debe aplicarse un muestreo tanto a las curvas meteorológicas como a las de turgencia, pudiendo emplearse frecuencias distintas para buscar la combinación con mejor comportamiento. De esta forma se obtiene la figura 4.12, que representa en un mapa de calor el acierto balanceado en función de ambas frecuencias.

En este caso, las frecuencias se han representado en función del periodo de muestreo en horas normalizadas, es decir, el número de horas entre muestras si se escala el eje temporal para hacer que el amanecer se produzca a las 6:00 y el atardecer a las 18:00, habiendo 12 horas normalizadas en el período diurno completo.

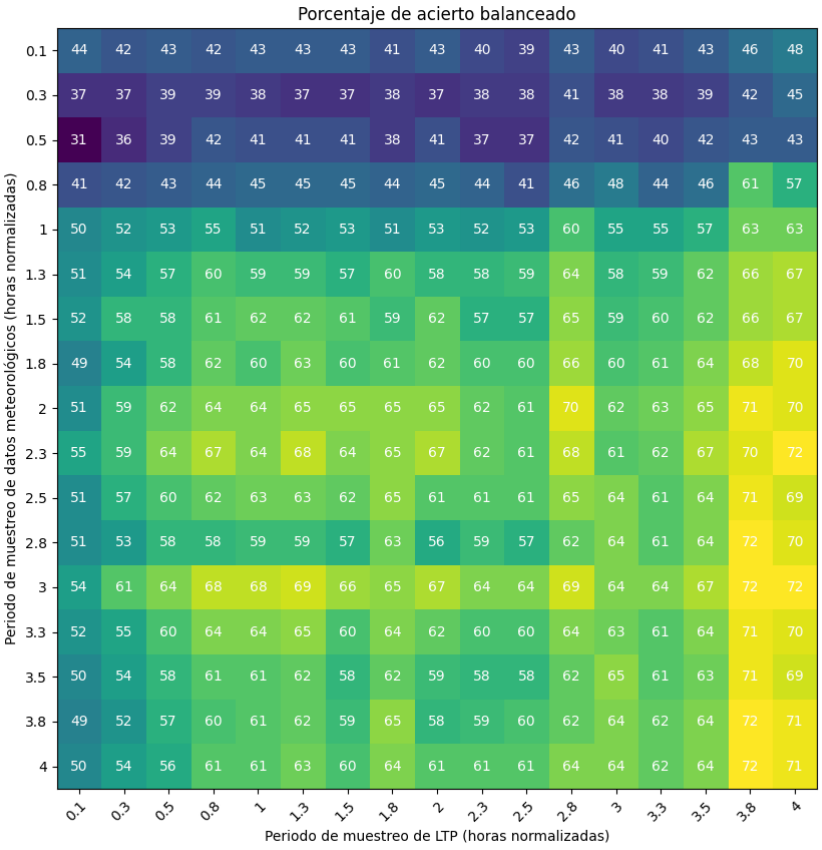

Figura 4.12 Acierto balanceado en función del periodo de muestreo, considerando datos meteorológicos..

Se observa que los resultados obtenidos son mejores que sin emplear los datos meteorológicos, incluso sin haber utilizado memoria ni preprocesado mediante PCA para mejorar el resultado, alcanzando un acierto

balanceado en torno al 71 % con el número de muestras adecuado, superando de nuevo los valores más altos alcanzados anteriormente.

En este caso, se pueden reducir las características a analizar minimizando el número de datos meteorológicos empleados, ya que existe una cierta correlación entre los mismos. Con este fin, se ha optado por realizar una prueba empleando únicamente la radiación neta media, ya que es la variable empleada para la obtención de las horas de amanecer y anochecer. Los resultados de esta prueba aparecen reflejados en la figura 4.13.

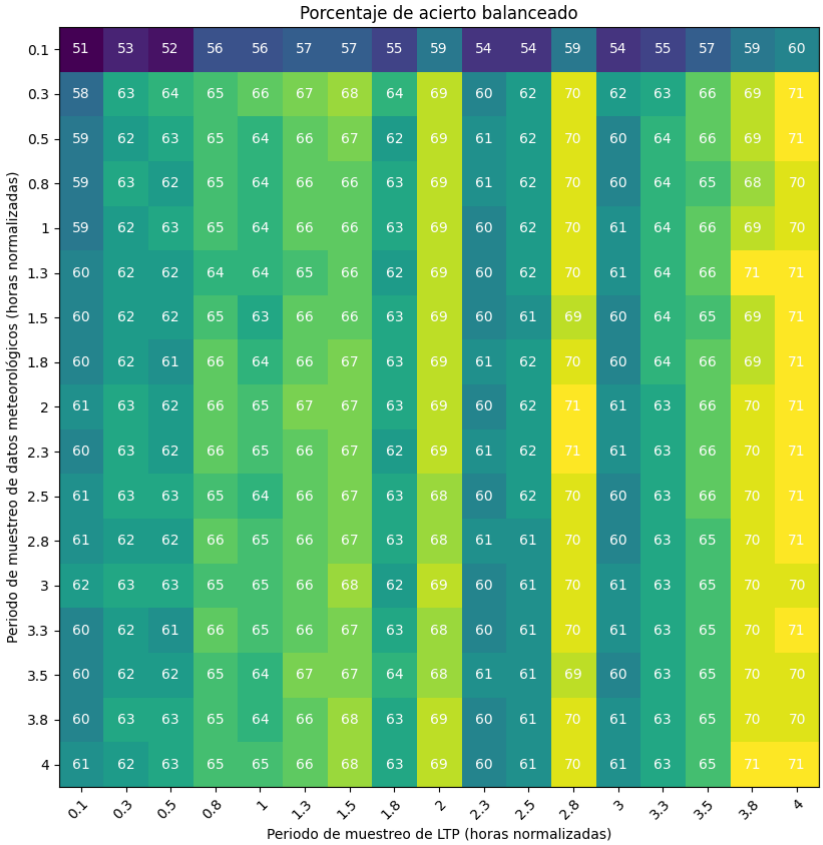

Figura 4.13 Acierto balanceado en función del periodo de muestreo, considerando la radiación neta media..

Se observa que los resultados son similares a los obtenidos al emplear el resto de datos meteorológicos, pero el entrenamiento soporta un mayor número de muestras, lo que indica que los datos adicionales no aportaban información nueva pero dificultaban el entrenamiento y aumentaban los tiempos de procesado. En concreto, utilizando ambos periodos de muestreo de 4 horas normalizadas, se ha obtenido un porcentaje de acierto balanceado del 71 % y una matriz de confusión:

$$C_{LDA,meteo} = \begin{bmatrix} 490 & 188 & 9 \\ 25 & 59 & 19 \\ 2 & 19 & 113 \end{bmatrix}. \tag{4.26}$$

Dado el valor del periodo de muestreo, en total se están empleando 6 características por muestra, correspondientes a las lecturas al amanecer, al atardecer y a mediodía del sensor LTP y de la radiación neta.

4.5.6 Implementación con preprocesado mediante PCA y acceso a datos meteorológicos

Combinando las técnicas anteriores puede obtenerse una mejora de los resultados obtenidos por cada una de ellas de forma independiente.

De esta forma, se ha implementado un clasificador con acceso a datos meteorológicos y preprocesado mediante PCA de todo el conjunto de características, que accede a datos actuales y de días anteriores.

Por otro lado, si sólo se proporciona la información de la radiación neta media, el comportamiento es muy parecido al conseguido por el clasificador LDA con acceso a la radiación, con un 71 % de acierto balanceado. Si se añade la temperatura y la humedad, el porcentaje de acierto se incrementa en un 1 %, alcanzando un 72 % de acierto balanceado. Añadir además el déficit de presión de vapor permite obtener cerca de un 73 % de acierto balanceado, el mejor resultado de entre todos los clasificadores probados en esta sección.

Añadir las radiaciones parcial y global empeora el resultado, probablemente por no aportar información que no se haya incluido ya con la radiación neta pero añadiendo características al entrenamiento.

La matriz de confusión correspondiente al mejor resultado, correspondiente al clasificador LDA con preprocesado basado en PCA limitado a las 4 componentes principales, y teniendo en cuenta los datos de humedad, temperatura, radiación neta media y déficit de presión de vapor con 3 muestras diarias de cada variable, es la siguiente:

$$C_{LDA,PCA,meteo,n} = \begin{bmatrix} 643 & 35 & 9 \\ 41 & 42 & 20 \\ 2 & 19 & 113 \end{bmatrix}. \tag{4.27}$$

Comparando esta matriz con la del clasificador basado en el error cuadrático respecto a curvas tipo de la ecuación 4.7, se observa que los estados 1 y 2 siguen siendo los menos distinguibles entre sí, pero en este caso el estado 2 es el que peor se clasifica. En términos globales, el acierto del clasificador se ha incrementado notablemente.

Si se representan los valores reales y predichos de las muestras analizadas, ordenadas por sensor y fecha, se obtiene la gráfica de la figura 4.14. También puede resultar de interés la figura 4.15, en la que se analiza el error cometido; un valor de error negativo representa que la planta se encuentra mejor hidratada que lo predicho.

Se observa que la mayoría de errores son puntuales, y tienden a aparecer en periodos en los que el nivel de estrés cambia con frecuencia o rápidamente, lo que dificulta obtener lecturas fiables.

Se ha comprobado que el acceso a datos de días anteriores sigue sin aportar una variación apreciable del porcentaje de acierto balanceado independientemente del número de días almacenados en memoria, al obtener exactamente los mismos resultados.

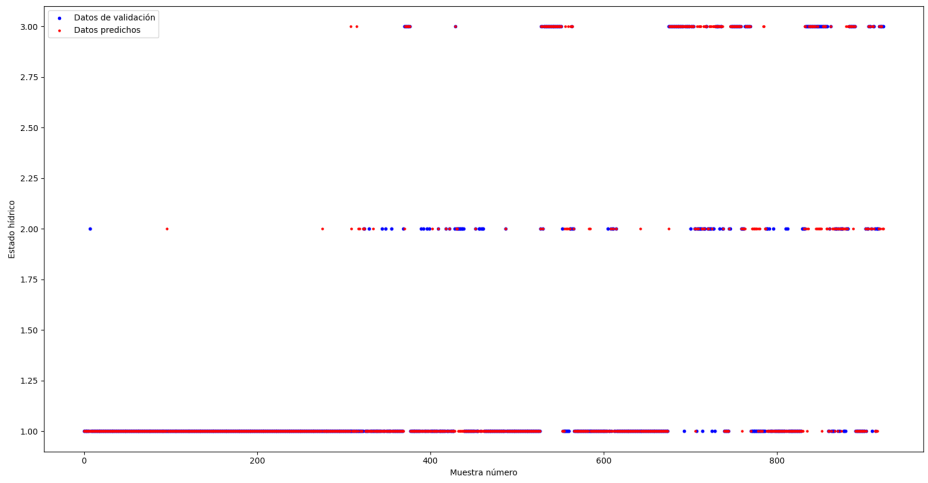

Figura 4.14 Estrés predicho frente a estrés real..

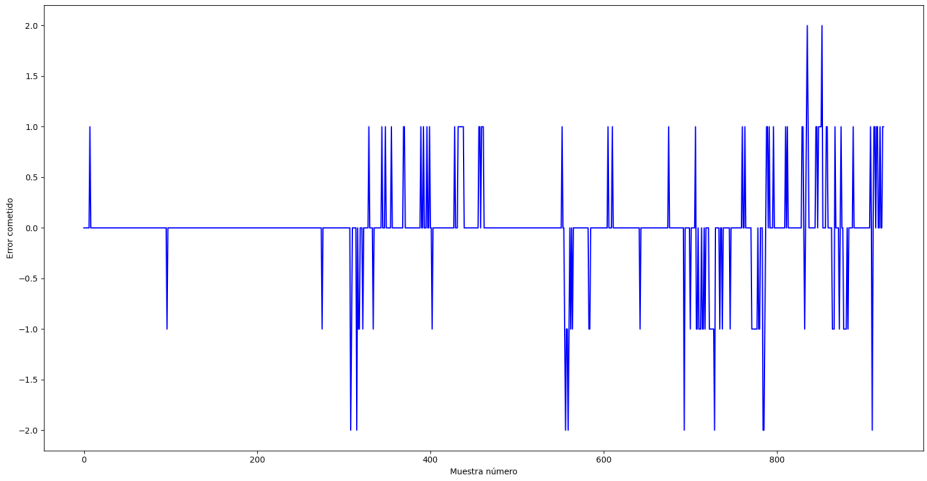

Figura 4.15 Error de predicción..

4.5.7 Otras lineas de investigación abiertas

Durante la elaboración del proyecto se han probado otros métodos de clasificación que no han dado buenos resultados a priori, pero que puede resultar de interés comprobar más adelante.

Por ejemplo, se ha comprobado como se comporta el clasificador si en lugar de la curva completa se hubiesen utilizado características basadas en los estimadores que se propusieron al comienzo del proyecto, como pueden ser el valor máximo y el valor mínimo, las horas a las que se alcanzan y el valor medio, buscando aportar la información relevante seleccionando estas características manualmente.

En este caso, empleando las características anteriormente mencionadas tanto para la curva LTP como para las de datos meteorológicos, y empleando el clasificador de la sección 4.5.6, se ha conseguido un porcentaje de acierto normalizado de sólo el 57 %, lo que indica que la pérdida de información es mayor que la ventaja obtenida al reducir el número de características de esta forma. El hecho de que una frecuencia de muestreo baja, de 3 o 4 muestras por día, proporcione un mejor resultado parece indicar que es más importante en este caso disponer de muestras espaciadas en el tiempo que analizar los valores extremos de la curva, aunque podrían existir otras causas.

También se ha comprobado la efectividad de añadir la curva correspondiente a la derivada de las lecturas LTP al clasificador de la sección 4.5.6, medida como el incremento entre muestras consecutivas. En este caso, el porcentaje de acierto ha sido del 70.6 %, peor que con el clasificador original, ya que los nuevos datos no aportan suficiente información relevante y requieren más datos de entrenamiento para obtener el mismo resultado. Si los datos nuevos realmente pueden ayudar a mejorar la clasificación, es necesario emplear un mayor número de datos de entrenamiento, lo que no es posible hasta haber recogido nuevos datos de los sensores instalados.

Por último, queda probar otros métodos de clasificación. Un clasificador que se probó inicialmente pero fue descartado es el clasificador bayesiano ingenuo. Este clasificador es muy empleado por su sencillez y por su robustez ante el sobreentrenamiento, pero requiere un gran número de datos de entrenamiento y las características deben ser discretas, además de asumir independencia estadística entre características, lo que no es aplicable realmente a los puntos de las curvas LTP. Es posible aplicarlo, no obstante, si se discretizan los valores leídos, lo que conlleva una cierta pérdida de información. En cualquier caso, el mejor resultado obtenido con este clasificador fue de un 44 % de acierto balanceado, resultando llamativo que ninguna de las muestras fuese clasificada en el estado 2, a pesar de haber empleado una versión del clasificador con soporte multiclase, es decir, que admite tres o más clases de clasificación.

5 Conclusiones generales del trabajo realizado

El uso de técnicas de aprendizaje automático para la interpretación de las lecturas de sensores ZIM con el fin de conocer el estado hídrico de las plantas puede resultar de gran utilidad en un futuro cercano para el desarrollo de sistemas de riego más eficientes y automatizables.

Este trabajo refleja la viabilidad del proyecto, así como un primer paso hacia la consecución de sus objetivos, pero los resultados obtenidos, aunque útiles y prometedores por sí mismos, aún parecen tener margen de mejora.

También se ha comprobado la eficacia del análisis discriminante lineal y del análisis de componentes principales como métodos probabilísticos para la clasificación de lecturas con un gran número de variables involucradas entre las que se dan relaciones demasiado complejas para otros tipos de análisis, como es frecuente en sistemas biológicos.

Queda pendiente para trabajos posteriores el análisis de nuevas formas de clasificación, así como refinar donde sea posible los clasificadores empleados, con el fin de obtener un clasificador con un mejor desempeño. Otra posible vía de mejora podría consistir en encontrar una forma de clasificación basada no en tres categorías separadas, como la empleada en este trabajo, si no en un espectro continuo, donde se podría considerar aceptable un cierto margen de error. Esta clasificación continua requeriría un análisis aun más detallado del estado hídrico de las plantas, pero, de ser viable, proporcionaría una información más precisa, lo que a su vez permitiría un control del riego aun más eficiente.

Otra manera de mejorar el sistema de clasificación consistiría en entrenarlo con datos de años distintos. Habiéndose entrenado con datos de un único año y comprobando la clasificación en un año distinto, con 5 años de diferencia, se ha podido comprobar que el clasificador es relativamente robusto ante los cambios climáticos y los ocurridos en las plantas del cultivo, pero disponer de un año adicional permitiría introducir en el set de entrenamiento diferencias interanuales que no se correspondiesen con las del set de validación, pudiendo comprobar la mejora asociada.

Referencias

[1] Kevin Dunn. «Process Improvement Using Data». En: *https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/index* (visitado el 7/8/2022). https://learnche.org/pid/contents, 2022. Cap. 6.5. Principal Component Analysis (PCA).

[2] J.E. Fernández. «Plant-based sensing to monitor water stress: Applicability to commercial orchards». En: *Agricultural Water Management* 142 (2014), págs. 99-109. ISSN: 0378-3774. DOI: *https://doi.org/10.1016/j.agwat.2014.04.017*. URL: *https://www.sciencedirect.com/science/article/pii/S037837741400136X*.

[3] M Govender et al. «Review of commonly used remote sensing and ground-based technologies to measure plant water stress». En: *Water Sa* 35.5 (2009).

[4] José Enrique Fernández Luque et al. *Estrategias y programación del riego: manual elaborado en el marco del Hito 1.3.2: Bases tecnológicas de riego deficitario*. 2015.

[5] EM Martínez et al. «Comparison of two techniques for measuring leaf water potential in Vitis vinifera var. Albariño». En: *Ciência e Técnica Vitivinícola* 28 (oct. de 2013), págs. 29-41.

[6] CM Padilla-Díaz et al. «Scheduling regulated deficit irrigation in a hedgerow olive orchard from leaf turgor pressure related measurements». En: *Agricultural Water Management* 164 (2016), págs. 28-37.

[7] F. Pedregosa et al. «Scikit-learn: Machine Learning in Python». En: *Journal of Machine Learning Research* 12 (2011), págs. 2825-2830.

[8] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. 2022. URL: *https://scikit-learn.org/stable/index.html* (visitado 01-08-2022).

A Fundamentos teóricos del clasificador LDA

Según la documentación de la librería scikit-learn[8], el análisis discriminante lineal (LDA) y el análisis discriminante cuadrático (QDA) pueden derivarse de un modelo probabilístico simple que modele la distribución condicional de cada dato $P(X|y=k)$ para cada clase k. A partir de este modelo, puede aplicarse la regla de Bayes para obtener la probabilidad posterior de cada clase en cada muestra, según la ecuación A.1.

$$P(y=k|x) = \frac{P(x|y=k)P(y=k)}{P(x)} = \frac{P(x|y=k)P(y=k)}{\sum_l P(x|y=l)\cdot P(y=l)}. \tag{A.1}$$

Es decir, la probabilidad de que una muestra y pertenezca a la clase k dadas unas características x equivale a la probabilidad de que se den unas características x cuando y pertenece a la clase k, multiplicada por la probabilidad de que y pertenezca a la clase k y dividida por la probabilidad de que se den las características x. La muestra se asumirá como perteneciente a la clase cuya probabilidad posterior sea mayor.

A su vez, la probabilidad $P(x|y=k)$ se modela como una distribución gaussiana multivariante, cuya densidad se expresa en la ecuación A.2.

$$P(x|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)\right). \tag{A.2}$$

Donde d es el número de características de cada muestra ($x \in \mathbb{R}^d$), μ_k es el valor medio de las características de la clase k y Σ_k es la matriz de covarianza de las características de la clase k.

Estas ecuaciones permiten implementar el análisis QDA. El análisis LDA es un caso particular en el que se asume que las funciones gaussianas de cada clase comparten las mismas matrices de covarianza: $\Sigma_k = \Sigma$ para cualquier valor de k.

Se puede simplificar el cálculo a partir del logaritmo de la probabilidad posterior y separando los términos constantes. En este caso, la muestra pertenecerá a la clase que devuelva el logaritmo de mayor valor.

$$lnP(y=k|x) = lnP(x|y=k) + lnP(y=k) - lnP(x). \tag{A.3}$$

$$lnP(y=k|x) = ln\left((2\pi)^{d/2}|\Sigma|^{1/2}\right) - \frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k) + lnP(y=k) - lnP(x). \tag{A.4}$$

Dado que el objetivo es buscar la clase que maximiza el logaritmo, puede omitirse el cálculo de los términos constantes *Cst* de la ecuación A.5, disminuyendo el tiempo de cálculo.

$$lnP(y=k|x) = -\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k) + lnP(y=k) + Cst. \tag{A.5}$$

Donde $(x-\mu_k)'\Sigma^{-1}(x-\mu_k)$ representa la distancia de Mahalanobis entre la muestra con características x y la media μ_k. Esta distancia mide la cercanía entre x y μ_k teniendo en cuenta la varianza de cada característica; puede considerarse que el LDA asigna cada muestra a la clase cuya media es más cercana, teniendo en cuenta también las probabilidades de partida de cada clase.

La ecuación A.5 también puede expresarse en función de una superficie de decisión lineal, como se refleja en la ecuación A.6, donde $w_k = \Sigma^{-1}\mu_k$ y $w_{k0} = -\frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + logP(y=k)$.

$$lnP(y=k|x) = w_k'x + w_{k0} + Cst. \tag{A.6}$$

Otra forma de interpretar el análisis, que permite visualizar mejor su uso para reducir la dimensionalidad del problema, consiste en separar el proceso en una fase de "esferificación", en la que la matriz de covarianza se convierte en la matriz identidad al aparecer dividiendo en la ecuación, y una segunda fase de proyección, en la que la distancia de Mahalanobis se ha convertido en una distancia Euclídea en un espacio d-dimensional con una dimensión por cada característica.

En esta segunda fase, el cálculo de la distancia Euclídea puede realizarse proyectando el espacio sobre un subespacio de dimensión $d-1$, obteniendo así un nuevo conjunto de características. Aplicando este método de forma recursiva, puede obtenerse un conjunto de características con la dimensión deseada o bien calcular la distancia empleada por el clasificador.

B Fundamentos teóricos de la descomposición PCA

El Análisis de Componentes Principales (PCA) es una técnica que se basa en descomponer un conjunto de datos multivariables en un conjunto de componentes ortogonales, de dimensión menor al número de variables originales, que expliquen la máxima cantidad de varianza. Para ello, se suele utilizar la Descomposición en Valores Singulares (SVD).

De acuerdo con "Process Improvement Using Data", de Kevin Dunn [1], el proceso es más fácil de comprender cuando se explica desde un punto de vista geométrico. En este sentido, se va a considerar unos datos de partida tridimensionales, con tres características por muestra X, x_1, x_2 y x_3, como los representados en la figura B.1. Todo lo que se explica a continuación es extrapolable a cualquier número de dimensiones K.

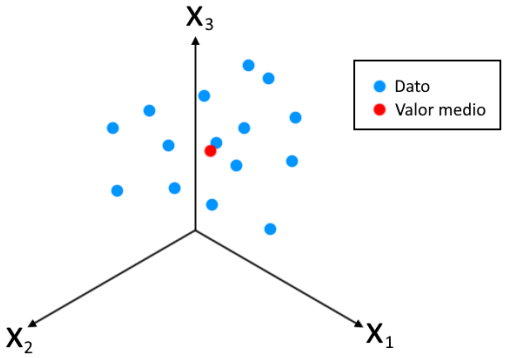

Figura B.1 Datos representados en el espacio tridimensional..

El primer paso al aplicar PCA consiste en desplazar el punto medio al origen, restando los valores medios a cada una de las muestras, como se ha representado en la figura B.2. También es habitual realizar un escalado para obtener varianza unidad en cada característica, aunque este paso es opcional al emplear la librería scikit-learn [8]. La necesidad del escalado dependerá del tratamiento que se pretenda utilizar tras aplicar PCA y de las características de los datos empleados.

La recta de mejor ajuste coincide con la dirección en la que la varianza de la nube de puntos es mayor. Esta recta también proporciona la primera componente principal, también conocida como la primera variable latente, que se compone de dos partes: el vector de dirección p_1 y el vector de puntuación t_1, como se ha representado en la figura B.3.

El vector de dirección p_1 tiene magnitud 1, dimensión K, siendo K el número de características por muestra, y coincide con la dirección de la recta de mejor ajuste. En el caso del ejemplo, $K = 3$.

El vector de puntuación t_1 tiene dimensión N, siendo N el número de muestras en la nube de puntos, y contiene la distancia entre el origen y la proyección ortogonal de cada punto sobre la recta, coincidiendo con el valor de la primera dimensión reducida para cada muestra. En el caso del ejemplo, $N = 15$.

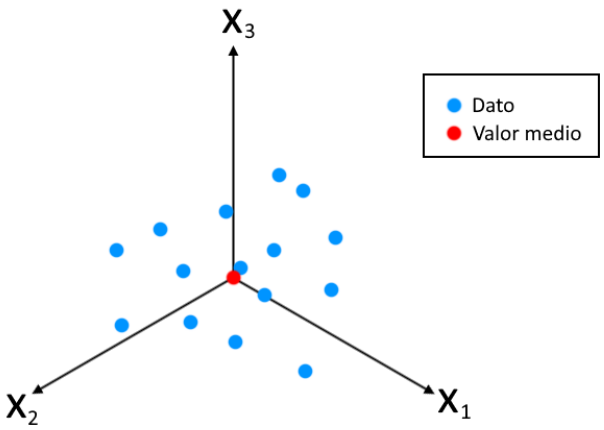

Figura B.2 Datos representados en el espacio tridimensional, con media nula..

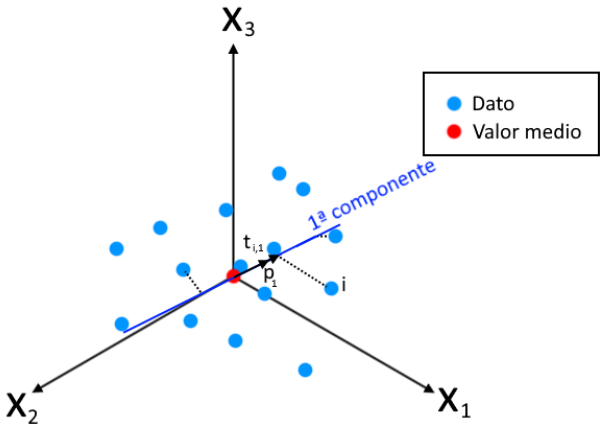

Figura B.3 Primera componente..

Una vez hallada la primera componente, la segunda componente es perpendicular a la primera. Para calcularla geométricamente, pueden proyectarse los datos sobre un subespacio de dimensión $K-1$ perpendicular a p_1 que pase por el origen, para a continuación calcular la recta de mejor ajuste dentro de ese subespacio.

Una vez obtenida la segunda recta de mejor ajuste, se puede regresar al espacio original para obtener p_2 y t_2 de forma análoga a p_1 y t_1.

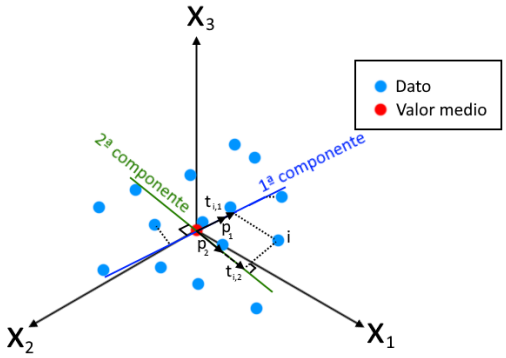

Figura B.4 Segunda componente..

Este proceso puede repetirse iterativamente para obtener las componentes necesarias, hasta un máximo

de $K-1$. El conjunto de vectores de dirección p representa entonces un hiperplano, que es el modelo de variables latentes buscado, y coincide con la mejor aproximación que puede realizarse de los datos originales para el número de componentes elegidas. La distancia entre cada punto y el hiperplano, medida de forma ortogonal a este último, es el error residual.

En el ejemplo, al calcular la segunda componente, se ha alcanzado el número máximo de componentes, quedando el modelo de la figura B.5.

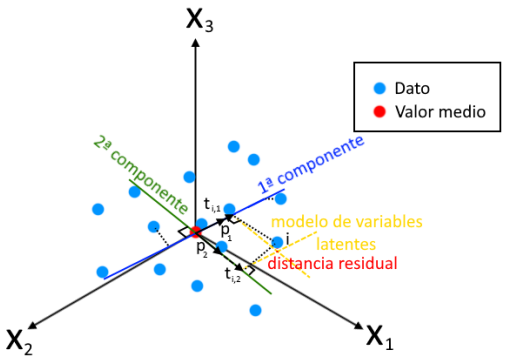

Figura B.5 Modelo de variables latentes..

El modelo puede calcularse a partir de un conjunto de datos de entrenamiento iniciales y luego aplicarse a nuevos conjuntos de datos.

Si se llama X a una matriz de datos, donde cada fila representa una muestra, $x_{i,k}$ al valor de la dimensión k de la muestra i, $t_{i,j}$ es la puntuación en la componente j-ésima de la muestra i y $p_{k,j}$ es el valor de la dimensión k del vector de dirección de la componente j-ésima, por geometría, se tiene que:

$$t_{i,j} = x_{i,1}p_{1,j} + x_{i,2}p_{2,j} + \ldots + x_{i,k}p_{k,j} + \ldots + x_{i,K}p_{K,j} = \sum_{k=K}^{1} x_{i,k}p_{k,j}. \tag{B.1}$$

Agrupando los términos anteriores en matrices, pueden calcularse todas las puntuaciones para una muestra dada:

$$t_i' = x_i'P. \tag{B.2}$$

De igual forma, si se agrupan todas las muestras en la matriz X, se pueden calcular todas las puntuaciones para todas las muestras de forma matricial:

$$T = XP. \tag{B.3}$$

Este método, a pesar de ser sencillo de comprender e interpretar, resulta muy lento cuando los datos de entrada son numerosos y de dimensión elevada, como suele ocurrir. Puede obtenerse un método más rápido operando matemáticamente sobre las definiciones anteriores.

Aplicando la descomposición en valores singulares (SVD) sobre X se obtiene su expresión en otras tres matrices:

$$X = U\Sigma V'. \tag{B.4}$$

Donde U y V son ortonormales, es decir, cada columna suma la unidad y es ortogonal al resto, mientras que Σ es una matriz diagonal. Despejando en la ecuación B.3 se obtiene:

$$X = TP'. \tag{B.5}$$

Donde P también es ortonormal. Se puede demostrar que, si se preprocesan los datos para obtener X anulando la media y ajustando la varianza, $P = V$ y $T = U\Sigma$ proporciona el modelo PCA con todas las componentes posibles, aunque no puede aplicarse si faltan datos en alguna de las muestras.

Dado que existen métodos muy eficientes para el cálculo de la SVD y que si se desea un menor número de componentes basta con seleccionarlas en orden decreciente de varianza de T, el método de cálculo mediante

SVD es uno de los más aplicados en la actualidad, y el implementado por defecto en la librería scikit-learn [8].