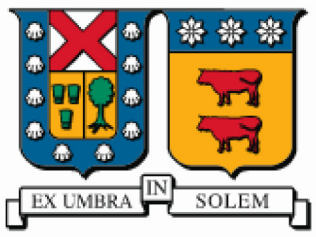


UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Predicción de días de tormenta dentro del territorio chileno mediante el uso de técnicas de Inteligencia Artificial

Autor:
Sergio Rosales Baros

Profesores patrocinadores:
Dr. Johny Montaña Chaparro
Dr. Carlos Valle Vidal

Correferente:
Dra. Diana Pozo Labrada

Memoria de Titulación para optar al Título de Ingeniero Electricista

en el

Departamento de Ingeniería Eléctrica
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Valparaíso, Chile

2 de agosto de 2023

« *Lo que sabemos es una gota de agua, lo que ignoramos es el océano.* »

— Isaac Newton

Para mi familia y amigos . . .

Agradecimientos

A mis padres, por brindarme su amor y aliento. A mis amigos, por el apoyo y tiempo compartido. A mis profesores, por sus conocimientos y dedicación. Y a las instituciones Dirección Meteorológica de Chile y World Wide Lightning Location Network por el acceso a la información.

Resumen

El presente documento es el informe de Memoria de Titulación presentado en cumplimiento parcial de los requisitos para optar al Título de Ingeniero Electricista en la Universidad Técnica Federico Santa María (UTFSM). Este informe, contiene un modelo de predicción de descargas atmosféricas para un lugar preestablecido dentro del territorio chileno mediante el uso de técnicas de Inteligencia Artificial (IA).

El estudio de los rayos o descargas eléctricas atmosféricas es un tema de interés mundial, en tanto, la predicción de este fenómeno natural es un problema multidisciplinario que afecta a varios sectores, ya sea en ámbitos sociales (como un complemento al Sistema de Alerta de Emergencia (SAE)) o económicos (actividades ligadas a faenas mineras, actividades deportivas, industria aeronáutica, etc.) así como también permite ampliar la frontera del saber y dar pie a futuros desarrollos o investigaciones.

En esta memoria, se presenta el diseño de un modelo que permite la predicción de días de tormenta eléctrica atmosférica. En el alcance de este estudio se considera una localidad dentro Chile, con la restricción de que sea una ubicación que haya registrado una alta cantidad de descargas atmosféricas y cuente con los suficientes registros meteorológicos entre los años 2012 y 2021 que permita entrenar una máquina de aprendizaje. La localidad escogida fue Visviri, pueblo ubicado en la comuna de General Lagos, región de Arica y Parinacota. Se empleó una metodología que consistió en crear un conjunto de datos, incluyendo un Análisis Exploratorio de Datos (EDA), la imputación de datos no disponibles, selección de atributos (Feature Engineer), reducción de la dimensional (Feature Selection), búsqueda de hiper-parámetros y un análisis de sensibilidad para el modelo que presente el mejor rendimiento mediante la puntuación F1.

El modelo realizado consistió en una red multicapa con función de activación ReLu y dropout que alcanzó un desempeño del 74.68 % en la puntuación F1 para el año 2021.

Índice general

Índice de tablas

Índice de figuras

Siglas y Acrónimos

IC Inter-cloud

CC Cloud-to-cloud

CG Cloud-to-ground

CA Cloud-to-air

DMC Dirección Meteorológica de Chile

WWLLN World Wide Lightning Location Network

IA Inteligencia Artificial

RNA Red Neuronal Artificial

MLP Multilayer Perceptron

ML Machine Learning

EDA Análisis Exploratorio de Datos

SAE Sistema de Alerta de Emergencia

UTFSM Universidad Técnica Federico Santa María

WRF Weather Research and Forecasting

HR Humedad relativa

RRR Precipitación

QFE Presión del aire a nivel de la estación

QFF Presión del aire a nivel del mar

QNH Presión atmosférica a nivel del mar mediante Atmósfera Estándar de la OACI

Td Temperatura del punto de rocío

Ts Temperatura del aire seco

ff Velocidad del viento

dd Dirección del viento

1 Introducción

Índice general

1.1 Contexto

El estudio de los rayos o descargas eléctricas atmosféricas de gran intensidad producidas durante una tormenta es un tema que ha captado el interés de varios investigadores a nivel mundial. En este contexto, la predicción de este fenómeno natural es un problema multidisciplinario que afecta a varios sectores, ya sea en ámbitos sociales (como un complemento al SAE) o económicos (actividades ligadas a faenas mineras, actividades deportivas, industria aeronáutica, etc.), así como también permite ampliar la frontera del saber y dar pie a futuros desarrollos o investigaciones. Si bien este problema se ha planteado desde un enfoque tradicional mediante el análisis del modelo físico, una alternativa que ha cobrado relevancia en los últimos años debido al avance computacional es el uso de máquinas de aprendizaje automático usando técnicas de IA. Estas máquinas son capaces de responder una problemática sin tener que modelar el fenómeno físico. Es cuestión de esta memoria diseñar un modelo que, al recibir un conjunto de datos determine la ocurrencia o no ocurrencia de días de tormenta en una ubicación particular dentro del territorio chileno.

1.2 Motivación

En la actualidad, no existe un modelo que pronostique con una capacidad de generalización[1] perfecta la ocurrencia de días de tormenta a nivel nacional ni mundial. En este trabajo se desarrolla un modelo para el territorio chileno con el propósito de que sea una herramienta de apoyo y ayude en la toma de decisiones para levantar las alertas y precauciones para evitar la pérdida de vidas humanas y equipos eléctricos. Existen algunos modelos que han sido desarrollados, como es el del caso del algoritmo basado en árboles de decisión para Mashhad, Irán que alcanzó un rendimiento del 86.8 % en la puntuación F1. Mientras que, en el ámbito nacional el esquema de predicción consiste en modelar las ecuaciones de flujo y observar la

[1] Véase 2.4.1.

carta sinóptica considerando 3 condiciones: inestabilidad atmosférica, humedad y presencia de nubes de 10 km de altura. Además, existe un sinnúmero de sensores y redes interconectadas que registran los rayos y los días de tormenta en todo el planeta. Una de las redes se encuentra ubicada en las dependencias de la UTFSM. Esto ha permitido caracterizar información de la ubicación de las descargas con una resolución de 10 km en todo Chile continental desde el año 2012.

1.3 Bibliografía relevante

Realizando una revisión del estado de la cuestión, se encontró que ya se han hecho estudios en otros países. Como el sistema de alertas de bajo costo desarrollado en el laboratorio de la Universidad de Moratuwa, Sri Lanka (Jayendra y cols., 2007). Este modelo consistió en un perceptrón que fue alimentado con 2 entradas (las señales de radiofrecuencia emitidas por los rayos dentro de la nube (Inter-cloud (IC), por sus siglas en inglés) y el campo eléctrico estático medido con un molino de campo eléctrico fabricado artesanalmente). Su salida determinó el grado de amenaza (alto, medio o nulo). El modelo completamente entrenado obtuvo resultados satisfactorios. Notar que el estudio no precisó más información como la función de activación del perceptrón.

En Malasia se han desarrollado varios modelos, uno de ellos fue una red neuronal desarrollada para la ciudad de Subang Jaya (Johari, Rahman, y Musirin, 2007). Esta red recibió 24 variables de entrada (8 variables meteorológicas, tales como: viento, punto de rocío, humedad, presión, entre otras y 1 indicador para cada mes y estación del año), 2 capas ocultas (8 neuronas en la primera capa y 5 en la segunda) con función de activación logaritmo-sigmoidea y una salida con función lineal que indica la ocurrencia o no ocurrencia de rayos. Para entrenar se usó el algoritmo de Levenberg-Marquardt, una tasa de aprendizaje de 0.4819 y una constante de momento de 0.0577. Se usaron 378 datos para entrenar y 197 datos para probar la red. El error de entrenamiento fue calculado mediante el error RMS (obteniendo un error del 0.41 %). El modelo completamente entrenado obtuvo un desempeño del 99.997 % en la correlación entre el valor esperado y el valor de la salida de la red. Notar que no se proporcionó información acerca de la ventana de tiempo con la que fueron medidos los datos ni el porcentaje de desequilibrio de la variable objetivo.

Otro estudio malayo consistió en un Perceptrón multicapa (Multilayer Perceptron (MLP), por sus siglas en inglés) desarrollado para el Aeropuerto Internacional de Kuala Lumpur (Ramzi, Adnan, Samad, y Ruslan, 2018). Esta red recibió 5 variables de entrada (Temperatura del aire seco (Ts), Humedad relativa (HR), Presión del aire a nivel del mar (QFF), Velocidad del viento (ff) y Precipitación (RRR)), 1 capa oculta (35 neuronas) con función de activación tangente hiperbólica sigmoidea y una salida con función lineal que indica el número de rayos ocurridos. Para entrenar se usó el algoritmo de Levenberg-Marquardt, una tasa de aprendizaje de 0.08 y una constante de momento de 0.95. Se usaron 288 datos para entrenar (desde enero-2010 hasta diciembre-2013) y 72 datos para probar la red (desde enero-2015 hasta diciembre-2015). Todos los datos fueron obtenidos del Departamento Meteorológico de Malasia. El error de entrenamiento fue calculado mediante el error RMS (obteniendo un error del 0.0786 %). La correlación entre el valor esperado y el valor de salida de la red resultó ser de un 99.999 %. El modelo completamente entrenado obtuvo un desempeño del 94.64 % en la correlación entre el valor esperado y el valor de la salida de la red. Notar que no se proporcionó información acerca del porcentaje de desequilibrio de la variable objetivo.

Otro estudio malayo consistió en otro MLP desarrollado para el aeropuerto Sultan Abdul Aziz Shah ubicado en la ciudad de Subang (Abdullah, Adnan, Samad, y Ahmat Ruslan, 2018). Esta red recibió 5 variables de entrada (temperatura, HR, QFF, ff y precipitación), 1 capa oculta y una salida que indica el número de rayos ocurridos. Se usaron 360 datos para entrenar (desde enero-2010 hasta diciembre-2014) y 72 datos para probar la red (desde enero-2015

hasta diciembre-2015). Todos los datos fueron obtenidos del Departamento Meteorológico de Malasia. El error de entrenamiento fue calculado mediante el error RMS (obteniendo un error del 11.05 %). La correlación entre el valor esperado y el valor de salida de la red resultó ser de un 99.990 %. El modelo completamente entrenado obtuvo un desempeño del 98.718 % en la correlación entre el valor esperado y el valor de la salida de la red. Notar que no se proporcionó información acerca de los resultados de la matriz de confusión.

Un estudio diferente, consistió en un modelo de aprendizaje automático (Machine Learning (ML), por sus siglas en inglés) desarrollado para 12 ubicaciones de Suiza, como las montañas de Säntis y Monte San Salvatore (Mostajabi, Finney, Rubinstein, y Rachidi, 2019). Este modelo recibió 4 variables de entrada meteorológicas (Presión del aire a nivel de la estación (QFE), Temperatura del aire a 2 m sobre el suelo, HR y ff), y una salida asociada a la ocurrencia del rayo. Se usaron los datos registrados cada 10 minutos desde 2006 hasta 2017. Este modelo de ML se contrastó con un método de pronóstico por persistencia, un modelo basado en el método de campo electrostático y un esquema basado en el umbral de la Energía potencial de convección disponible (CAPE, por sus siglas en inglés) resultando en que el modelo de ML obtuvo los mejores resultados. Los resultados del modelo completamente entrenado los puede observar en la tabla 1.1. Notar que no se proporcionó más información sobre la arquitectura de la red, la fuente de los datos meteorológicos, el conjunto de entrenamiento, compilación, optimizadores, ni tampoco sobre la función de pérdida.

	POD	FAR	CSI	HSS
Säntis	71 %	9 %	67 %	80 %
Monte San Salvatore	81 %	3 %	81 %	90 %

Tabla 1.1: Resultados del modelo suizo a un horizonte de tiempo de 10 minutos

Donde[2]:

- POD: Probabilidad de detección
- FAR: Razón de falsas alarmas
- CSI: Índice crítico de éxito
- HSS: Puntuación de habilidad de Heidke

Otro estudio, consistió en un ensamblado entre redes neuronales y árboles de decisión desarrollado para las ciudades de Mashhad, Neyshābūr y Qūchān en la provincia de Jorasán Razaví, Irán (Pakdaman, Naghab, Khazanedari, Malbousi, y Falamarzi, 2020). Estos datos se caracterizan por estar desequilibrados, el cual se abordó mediante undersampling. Esta red recibió variables de entrada temporales (año, mes, día y hora) y meteorológicas (visibilidad, nubosidad, Dirección del viento (dd), ff, temperatura de aire seco, temperatura de rocío, QFE, QFF, precipitación, temperatura ambiente, humedad, nubosidad baja, tipo de nubes bajas, entre otras), 1 capa oculta con función de activación sigmoidea en el caso de la red neuronal y una salida con función hard-limit que indica la presencia de rayos. Se usaron los datos desde 1992 hasta 2018 en los que aleatoriamente se entrenó con el 85 % (70 % entrenamiento + 15 % validación) y se probó la red con el 15 % restante. Todos los datos fueron obtenidos de la Organización Meteorológica de Irán con una ventana de tiempo de 3 horas. Según los resultados obtenidos, el árbol de decisión superó a las redes neuronales en todos los conjuntos de datos. El mejor modelo completamente entrenado obtuvo un desempeño del 86.8 % en la puntuación F1 para el conjunto de datos de Mashhad, un 85.6 % para Neyshābūr y un 85.6 % para Qūchān. Notar que no se proporcionó más información sobre la compilación de la red, ni los optimizadores, ni la función de pérdida.

[2] Véase 2.4.6.

1.4 Presentación del tema

En esta memoria se persigue como principal objetivo proponer el diseño de un modelo que permita la predicción de días de tormenta eléctrica atmosférica en lugares preestablecidos dentro del territorio chileno mediante el uso de técnicas de IA, abordando tanto el problema de los datos perdidos como el de los datos desequilibrados. En el alcance de este estudio se considera una localidad dentro de Chile, con la restricción de que sea una ubicación que haya registrado una alta cantidad cantidad de descargas atmosféricas y cuente con los suficientes registros meteorológicos entre los años 2012 y 2021 que permita entrenar una máquina de aprendizaje. Los objetivos específicos son los siguientes:

1. Obtener y analizar datos de estaciones meteorológicas y datos de la red World Wide Lightning Location Network (WWLLN) del territorio chileno entre los años 2012 y 2021 mediante consulta en sitios especializados.

2. Seleccionar variables meteorológicas mediante técnicas de procesamiento de información (Feature Engineer y Feature Selection) con objeto de la creación de una base de datos que alimente el modelo de predicción.

3. Contrastar literatura sobre el uso de Inteligencia Artificial en la predicción de rayos, para recopilar información con el fin de desarrollar un modelo que mejore la predicción de días de tormenta en la métrica F1, mediante comparación de indicadores del desempeño.

4. Diseñar y entrenar varias arquitecturas de redes neuronales profundas, seleccionando una por medio de búsqueda de hiper-parámetros validándola con el pronóstico del año 2021.

1.5 Presentación del documento

Esta memoria consta de 5 capítulos, en los cuales se desarrolla un marco teórico, se presenta la metodología empleada, los resultados y análisis y se finaliza con conclusiones y recomendaciones. Además, en el anexo se pueden encontrar toda las figuras que son complementarias a este trabajo.

En el Capítulo 2 se incluye una revisión del estado del arte, abordando la formación de las tormentas eléctricas, las instituciones consultadas, las series temporales, las redes neuronales y los software's utilizados. En el Capítulo 3 se incluyen la metodología empleada para abordar la decisión del lugar a evaluar, el EDA, el tratamiento con los valores atípicos y no disponibles, la división del conjunto de entrenamiento y de prueba, la selección de la máquina de aprendizaje automático, la selección de hiper-parámetros, el análisis de sensibilidad y el tratamiento a posterior de los datos entregados por el modelo para ajustarse a los objetivos. En el Capítulo 4 se incluyen los resultados y análisis obtenidos mediante la metodología empleada. Y finalmente, en el Capítulo 5 se incluyen las conclusiones y recomendaciones asociadas con esta memoria.

2 Estado del arte

Índice general

2.1 La tormenta eléctrica

Una tormenta eléctrica es una serie de descargas eléctricas repentinas provocadas por condiciones atmosféricas. Estas descargas eléctricas dan lugar a repentinos destellos de luz y estruendosas ondas sonoras, conocidas como rayos y truenos respectivamente. La tormenta eléctrica es un fenómeno meteorológico de meso-escala[1] con una escala espacial que abarca desde unos pocos kilómetros hasta un par de cientos de kilómetros y una escala temporal que abarca de menos desde una hora hasta varias horas y que se produce estacionalmente. Las tormentas eléctricas pueden llegar acompañadas lluvias torrenciales, fuertes ráfagas de viento, granizo ocasional y tornados (Chaudhuri, 2011).

En Chile, son varias las condiciones atmosféricas que propician las tormentas eléctricas en la cordillera. En un 80 % de los casos, las tormentas se asocian a sistemas meteorológicos de niveles medios que generan un flujo desde el Este sobre la zona centro-sur. Esto es de suma

[1] En meteorología, un fenómeno de meso-escala es un sistema atmosférico que oscila espacialmente desde los 2 km hasta los 2000 km y temporalmente desde 1 hora hasta varias horas (Orlanski, 1975).

importancia, ya que nuestro país está sometido a un constante viento del Oeste y este cambio de dirección del viento desencadena el trasporte de humedad desde Argentina hacia nuestro lado de la cordillera. El otro 20 % está asociado a sistemas meteorológicos provenientes desde el Océano Pacífico, transportando humedad con vientos del Oeste intensificados (Viale y Garreaud, 2013).

En general, el alto contenido de humedad y el ascenso del aire desempeñan un papel importante en la formación de tormentas eléctricas. Estas son nubes cúmulo que crecen verticalmente en lugar de hacerlo horizontalmente como puede ver en la figura 2.1 (Cooper y Holle, 2019). Existen varias causas que provocan la elevación de aire cálido y húmedo, como el calentamiento del aire por la radiación solar, el encuentro de dos corrientes de aire diferentes, la proximidad de un canal de baja presión, etc. Cuando el aire húmedo se eleva y se enfría, la humedad del aire se condensa y forma nubes. Debido a la elevación del aire húmedo, la nube se hace más grande y las gotas de agua siguen creciendo en la nube y se congelan hasta formar cristales de hielo. Tan pronto las gotas de agua se vuelven pesadas, caen en forma de granizo. El granizo adquiere carga negativa debido al roce con los cristales de hielo de las nubes. Así, las cargas negativas se acumulan en la base de la nube y se crean cargas positivas en la parte superior de la nube. Estas cargas negativas son atraídas por otras nubes, objetos y la Tierra. Cuando la atracción aumenta, las cargas negativas y positivas se descargan o se unen para formar un rayo o relámpago. Los rayos calientan y expanden el aire, lo que produce los truenos. Los rayos se clasifican generalmente en cuatro categorías (Bala, Choubey, y Paul, 2017). Estas categorías son:

- Entre nubes (IC, por sus siglas en inglés).
- Nube a nube (Cloud-to-cloud (CC), por sus siglas en inglés).
- Nube a tierra (Cloud-to-ground (CG), por sus siglas en inglés).
- Nube a aire (Cloud-to-air (CA), por sus siglas en inglés).

Figura 2.1: Columnas de cúmulos con corrientes ascendentes indicadas con flechas rojas. Colorado, Estados Unidos (Bala, Choubey, y Paul, 2017)

Otra clasificación sobre los tipos de rayos es referente a su polaridad y dirección, estos pudiendo ser: positivos, negativos, ascendentes y descendentes (Tomas, 2004).

Un factor que interviene en el clima son los rayos que producen NOx (gas de efecto invernadero). El aumento de las concentraciones de aerosoles y su distribución en la troposfera también afectan al clima y pueden provocar un aumento de la actividad de los rayos (Singh y cols., 2011). La NASA ha reportado que la aparición de rayos se ha incrementado entre un 5 y un 6 % cuando la temperatura global del planeta se ha incrementado en 1 °C (Romps, Seeley, Vollaro, y Molinari, 2014). Además, existe un variado listado de consecuencias negativas asociadas a la ocurrencia de rayos (Price y Rind, 1992) (Carey, Rutledge, y Petersen, 2003). Tales como:

- Fuente de óxidos de nitrógeno hacia la atmósfera.
- Incendios forestales y contribuyente de aerosoles que queman biomasa.
- Lesiones y muertes humanas y de ganado.
- Fuente de interferencia electromagnética.
- Daños a líneas de transmisión, turbinas eólicas y paneles fotovoltaicos.
- Impactos adversos en la industria de la aviación.
- En los centros espaciales, son un peligro para las tripulaciones de combustible, las operaciones terrestres y las operaciones de lanzamiento de cohetes.
- Consecuencias económicas asociadas a la pérdida de producción de energía, costos adicionales de mantenimiento o incluso la pérdida del equipo operativo.

2.2 La obtención de los datos

2.2.1 Dirección Meteorológica de Chile

La Dirección Meteorológica de Chile (DMC) es el organismo público chileno responsable del quehacer meteorológico en el país, cuyo propósito es satisfacer las necesidades de información y previsión meteorológica de todas las actividades nacionales. Depende de la Dirección General de Aeronáutica Civil. Sus funciones básicas son proporcionar la información meteorológica que requiere la Aeronáutica, proveer servicios meteorológicos y climatológicos a las diferentes actividades socio-económicas que requiere el país para su desarrollo, realizar investigación meteorológica en coordinación con organismos nacionales e internacionales y administrar el Banco Nacional de Datos Meteorológicos. Su objetivo es mitigar los daños por fenómenos meteorológicos extremos para contribuir a la protección de las personas, sus bienes y aportar al desarrollo socio-económico del país en un marco de eficiencia, eficacia y de acuerdo a estándares de calidad, la DMC entrega y elabora diversos productos y servicios (*Dirección Meteorológica de Chile*, s.f.). Tales como:

- Realizar el pronóstico meteorológico oficial para todo el territorio nacional.
- Levantar avisos e información de tiempo y clima para los distintos sectores socio-económicos como la aeronáutica, agricultura, protección civil, medio ambiente, salud, y otros.
- Realizar estudios e investigaciones sobre cambio climático.
- Establecer y mantener enlace con organismos e instituciones científicas internacionales de meteorología, con el fin de optimizar la gestión y proyectar la Dirección en el ámbito científico internacional.
- Administrar el Banco Nacional de Datos Meteorológicos.
- Elaborar la normativa subsidiaria en relación con el reglamento Servicio Meteorológico para la Navegación Aérea.
- Instalar, mantener y operar los sistemas meteorológicos implementados para la navegación aérea y otras actividades.

2.2.2 World Wide Lightning Location Network

La WWLLN (pronunciada como 'woollen') es una red mundial de sensores de rayos por radio de muy baja frecuencia operada por la Universidad de Washington en Seattle, Estados Unidos. Es capaz de generar cada día después de la medianoche UTC, un mapa de densidad de los relámpagos registrados en sus sensores durante el día anterior. Puede realizar animaciones de vídeo superponiendo los datos registrados en imágenes de satélite, donde los vídeos se actualizan cada 20 minutos. Y puede disponer de espectrograma a muy baja frecuencia. La mayoría de las observaciones terrestres ocurren la banda de muy baja frecuencia (desde 3 hasta 30 kHz) y están dominadas por señales impulsivas de descargas de rayos denominadas "sferics". Existe una potencia electromagnética radiada significativa desde unos pocos Hz hasta varios

cientos de MHz, con la mayor parte de la energía radiada en muy baja frecuencia. Es capaz de elaborar mapas periódicos de la actividad de los rayos en todo el planeta (*WWLLN*, s.f.).

Funciona de la siguiente manera. Todos los "anfitriones" reciben los datos mundiales para su propia investigación. A cambio, cada anfitrión proporciona el ordenador y se hace cargo de los gastos locales, como electricidad, Internet y mantenimiento. Sin embargo, un único sensor no va a proporcionar los datos sobre la localización de los rayos por sí solo. Eso sólo lo hace el conjunto de la red. Cada localización de un rayo requiere la hora de llegada de un grupo de al menos 5 sensores. Estos sensores pueden encontrarse a varios miles de kilómetros del rayo. La disposición geográfica de los sensores es importante, pues un rayo que está rodeado de sensores se localiza con mayor precisión que uno que no lo está. Además, como la Tierra es esférica, no hay bordes, por lo que cada rayo está rodeado de sensores, pero no necesariamente por los sensores que lo detectan. Normalmente, sólo entre el 15 y el 30 % de los rayos detectados por un sensor son detectados por 5 o más sensores. Investigaciones recientes indican que la eficacia de detección de las corrientes de unos 30 kA es de aproximadamente un 30 % a nivel mundial (*WWLLN*, s.f.).

2.3 Las series temporales

Una serie de tiempo o serie temporal es una realización de un proceso estocástico que tiene un conjunto de variables aleatorias indexadas en el tiempo. El proceso estocástico hace referencia a variables aleatorias reales denotadas como $\{X_i : i \in I\}$. Los valores observados en un proceso estocástico, serán denotados como $\{X_i(\omega) : i \in T\}$, donde $T \in \mathbb{N}$. Entonces, se puede presentar la serie temporal como $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_T\}$, donde cada x_i corresponde a una observación en el i-ésimo instante. De esta forma, el conjunto de datos vendrá dado por $D_l = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$, donde cada $\mathbf{x}_i$ tiene largo T y está dado por $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{i,T}\}$, con $x_{i,j} \in \mathbb{R}$ e $y_i \in Y \equiv [-1, 1]^K$ (Arnold, 1974). El término serie temporal es común usarlo tanto para referirse al proceso como a una realización concreta, y no se hará una distinción entre ambos conceptos (Robert H. Shumway, 2017). A continuación, se presentan definiciones importantes.

El aspecto más importante de los valores de las series temporales es su su correlación con pasos temporales anteriores (denominados desfases, retrasos o rezagos), en palabras sencillas, su dependencia de valores anteriores. En este sentido, la *función de autocorrelación*[2] es una función que mide la predictibilidad lineal de la serie en el tiempo t, digamos x_t, utilizando sólo el valor de x_s. Viene dada por la ecuación 2.1. Es posible demostrar que $-1 \leq \rho(s,t) \leq 1$ utilizando la desigualdad de Cauchy-Schwarz[3] (Robert H. Shumway, 2017).

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{2.1}$$

Donde γ es la *función de autocovarianza*. Esta función está definida como el producto de segundo momento para todo s y t. Mide la dependencia lineal entre dos puntos de la misma serie, pero observados en momentos diferentes. Viene dada por la ecuación 2.2. Las series muy suaves presentan funciones de autocovarianza que siguen siendo grandes incluso cuando t y s están muy separadas, mientras que las series entrecortadas tienden a tener funciones de autocovarianza cercanas a cero para grandes separaciones (Robert H. Shumway, 2017).

$$\gamma(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{2.2}$$

donde μ es la *función de media*. Esta función se define por la ecuación 2.3. Y el operador E denota el valor esperado o la esperanza (Robert H. Shumway, 2017).

[2] Dado que la correlación de las observaciones de las series temporales se calcula con valores de la misma serie en momentos anteriores, se denomina autocorrelación o correlación serial.

[3] $|\gamma(s,t)|^2 \leq \gamma(s,s)\gamma(t,t)$

$$\mu_t = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx. \tag{2.3}$$

Una serie temporal *estrictamente estacionaria* es aquella para la que el comportamiento probabilístico de cada colección de valores $x_{t_1}, x_{t_2}, \ldots, x_{t_k}$ es idéntico al del conjunto desplazado en el tiempo $x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}$ para todos los desplazamientos en h unidades. Una serie temporal *débilmente estacionaria*, x_t, es un proceso de varianza finita tal que la función de valor medio, μ_t, definida en 2.3 es constante y no depende del tiempo t, y la función de autocovarianza, $\gamma(s,t)$, definida en 2.1 depende de s y t únicamente a través de su diferencia $|s-t|$. En lo sucesivo, utilizaremos el término estacionario para referirnos a débilmente estacionario; si un proceso es estacionario en sentido estricto, utilizaremos el término estrictamente estacionario (Robert H. Shumway, 2017).

Para realizar cualquier análisis estadístico significativo de datos en series de tiempo, al menos, la media y las funciones de autocovarianza satisfacen las condiciones de estacionariedad durante un período razonable de tiempo. Estas condiciones son esperanza constante y autocovarianza constante. Con series temporales, los supuestos de la estadística descriptiva (normalidad, independencia e idéntica distribución de los datos) pierden validez. La descomposición de una serie temporal es una tarea estadística que descompone una serie temporal en 4 componente, cada uno de las cuales se describe a continuación (Athanasopoulos, 2018).

1. **Tendencia**: La tendencia representa el cambio en las variables dependientes con respecto al tiempo desde el principio hasta el final. En caso de tendencia creciente, la variable dependiente aumentará con el tiempo y viceversa. No es necesario tener una tendencia definida en las series temporales, podemos tener una única serie temporal con tendencia creciente y decreciente. En resumen, la tendencia representa la media variable de los datos de las series temporales.
2. **Estacionalidad**: Si después de un intervalo de tiempo fijo, las observaciones mantienen su media y varianza, se denominan observaciones estacionales. No necesariamente deben repetirse los mismos valores. Estos cambios estacionales en los datos pueden ocurrir debido a acontecimientos naturales o provocados por el hombre.
3. **Irregularidades**: También conocidas como ruido. Son saltos y caídas extrañas en los datos. Estas fluctuaciones son causadas por acontecimientos incontrolables.
4. **Ciclicidad**: La ciclicidad se produce cuando las observaciones de la serie se repiten siguiendo un patrón aleatorio. Tenga en cuenta que, si existe un patrón fijo se convierte en estacionalidad. En el caso de la ciclicidad, las observaciones pueden repetirse al cabo de una semana, unos meses o incluso un año. Este tipo de patrones son mucho más difíciles de predecir.

2.4 Las redes neuronales artificiales

Las Redes Neuronales Artificiales (RNA) son modelos computacionales que pertenecen a los modelos del Aprendizaje Automático (o Machine Learning). Si bien su desarrollo se remonta a principios de los años 40, su popularidad aumentó a finales los años 80 a consecuencia del descubrimiento de nuevas técnicas y avances en la tecnología del hardware informático. Su principal inspiración reside del deseo de producir sistemas artificiales capaces de realizar cálculos sofisticados, tal vez *inteligentes*, similares a los que realiza el cerebro humano (Mitchell, 1997).

> « Se dice que un programa informático aprende de la experiencia E con respecto a una clase de tareas T y una medida de rendimiento P, si su rendimiento en las tareas de T, medido por P, mejora con la experiencia E. »
>
> — Tom Mitchell

La mayoría de las RNA tienen algún tipo de regla de *entrenamiento*. En otras palabras, *aprenden* a partir de ejemplos y muestran cierta *capacidad de generalización*[4] más allá de los datos de entrenamiento. A diferencia de la programación explícita, las redes neuronales no necesitan analizar el problema a resolver pues se adaptan durante un *periodo de entrenamiento*, basándose en ejemplos de problemas similares incluso sin una solución deseada para cada problema. Tras el entrenamiento suficiente, la RNA es capaz de relacionar características de entrada (*input features*) con las salidas objetivo (*output target*). Puede observar una comparativa frente a la programación tradicional en la tabla 2.1 (Kruse, Borgelt, Braune, Mostaghim, y Steinbrecher, 2016).

Programación tradicional	**Redes neuronales**
Razonamiento deductivo	Razonamiento inductivo
Computación centralizada, sincrónica y en serie	Computación colectiva, asíncrona y en paralelo
Memoria empaquetada, almacenada literalmente y direccionable por ubicación	Memoria distribuida, internalizada y contenido direccionable.
No tolera fallos	Tolera fallos
Exacta	Inexacta
Conectividad estática	Conectividad dinámica
Reglas bien definidas y datos de entrada precisos	Reglas desconocidas o complicadas y datos de entrada ruidosos o parciales

Tabla 2.1: Comparativa entre Programación tradicional y Redes neuronales

La unidad básica de una RNA es la neurona, donde cada neurona puede recibir una o varias entradas, pero solo emite una salida, como puede observar en la figura 2.2. Para formar una RNA las neuronas se conectan entre sí. En cada neurona, cada entrada tiene un *peso* asociado que modifica la fuerza de cada entrada. La neurona suma todas las entradas y calcula una salida que se transmitirá a la siguiente neurona (Kruse y cols., 2016). Una de las primeras RNA fue desarrollada por McCulloch y Pitts en 1943 (McCulloch y Pitts, 2021). Podemos ver la estructura de una RNA en la figura 2.3, donde cada "bias" o "sesgo" vale 1, es decir, $h_0^i = 1, \forall i$.

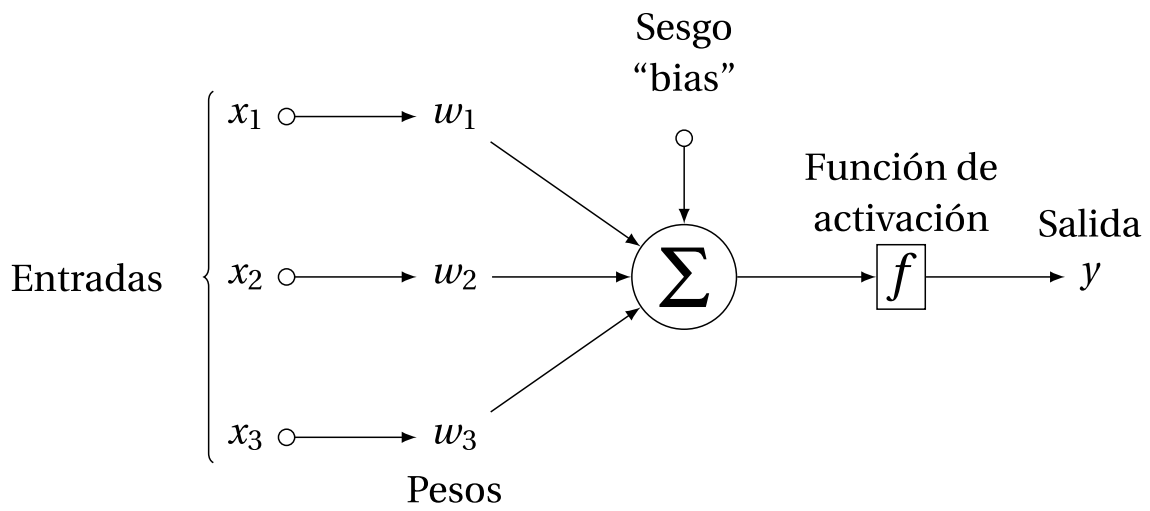


Figura 2.2: Estructura de una neurona

[4] Véase 2.4.1.

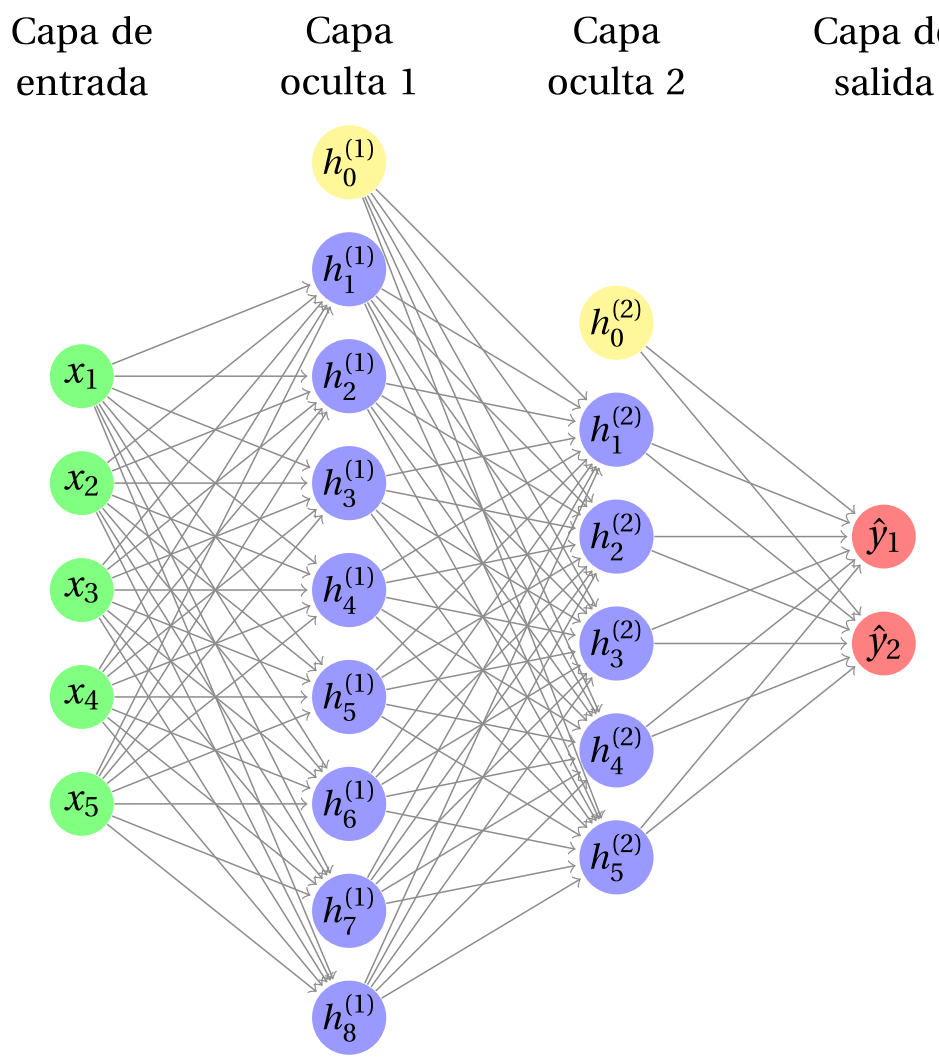


Figura 2.3: Estructura de una Red neuronal

Llamaremos al par $(\mathbf{x}_m, y_m)$ como el ejemplo de entrenamiento, donde $\mathbf{x}_m$ corresponde al m-ésimo valor de las I características de entrada e y_m al m-ésimo valor de la salida objetivo. Un conjunto de entrenamiento S_M es un grupo de M ejemplos de entrenamiento, $S_M = (\mathbf{x}_m, y_m), m = 1, \ldots, M$. El objetivo es encontrar la función tal que, del espacio de características reproduzca el espacio de salida. Llamaremos a esta función como la *hipótesis* o *learner*. En otras palabras, una RNA es un modelo no lineal que recibe entradas que pueden ser numéricas o categóricas y devuelve salidas. Si la salida es continua, es un problema de regresión, pero si la salida tiene un número finito de K valores discretos es un problema de clasificación. En particular, si $K = 2$ es un problema de clasificación binaria (Goodfellow, Bengio, y Courville, 2016).

Por otro lado, la *función de pérdida* ℓ está definida tal que $\ell : R \times Y \to [0, \infty[$ y cuantifica el rendimiento de la respuesta obtenida de $\hat{y} = f(\mathbf{x})$ respecto de la respuesta verdadera o deseada y. Es decir, representa la calidad de una hipótesis h. Por ejemplo, una función de pérdida es la función de pérdida cuadrática que se expresa en la ecuación 2.4 (Goodfellow y cols., 2016).

$$\ell(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2. \tag{2.4}$$

De esta forma. Sea $w = [w_0, w_1, \ldots, w_I]^T$ el vector de los parámetros o *pesos* que determinan cómo afecta la predicción de un modelo lineal definido por la ecuación 2.5, tal que $x^{(0)} = 1$ e I es el número de características de entrada (Goodfellow y cols., 2016).

$$f(x) = \sum_{i=0}^{I} w_i x^{(i)} = w^T \mathbf{x}. \tag{2.5}$$

El valor de los parámetros w que minimizan la forma matricial de la función cuadrática de pérdida al minimizar la ecuación 2.6[5]

[5] Por comodidad usamos $m = 2$.

$$J(w) = \frac{1}{m}\sum_{m=1}^{M}(y_m - f(\mathbf{x}_m))^2 \rightarrow J(w) = \frac{1}{2}(Y - Xw)^T(Y - Xw), \tag{2.6}$$

quedan definidos por la ecuación 2.7 (Goodfellow y cols., 2016).

$$w_{LMS} = (X^T X)^{-1} X^T Y. \tag{2.7}$$

2.4.1 Capacidad de generalización, overfitting y underfitting

Normalmente, al entrenar un modelo de ML, se tiene acceso a un conjunto de entrenamiento[6] y calculamos alguna medida para el error sobre el conjunto de entrenamiento denominada *error de entrenamiento*. Lo que separa el aprendizaje automático de la optimización es que queremos que el *error de generalización* o *error de prueba* sea también bajo. El error de generalización se define como el valor esperado del error para una nueva entrada (Goodfellow y cols., 2016).

Normalmente, se estima el error de generalización de un modelo de ML midiendo su rendimiento en un *conjunto de prueba* de ejemplos que se recogieron por separado del conjunto de entrenamiento. Esta forma de influir en el rendimiento del conjunto de pruebas cuando sólo podemos observar el conjunto de entrenamiento introduce tres conceptos importantes. La capacidad de generalización, el overfitting y el underfitting (Goodfellow y cols., 2016).

La *capacidad de generalización* está definida como el método o algoritmo automático capaz de estimar ejemplos futuros basándose en el fenómeno observado en el conjunto de entrenamiento. El *sobreajuste* u *overfitting* ocurre cuando los algoritmos que memorizan las muestras del entrenamiento tienen un pobre rendimiento predictivo con ejemplos desconocidos. Y el *infraajuste* o *underfitting* ocurre cuando el modelo no es capaz de obtener un valor de error suficientemente bajo en el conjunto de entrenamiento. El underfitting se produce cuando la diferencia entre el error de entrenamiento y el error de prueba es demasiado grande. En la figura 2.4 podemos ilustrar los tres conceptos (Goodfellow y cols., 2016).

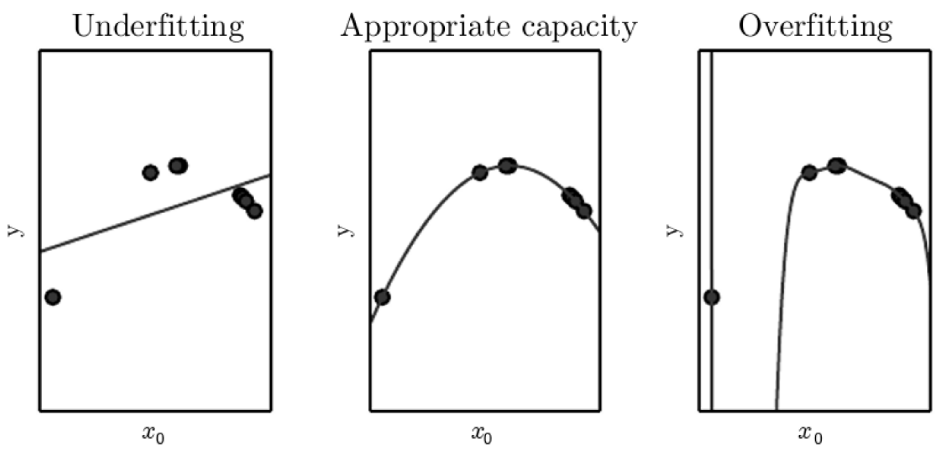


Figura 2.4: Modelo con underfitting, generalizado y con overfitting (Goodfellow y cols, 2016)

Debemos recordar que, aunque las funciones más sencillas tienen más probabilidades de generalizar, debemos elegir una hipótesis lo suficientemente compleja para lograr un error de entrenamiento bajo. Normalmente, el error de entrenamiento suele disminuir hasta alcanzar el mínimo valor de error posible a medida que aumenta la capacidad del modelo. Típicamente el error de generalización tiene una curva en forma de "U" en función de la capacidad de generalización del modelo. Esto se ilustra en la figura 2.5 donde el error de entrenamiento y el de prueba se comportan de forma diferente[7]. En el extremo izquierdo la figura, tanto el

[6] El conjunto original de datos se divide en conjunto de entrenamiento y conjunto de prueba

[7] Aunque suena contradictorio con el razonamiento lógico (cuanto más entreno, mejor es la capacidad de generalizar). El razonamiento inductivo, o de inferir reglas generales a partir de un conjunto limitado de ejemplos, no es lógicamente válido. Para inferir lógicamente una regla que describa a todos los miembros de un conjunto, se debe

error de entrenamiento como el de generalización son elevados. Este es el underfitting. A medida que aumentamos la capacidad (número de épocas), el error de entrenamiento disminuye, pero aumenta la diferencia entre el error de entrenamiento y el de generalización. Finalmente, el tamaño de esta diferencia supera la disminución del error de entrenamiento y entramos en el régimen de sobreajuste, en el que la capacidad es demasiado grande, por encima de la capacidad óptima (Goodfellow y cols., 2016).

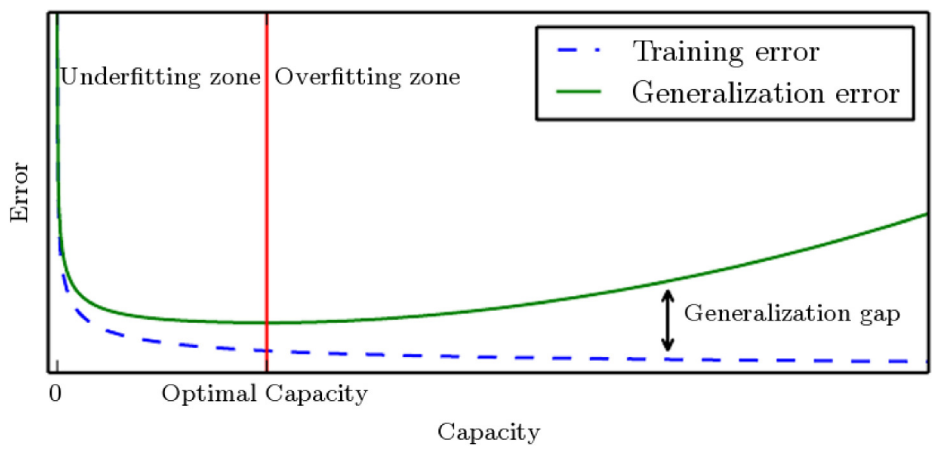


Figura 2.5: Relación típica entre la capacidad de generalización y el error (Goodfellow y cols, 2016)

2.4.2 Regularización

Un problema *bien definido* o *bien propuesto* es un problema que consiste en resolver una ecuación diferencial sujeta a condiciones de frontera o de valor inicial cuando una de las variables que la definen toma cierto valor. Las soluciones de estos problemas tienen una estructura que suelen incluir que *existe una solución*, la *solución es única* y una *solución es dependiente de las condiciones iniciales de manera continua*. Si no está bien planteado, habrá que formularlo de nuevo para su tratamiento numérico. Normalmente, esto implica incluir supuestos adicionales, como la suavidad de la solución. Este proceso se conoce como regularización. La regularización de Tikhonov es una de las más utilizadas para la regularización de problemas lineales mal planteados[8] (Willoughby, 1979).

Para diseñar algoritmos que funcionen bien en una tarea específica, se incorpora un conjunto de preferencias en el algoritmo de aprendizaje. Cuando estas preferencias se alinean con los problemas de aprendizaje que le pedimos al algoritmo que resuelva, éste rinde mejor. El comportamiento del algoritmo se ve fuertemente afectado no sólo por lo grande del conjunto de funciones permitidas en su espacio de hipótesis, sino por la identidad específica de esas funciones. El algoritmo de aprendizaje estudiado (regresión lineal) tiene un espacio de hipótesis que consiste en el conjunto de funciones lineales de su entrada. Estas funciones lineales pueden ser muy útiles para problemas en los que la relación entre entradas y salidas es cercana a la lineal. Son menos útiles para problemas que se comportan de manera no lineal[9]. Así, se controla el rendimiento del algoritmo eligiendo el tipo de función para permitir extraer soluciones (Goodfellow y cols., 2016).

Se puede hacer que un algoritmo de aprendizaje tome la preferencia por una solución en su espacio de hipótesis frente a otra solución. Esto significa que ambas funciones son elegi-

tener información sobre cada miembro de ese conjunto. En parte, del aprendizaje automático evitar este problema ofreciendo sólo reglas probabilísticas, en lugar de las reglas totalmente seguras utilizadas en el razonamiento puramente lógicos.

[8] En estadística, el método se conoce como regresión Ridge, en aprendizaje automático se conocen como decaimiento del peso (weight decay), y con múltiples descubrimientos independientes, también se conoce como método Tikhonov-Miller, método Phillips-Twomey, método de inversión lineal restringida, regularización L^2 y método de regularización lineal. Está relacionado con el algoritmo de Levenberg-Marquardt para problemas de mínimos cuadrados no lineales.

[9] La regresión lineal no funciona bien si se intenta utilizar para predecir, por ejemplo, $\sin(x)$ a partir de x

bles, pero se prefiere una. La solución no preferida se elegirá sólo si se ajusta significativamente mejor a los datos de entrenamiento frente a la solución preferida. Por ejemplo, al modificar el criterio de entrenamiento de la regresión lineal para incluir el decaimiento de los pesos (*weight decay*). Para realizar una regresión lineal con caída de pesos, minimizamos una suma que comprende tanto el error cuadrático medio en el entrenamiento, como un criterio $J(w)$ que expresa una preferencia para que los pesos tengan una norma L^2 más pequeña (Goodfellow y cols., 2016). Tal cómo se muestra en la ecuación 2.8,

$$J(w) = (Y - Xw)^T (Y - Xw) + \lambda w^T w, \tag{2.8}$$

donde λ es un valor elegido de antemano que controla la fuerza de la preferencia por pesos más pequeños. Cuando $\lambda = 0$, no se impone preferencia, y cuando λ es mayor se obliga a que las ponderaciones sean menores. Minimizar $J(w)$ resulta en una elección de pesos que hacen un compromiso entre ajustarse a los datos de entrenamiento y que sean pequeños. Esto nos da soluciones que tienen una pendiente más pequeña, o poner peso a un grupo menor de características. Como ejemplo de cómo podemos controlar la tendencia de un modelo a sobre ajustarse o desajustarse mediante el decaimiento del peso, podemos entrenar un modelo de regresión polinómica de alto grado con distintos valores de λ, tal como puede ver en figura 2.6 (Goodfellow y cols., 2016).

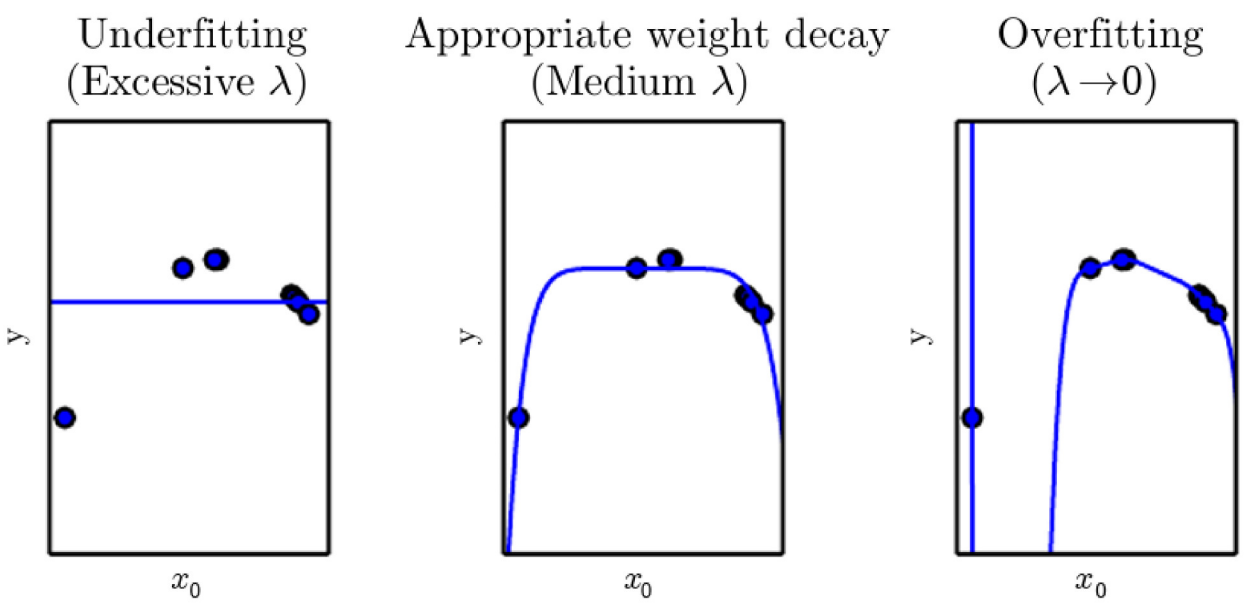


Figura 2.6: Modelos entrenados con diferentes valores de λ (Goodfellow y cols., 2016)

De forma más general, podemos regularizar un modelo que aprende una función $f(\mathbf{x};\theta)$ añadiendo una penalización llamada regularizador a la función de coste. En el caso del decaimiento del peso, el regularizador es $\Omega(w) = w^T w$. La regularización es cualquier modificación del algoritmo de ML que pretenda reducir su error de generalización, pero no su error de entrenamiento. La regularización es una de las preocupaciones centrales del campo del ML, rivalizando en importancia sólo con la optimización (Goodfellow y cols., 2016).

2.4.3 Hiper-parámetros y conjunto de validación

La mayoría de los algoritmos de aprendizaje automático tienen varios ajustes que podemos utilizar para controlar el comportamiento del algoritmo de aprendizaje. Estos ajustes se denominan hiper-parámetros. Los valores de los hiper-parámetros no los adapta el propio algoritmo de aprendizaje[10]. A veces se elige un ajuste que es un hiper-parámetro que el algoritmo de aprendizaje no aprende porque es difícil de optimizar. Esto se aplica a todos los hiper-parámetros que controlan la capacidad de generalización del modelo. Si se aprenden en

[10] aunque podemos diseñar un procedimiento de aprendizaje anidado en el que un algoritmo de aprendizaje aprenda los mejores hiper-parámetros para otro algoritmo de ML

el conjunto de entrenamiento, dichos hiper-parámetros siempre elegirían la máxima capacidad de generalización posible del modelo, lo que daría lugar a un sobre ajuste. Por ejemplo, siempre podemos ajustar mejor el conjunto de entrenamiento con un polinomio de grado más alto y un ajuste de caída de peso de $\lambda = 0$ que con un polinomio de grado más bajo y un ajuste de caída de peso positivo. Para resolver este problema, necesitamos un *conjunto de validación* de ejemplos que el algoritmo de entrenamiento no observe (Goodfellow y cols., 2016).

El conjunto de validación es un subconjunto del conjunto de entrenamiento que funciona cómo un *conjunto de prueba retenido*, compuesto por ejemplos procedentes de la misma distribución que el conjunto de entrenamiento. Este conjunto de validación puede utilizarse para estimar el error de generalización de un modelo, una vez completado el proceso de aprendizaje. Es importante que los ejemplos de prueba no se utilicen en modo alguno para tomar decisiones sobre el modelo, incluidos sus hiper-parámetros. Por esta razón, ningún ejemplo del conjunto de prueba puede utilizarse en el conjunto de validación. Por lo tanto, siempre construimos el conjunto de validación a partir de los datos de entrenamiento. En concreto, dividimos los datos de entrenamiento en dos subconjuntos disjuntos. Uno de estos conjuntos se utiliza para aprender los parámetros y el otro conjunto es nuestro conjunto de validación, utilizado para estimar el error de generalización durante o después del entrenamiento, permitiendo que los hiper-parámetros se actualicen en consecuencia. El subconjunto de datos utilizado para aprender los parámetros suele denominarse conjunto de entrenamiento, aunque pueda confundirse con el conjunto de datos más amplio utilizado para todo el proceso de entrenamiento. Dado que el conjunto de validación se utiliza para "entrenar" los hiper-parámetros, el error del conjunto de validación subestimará el error de generalización, aunque normalmente por una cantidad menor que el error de entrenamiento. Una vez completada la optimización de los hiper-parámetros, el error de generalización puede estimarse utilizando el conjunto de prueba (Goodfellow y cols., 2016).

La *validación cruzada* se utiliza cuando dividir el conjunto de datos en un conjunto fijo de entrenamiento y un conjunto fijo de prueba puede ser problemático si el conjunto de prueba es pequeño. Un conjunto de pruebas pequeño implica incertidumbre estadística en torno al error de prueba medio estimado, lo que hace difícil afirmar que cierto "algoritmo A" funcione mejor que otro "algoritmo B" en la tarea en cuestión. Cuando el conjunto de datos es demasiado pequeño, hay procedimientos alternativos que permiten utilizar todos los ejemplos en la estimación del error medio de la prueba, al precio de un mayor coste computacional. Estos procedimientos se basan en la idea de repetir el cálculo de entrenamiento y prueba en diferentes subconjuntos o divisiones elegidos al azar del conjunto de datos original. El algoritmo de validación cruzada k-fold estratificada es uno de estos procedimientos, en el que se forma una partición del conjunto de datos dividiéndolo en k subconjuntos que no se solapan y mantiene la proporción de las clases. El error de la prueba puede estimarse tomando la media del error de la prueba en k ensayos. En el enésimo ensayo, el subconjunto enésimo de los datos se utiliza como conjunto de prueba y el resto de los datos como conjunto de entrenamiento (Goodfellow y cols., 2016).

2.4.4 Descenso del gradiente estocástico

Un problema recurrente en el aprendizaje automático es que se necesitan grandes conjuntos de entrenamiento para una buena generalización, pero entrenar estos conjuntos también es más complejo desde el punto de vista computacional. La función de coste utilizada por un algoritmo de ML a menudo se descompone como una suma sobre la cantidad de ejemplos de entrenamiento de alguna función de pérdida, por ejemplo, la log-verosimilitud condicional negativa de los datos de entrenamiento puede escribirse como

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} L(\mathbf{x}^{(i)}, y^{(i)}, \theta), \tag{2.9}$$

donde L es la pérdida por cada ejemplo $L(\mathbf{x}, y, \theta) = -\log p(y|x;\theta)$ (Goodfellow y cols., 2016).

Para estas funciones de coste aditivo, el descenso gradiente requiere calcular el gradiente $\nabla_\theta J(\theta)$. A medida que el tamaño del conjunto de entrenamiento aumenta, el tiempo necesario para dar un solo paso de gradiente se vuelve prohibitivo. El coste computacional de esta operación es O(m) (Goodfellow y cols., 2016).

La idea del descenso de gradiente estocástico es que el gradiente es una expectativa. La expectativa puede estimarse aproximadamente utilizando un pequeño conjunto de muestras. Específicamente, en cada paso del algoritmo, podemos muestrear un pequeño lote de ejemplos (minibatch) B = $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m')}\}$ extraídos uniformemente del conjunto de entrenamiento. El tamaño del pequeño lote m' suele elegirse como un número relativamente pequeño de ejemplos, entre 1 y varios cientos. Lo más importante es que m' suele mantenerse fijo a medida que aumenta el tamaño m del conjunto de entrenamiento. Podemos ajustar un conjunto de entrenamiento con miles de millones de ejemplos utilizando actualizaciones calculadas en sólo cien ejemplos (Goodfellow y cols., 2016).

La estimación del gradiente se forma como

$$g = \frac{1}{m'} \nabla_\theta \sum_{i=1}^{m'} L(\mathbf{x}^{(i)}, y^{(i)}, \theta), \tag{2.10}$$

utilizando ejemplos del pequeño lote B. A continuación, el algoritmo de descenso de gradiente estocástico sigue el gradiente estimado cuesta abajo como (Goodfellow y cols., 2016)

$$w_i^{p+1} = w_i^p - \alpha \frac{1}{M} \sum_{m=1}^{M} (f(\mathbf{x}_m) - y_m)\mathbf{x}_m^{(i)}. \tag{2.11}$$

En general, el descenso por gradiente puede alcanzar un mínimo local. Sin embargo, J es una función convexa. Por lo tanto, el problema de optimización tiene sólo un óptimo global. El descenso de gradiente estocástico tiene muchos usos importantes fuera del contexto del aprendizaje profundo. Es la principal forma de entrenar grandes modelos lineales en conjuntos de datos muy grandes. Para un tamaño de modelo fijo, el coste por actualización del descenso de gradiente estocástico no depende del tamaño del conjunto de entrenamiento m. En la práctica, a menudo se utiliza un modelo más grande a medida que aumenta el tamaño del conjunto de entrenamiento, pero no se está obligado a hacerlo. El número de actualizaciones necesarias para alcanzar la convergencia suele aumentar con el tamaño del conjunto de entrenamiento. Sin embargo, a medida que m se acerca a infinito, el modelo acabará convergiendo a su mejor error de prueba posible antes de que el descenso de gradiente estocástico haya muestreado todos los ejemplos del conjunto de entrenamiento. Aumentar m no aumentará el tiempo de entrenamiento necesario para alcanzar el mejor error de prueba posible del modelo. Desde este punto de vista, se puede argumentar que el coste asintótico de entrenar un modelo con descenso de gradiente estocástico es O(1) en función de m (Goodfellow y cols., 2016).

2.4.5 El problema de la clasificación binaria y los datos desequilibrados

La idea principal del *submuestreo* o *under bagging* es generar subconjuntos de datos más equilibrados a partir del conjunto de datos originalmente desequilibrados mediante una muestra bootstrap que contenga la clase minoritaria. Después se aplican los algoritmos de clasificación binaria a cada subconjunto de datos (Leo, 1996).

2.4.6 Evaluación del rendimiento

Para evaluar el rendimiento de los algoritmos, se introduce la matriz de confusión, que se presenta en la figura 2.7 (Aggarwal, 2014).

Donde:

		Resultado de la predicción		
		No ocurrencia	**Ocurrencia**	**Total**
Valor real	**No ocurrencia**	True Negative	False Positive	N′
	Ocurrencia	False Negative	True Positive	P′
	Total	N	P	

Figura 2.7: Matriz de confusión

- **True Positive o Aciertos**: Valores verdaderos positivos o número de muestras observadas como rayos que fueron identificadas correctamente por el clasificador.
- **False Negative o Fallos**: Valores falsos negativos o número de muestras observadas como rayos que fueron identificadas erróneamente como cielos despejados por el clasificador. También llamado error tipo II.
- **False Positive o Falsa alarma**: Valores falsos positivos o número de muestras observadas como cielos despejados que fueron identificadas erróneamente como rayos por el clasificador. También llamado error tipo I.
- **True Negative o Desestimación correcta**: Valores verdaderos negativos o número de muestras observadas como cielos despejados que fueron identificadas correctamente por el clasificador.

Esta clasificación de los datos permite obtener ciertos indicadores o métricas para la evaluación del desempeño del modelo. A continuación, presento la definición de algunas de estas métricas:

La *precisión*, que viene dada por la ecuación 2.12.

$$Precision = \frac{TP}{TP + FP}. \tag{2.12}$$

El *recall* o probabilidad de detección, que es la proporción de muestras observadas con actividad de rayos identificadas correctamente por el clasificador. Es decir, es la proporción de eventos positivos que se identificaron correctamente. Viene dada por la ecuación 2.13.

$$True\ positive\ rate\ o\ Recall = \frac{TP}{TP + FN}. \tag{2.13}$$

La *tasa de falsas alarmas*, que es la proporción de muestras observadas de cielos despejados clasificados erróneamente como rayos por el clasificador. Viene dada por la ecuación 2.14.

$$False\ alarm\ ratio = \frac{FP}{TP + FP}. \tag{2.14}$$

Un buen rendimiento en precisión no garantiza necesariamente un buen rendimiento en recall, y viceversa. Debido a eso, utilizamos la media armónica entre la precisión el recall. Esta es la *puntuación F1*. Esta viene dada por la ecuación 2.15.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{2.15}$$

El *índice de éxito crítico*, que es la relación entre las previsiones de sucesos realizadas con éxito y el número total de previsiones de sucesos realizadas (TP + FP) y necesarias (FN). Viene dado por la ecuación 2.16.

$$Threat\ index = \frac{TP}{TP + FP + FN}. \tag{2.16}$$

La *Puntuación de destreza de Heidke*, que mide la mejora fraccionaria del pronóstico con respecto al pronóstico estándar. Su rango es $]-\infty, 1]$. Los valores negativos indican que el pronóstico por azar es mejor y 0 significa que no hay habilidad. Viene dada por la ecuación 2.17.

$$Heidke\ Skill\ Score = 2\frac{TP \cdot TN - FN \cdot FP}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}. \tag{2.17}$$

Debe tenerse en cuenta que el mejor valor de los indicadores mencionados es 1 (Wilks, 1995).

2.5 Lenguaje, programas y bibliotecas

2.5.1 Python

Python es un lenguaje de programación de alto nivel que soporta la orientación a objetos, programación imperativa y la programación funcional. Posee una licencia de código abierto, administrada por Python Software Foundation. Fue creado a principios de los años 90 por Guido van Rossum en el Stichting Mathematisch Centrum, Países Bajos (*Python*, s.f.).

Aunque, Python fue creado como lenguaje de programación de uso general, cuenta con una serie de librerías y entornos de desarrollo para cada una de las fases del proceso de la ciencia de datos, lo que sumado a su característica de código abierto le ha llevado a tomar la delantera frente a otros lenguajes como pueden SAS y R (*Wayback*, s.f.).

2.5.2 Jupyter Lab

Jupyter Lab es una interfaz de usuario para el Proyecto Jupyter que ofrece todos los bloques de construcción basado en Jupyter Notebook. Puede organizar varios documentos y actividades uno junto a otro en el área de trabajo mediante pestañas y divisores. Los documentos y las actividades se integran entre sí, lo que permite nuevos flujos de trabajo para la informática interactiva (*JupyterLab Documentation*, s.f.).

2.5.3 Anaconda

Anaconda es una plataforma para desarrollar códigos de Python. Es utilizada en la ciencia de datos y ML. Esta distribución es utilizada por 6 millones de usuarios e incluye más de 250 paquetes de ciencia de datos válidos para Windows, Linux y MacOS. Sus diferentes versiones de paquetes se administran mediante el sistema de gestión de paquetes conda, el cual permite instalar, correr, y actualizar software de ciencia de datos y ML como Scikit-learn, TensorFlow y SciPy (*Anaconda*, s.f.).

2.5.4 Scikit-learn

Scikit-learn es una biblioteca para el aprendizaje automático que cuenta con varios algoritmos de clasificación, regresión, agrupamiento, reducción de la dimensión, selección de modelos y procesamiento previo de datos. Se define como un conjunto herramientas sencillas y eficaces para el análisis de predicción de datos. Posee una licencia de código abierto.

Su lanzamiento inicial ocurrió en 2007 como proyecto de David Cournapeau para el programa Google Summer of Code (*Scikit-learn*, s.f.). Todas las máquinas de aprendizaje utilizadas se encuentran en 3.6.

2.5.5 TensorFlow

TensorFlow es una biblioteca para el aprendizaje automático enfocada especialmente para las redes neuronales. Se define como un ecosistema completo para ayudar a resolver problemas complejos del mundo real con aprendizaje automático. Posee una licencia de código abierto, administrada por Google Brain. Su lanzamiento inicial ocurrió el 9 de noviembre de 2015 (*TensorFlow*, s.f.). Todas las capas de las redes neuronales utilizadas se encuentran en 3.6.

3 Metodología

Índice general

3.1 Selección de la ubicación

La DMC cuenta con 1423 estaciones meteorológicas automáticas a lo largo de todo Chile. Estas estaciones dependen de siete organismos[1] que capturan diferentes variables meteorológicas, tales como, presión a nivel del mar, humedad relativa, temperatura del aire seco, radiación, entre otras. Entonces, al consultar estas estaciones en su sitio web se fue seleccionando las estaciones en las que en un área cuadrada de $30 \times 30 km^2$ contengan tanto la mayor cantidad de ejemplos como la mayor variedad de características (meteorológicas y registros de rayos) en el norte de Chile. Pues son 15 km la distancia a la que un observador puede detectar la presencia de un rayo. Y dentro de Chile, las tormentas eléctricas ocurren con mayor frecuencia en el norte. En la tabla 3.1 puede ver algunas de las Estaciónes Meteorológicas Automáticas (EMA) consultadas.

Realizar la intersección de los conjuntos de datos cercanos entre sí crea algunas problemáticas. Estos problemas son los valores atípicos, los valores no disponibles y las características repetidas, los 2 primeros serán abordados en los apartados 3.3.1 y 3.3.3, respectivamente. Para el tercero no es necesario abordarlo mediante una sección, pues en el párrafo siguiente se aborda esta problemática para una localidad en particular. En la tabla 3.2 se puede observar la correlación existente entre las características de igual nombre para la localidad de Visviri, Chile.

[1] La Dirección Meteorológica de Chile, la Dirección de General de Aguas, el Instituto de Investigaciones Agropecuarias, la Armada de Chile, la Fundación para el Desarrollo Frutícola, el Observatorio Europeo Austral, la Universidad de Santiago de Chile, la División Andina - Codelco y el Instituto Antártico Chileno

Código Nacional	**Nombre de la Estación**	**Zona Geográfica**	**Altitud (m)**
170001	Visviri Tenencia	Cordillera	4084
170005	Visviri DGA	Cordillera	4080
170007	Visviri INIA	Cordillera	4122
180017	Putre	PreCordillera	3532
180018	Defensa Civil, Arica	Litoral	71
200010	UNAP (Universidad Arturo Prat), Iquique	Litoral	30
210901	Ollagüe	Cordillera	3708
220002	El Loa, Calama Ad.	PreCordillera	4880
230004	Toconao	Cordillera	2495
230021	Llano de Chajnantor, observatorio APEX	Valle	5000
270009	Copiapó Universidad de Atacama	Valle	362

Tabla 3.1: Estaciones meteorológicas automáticas

	RR	**HR**	**QFE**	**RadGInst**
Correlación	0.99	1.00	0.99	0.93

Tabla 3.2: Correlación entre las características de igual nombre para la localidad de Visviri, Chile

Luego de encontrar la ubicación donde ocurre tanto la mayor cantidad de rayos como la mayor cantidad de registros meteorológicos, se usó el valor medio de las características meteorológicas repetidas. Luego, se distribuyeron homogéneamente los datos, es decir, datos como la precipitación que se registraron cada 6 horas, se distribuye 1/6 en cada hora[2]. Esto significó, aproximar los registros de rayos a la hora más cercana. Posteriormente, se programó un cuadrado de lado 30000 m en el sistema de coordenadas universal transversal de Mercator, pues a diferencia del sistema de coordenadas geográficas (expresadas en longitud y latitud), estas magnitudes permiten realizar operaciones en metros.

A continuación, se presenta el listado de características relacionadas a la unión de los conjuntos de datos para el pueblo de Visviri, Chile.

- Fecha y hora, Datetime (dd/mm/aaaa h).
- Precipitación acumulada durante 1 hora, RR (mm).
- Humedad relativa del aire, HR (%).
- Presión atmosférica a nivel de la estación, QFE (hPa).
- Presión atmosférica a nivel del mar, QFF (hPa).
- Presión atmosférica a nivel del mar mediante Atmósfera Estándar de la OACI, QNH (hPa).
- Radiación Solar Global Instantánea, RadGInst (W/m^2).
- Temperatura del punto de rocío, Td (°C).
- Temperatura del aire seco, Ts (°C).
- Dirección del viento a 10 m de altura, dd$_{10m}$ (°).
- Intensidad del viento a 10 m de altura, ff$_{10m}$ (kt).
- Dirección del viento promedio cada 2 minutos, dd$_{2min}$ (°).
- Intensidad del viento promedio cada 2 minutos, ff$_{2min}$ (kt).
- Dirección del viento a 2 m de altura, dd$_{2m}$ (°).
- Intensidad del viento a 2 m de altura, intensidad 2 metros, ff$_{2m}$ (kt).

Finalmente, los registros asociados a alguna periodización como por ejemplo los grados de la dirección del viento fueron descompuestos en sus componentes cartesianas para que todos los valores se encuentren separados la misma distancia.

[2]La mayor unidad de tiempo mínima recabada en los sitios consultados fue de 1 hora.

3.2 Análisis exploratorio de datos

Es una práctica habitual realizar un análisis descriptivo de los datos, esto con el propósito de obtener una visión general de la distribución de las características, comprobar si hay valores atípicos, no disponibles u otras anomalías y descubrir patrones o relaciones entre las variables en el conjunto de datos que ayuden en la predicción de días de tormenta o no (también llamados como *cielos despejados* para evitar confusiones) para la localidad seleccionada dentro del territorio chileno. Existen 3 tipos de análisis:

- Análisis descriptivo, que como su nombre lo indica, describe las características de los datos.
- Análisis explicativo, que establece las relaciones causa-efecto entre las características.
- Análisis de predicción que se basan en dos tipos de modelos, los modelos matemáticos[3] y los modelos basados en datos[4].

El *análisis descriptivo de las características numéricas de una variable* permite describir los datos numéricos, ya sea si son *discretos* o *continuos*. Un conjunto básico de estadísticas para desarrollar una comprensión sobre las características numéricas son las siguientes:

- Cantidad de observaciones.
- Medidas de tendencia central como la media o la mediana.
- Medidas de variabilidad como la desviación estándar y el rango entre intervalos.

En la tabla 3.3 puede observar las estadísticas descriptivas que resumen la cantidad, la tendencia central y la variabilidad de los datos disponibles para la localidad compuesta por la unión de los conjuntos de las estaciones ubicadas en Visviri, Chile. Los valores de la característica “Strokes” deben registrar únicamente 2 posibles valores, cuando se registra una descarga y cuando no se registra; para problemas de clasificación binaria se suele usar 1 y -1, respectivamente[5]. Notar que los bajos valores de HR se reflejan en valores negativos de Temperatura del punto de rocío (Td).

Y para la visualización de estos datos se puede utilizar histogramas y diagramas de cajas. De esta forma, en la figura 1 del Anexo puede visualizar la forma de las distribuciones para cada una de las características en función de las horas del día y también de las semanas del año, además de los diagramas de cajas y de violín para cada una de las 18 características para la localidad compuesta por la unión de los conjuntos de las estaciones ubicadas en Visviri, Chile.

En la tabla 3.4 puede ver un resumen con la cardinalidad, los valores de la asimetría y curtosis[6] con el propósito de describir la forma de la distribución de cada una de las características registradas para la unión de los conjuntos de las estaciones ubicadas en Visviri, Chile.

Posteriormente, se realiza un *análisis descriptivo de las características numéricas de dos variables*, con el propósito de describir los datos numéricos por separado para los casos cuando ocurre tormenta y cuando ocurren cielos despejados. En las figuras 2 del Anexo puede visualizar la forma de las distribuciones para cada una de las características cuando ocurre tormenta y cuando ocurren cielos despejados.

[3] Métodos basados en modelos estadísticos o matemático (model-based), para explicar el proceso de generación de las series de tiempo (regresión lineal, polinómica o exponencial, modelos auto-regresivos (AR), modelos de media móvil (moving average, MA), modelos AR + MA (ARMA), modelos de media móvil auto-regresivos e integrados (auto-regresive integrated model-average, ARIMA)), etc.

[4] Métodos basados en datos (data-driven) que aprendan a identificar patrones en los datos (problemas de aprendizaje automático de regresión o clasificación).

[5] Otra forma de clasificar es usar 1 y 0, para registrar cuando ocurre tormenta y cuando no, respectivamente.

[6] La curtosis de una variable estadística que caracteriza la forma de los datos según sea su distribución de frecuencias. Mide la concentración de los valores.

	RR	HR	QFE	QFF	QNH	RadGInst	Td	Ts	cos dd_{10m}
cantidad	74,352	71,925	71,926	57,790	58,861	73,417	58,864	71,924	58,864
media	0.03	38.18	627.21	1,028.21	1,036.46	286.66	-9.80	6.18	-0.15
desviación	0.17	25.49	1.53	13.26	2.41	391.04	9.18	7.12	0.65
mínimo	0.0	1.0	617.3	830.8	845.2	0.0	-38.7	-19.1	-1.00
25%	0.0	17.0	626.2	1,017.8	1,035.0	0.0	-16.6	1.5	-0.75
50%	0.0	31.0	627.3	1,028.8	1,036.5	5.5	-10.0	6.0	-0.28
75%	0.0	56.6	628.3	1,037.2	1,038.0	574.0	-1.80	11.9	0.42
máximo	3.7	99.0	632.7	1,060.0	1,044.7	1,634.0	8.6	23.2	1.00

	sin dd_{10m}	ff_{10m}	cos dd_{2min}	sin dd_{2min}	ff_{2min}	cos dd_{2m}	sin dd_{2m}	ff_{2m}	Strokes
cantidad	58,864	58,864	60,295	60,295	60,295	24,889	24,889	24,889	75,844
media	-0.33	6.21	-0.11	-0.32	6.15	0.22	-0.32	4.34	-0.93
desviación	0.66	5.41	0.68	0.65	5.53	0.62	0.68	3.80	0.37
mínimo	-1.00	0	-1.00	-1.00	0	-1.00	-1.00	0	-1
25%	-0.88	2	-0.75	-0.87	2	-0.28	-0.95	2	-1
50%	-0.59	4	-0.22	-0.56	4	0.21	-0.56	3	-1
75%	0.03	9	0.52	0.00	9	0.91	0.00	7	-1
máximo	1.00	29	1.00	1.00	33	1.00	1.00	21	1

Tabla 3.3: Estadísticas descriptivas para Visviri, Chile

	Cardinalidad	**Asimetría**	**Curtosis**
RR	171	9.26	111.58
HR	2292	0.66	-0.75
QFE	283	-0.17	-0.03
QFF	639	0.03	0.23
QNH	162	-8.59	669.58
RadGInst	14530	1.08	-0.23
Td	446	-0.18	-0.82
Ts	903	-0.24	-0.46
cos dd_{10m}	323	0.39	-1.24
sin dd_{10m}	330	0.84	-0.75
ff_{10m}	30	1.31	1.09
cos dd_{2min}	323	0.34	-1.33
sin dd_{2min}	330	0.79	-0.74
ff_{2min}	34	1.33	1.29
cos dd_{2m}	323	-0.22	-1.13
sin dd_{2m}	330	0.61	-0.99
ff_{2m}	22	0.90	0.13
Strokes	2	5.02	23.19

Tabla 3.4: Cardinalidad, asimetría y curtosis de las características de Visviri

Continuando, se realiza un *análisis explicativo de las características numéricas* con el propósito de encontrar las relaciones causa-efecto entre cada una de las características observadas. Un conjunto básico de estadísticas para desarrollar una comprensión sobre las características numéricas es mediante el *análisis multivariable* y la *correlación entre variables.*

En la figura 3.1 se puede ver el mapa de calor con las correlaciones entre cada una de las características registradas para la localidad de Visviri, Chile.

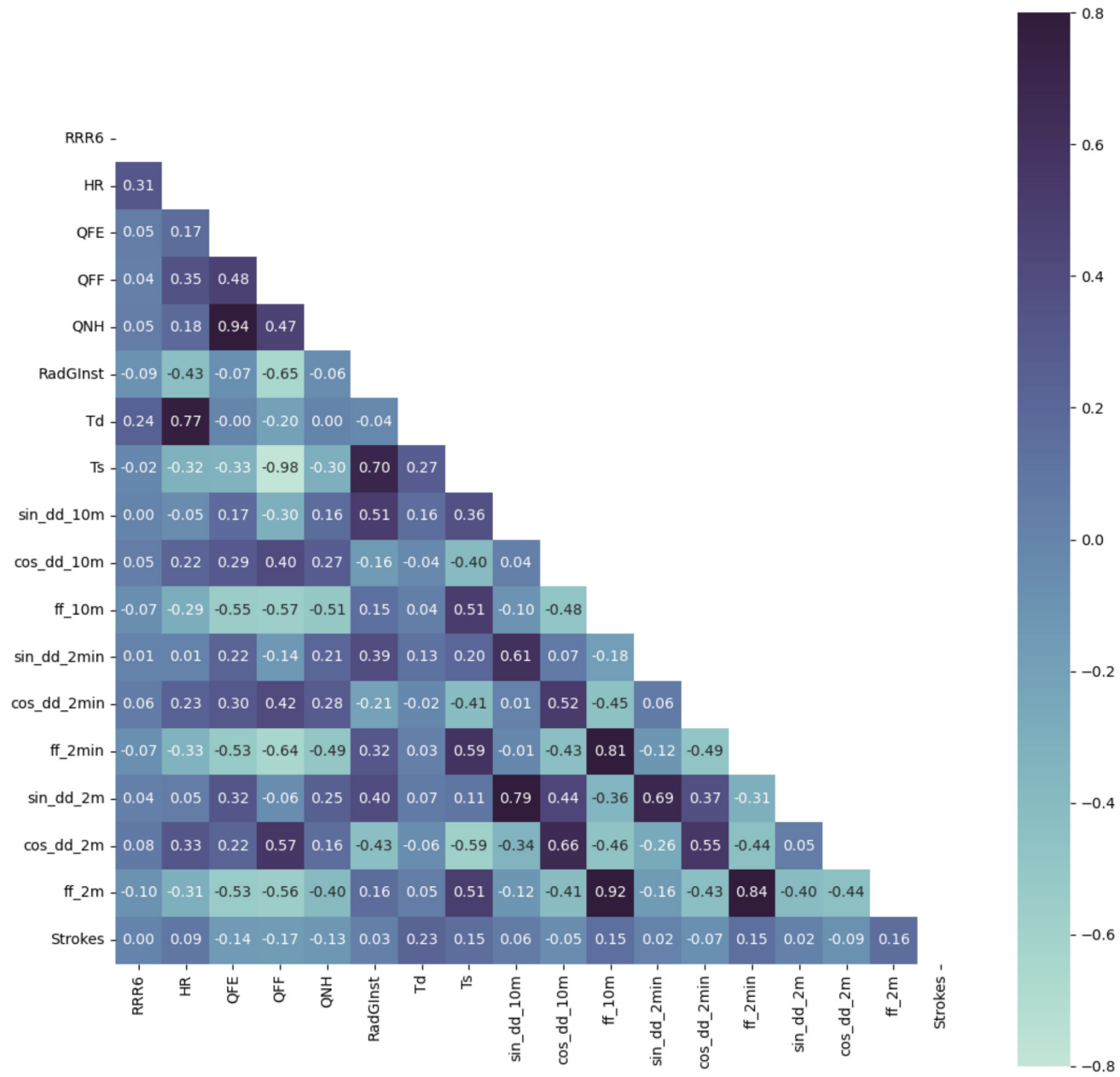


Figura 3.1: Mapa de calor para Visviri, Chile

3.3 Tratamiento previo del conjunto de datos

Realizar un tratamiento previo al conjunto de datos permite transformar los datos a un formato que pueda ser entendible por las máquinas de aprendizaje. Pues, los datos crudos no son comparables, excepto dentro de la misma característica. Además, pueden contener errores e inconsistencias o estar incompletos. De esta forma, el tratamiento de valores incoherentes y no disponibles.

3.3.1 Valores atípicos

Los valores atípicos se deben de tratar cuidadosamente pues afectan los modelos a la vez que pueden ser una valiosa fuente de información. Brindan información sobre comportamientos específicos. La principal preocupación es establecer un criterio que defina lo que es una observación informativa respecto de lo que es un valor desechable o reemplazable por una mejor aproximación. Si bien no existe un criterio universal para identificar cuando un valor deja de aportar información, la forma en que se aborda este problema consistió en realizar la observación de los diagramas de cajas y ajustar un número de veces la desviación estándar de la característica como umbral. Este fue un proceso heurístico. Complementariamente se

realizó el trazado de los registros en el tiempo y se observó cuando un valor estaba fuera de los registros normales. Los valores no disponibles serán abordados más tarde en la sección 3.3.3.

Para el caso de la localidad de Visviri, debido a que demasiados valores registrados tienen registros de magnitud cero resulta que, para aislar los valores atípicos la desviación estándar debe aumentarse varias veces. En este caso, los valores mantenidos fueron los que se encuentran entre los intervalos 0.1 % y 99.9 %. De esta forma, solo existió 1 valor atípico que fue reemplazado mediante el valor medio entre las horas colindantes porque los cambios de presión son procesos de algunas horas y en este caso en 1 hora se produjo una variación de 2.8 veces por debajo de la variación de presión QFF que ocurre durante 1 día promedio[7].

3.3.2 Ingeniería de características

En esta sección se usan las características existentes para crear nuevas características que sean de utilidad tanto para aportar nueva información, como para ayudar en el proceso de imputación de los datos no disponibles mediante el conocimiento del contexto del conjunto de los datos (dominio de la información).

Debido a que nuestro país está sometido a un constante viento del Oeste es importante generar esta nueva característica[8]. Sin embargo, en el caso de la localidad de Visviri, existen 3 subíndices de características asociadas al viento.

Los datos de viento que son obtenidos a mayor altura son de mejor calidad, por lo tanto, una estrategia apropiada es darle todo el peso a esto datos de mejor calidad frente a los otros 2. Y en caso de que ambos restantes puedan imputar el espacio disponible, el valor de consenso es la media entre estos valores registrados.

Continuando con la ingeniería, se observó que los datos obtenidos son registrados a lo largo del año por lo que sería interesante rescatar cada estación del año[9] al aproximar las fechas de los solsticios y equinoccios a la hora más cercana, también la semana del año y la hora del día puede ser una opción que mejore el desempeño del modelo. Ya que las 2 últimas nuevas características presentan tanto una cardinalidad muy alta como una periodicidad entre sus extremos, se decidió descomponerlas en un plano cartesiano.

Y para finalizar, se realizó una limpieza de características ya que, existen características que debido a que están contenidas en otras y para que no generen una doble correlación al realizar el entrenamiento del algoritmo, serán eliminadas dejando las que son de una mejor calidad. En general, la Td está compuesta por Ts y HR, por ejemplo, cuando Td es igual a la Ts, significa que HR es 100 %. Entonces para evitar la duplicidad de información, no se usó Td siempre que pueda rescatar HR. También ocurre algo similar con la presión, pues la QFF depende de la QFE y la Ts y la Presión atmosférica a nivel del mar mediante Atmósfera Estándar de la OACI (QNH) no se recomienda en análisis meteorológicos porque no considera las variaciones de temperatura.

3.3.3 Datos no disponibles

Tras el tratamiento previo de los datos para la localidad, continúa la imputación de los datos no disponibles. En la figura 3 del Anexo, puede observar cómo quedaría la cantidad de valores no disponibles para la localidad de Visviri tras realizar la ingeniería de características. En la tabla 3.5 puede ver el porcentaje de datos no disponibles para cada una de las características ordenadas en orden descendente para la localidad de Visviri, Chile.

El tratamiento de los datos no disponibles es toda un área de estudio que para los alcances de este trabajo será abordado de la siguiente manera.

[7] Para estos días las presiones QFF variaban entre 1000 y 1060 hPa. La presión atípica fue de 830.8 hPa.
[8] También se adiciona la componente Norte-Sur por completitud.
[9] Mediante una técnica llamada One-Hot Encoding.

	Cantidad de datos no disponibles	Porcentaje (%)
QFF	18054	23.80
QNH	16983	22.40
Td	16980	22.40
Ts	3920	5.20
HR	3919	5.20
QFE	3918	5.20
EO	3917	5.20
NS	3917	5.20
ff	3917	5.20
$\cos_{dd}$	3917	5.20
$\sin_{dd}$	3917	5.20
RadGInst	2427	3.20
RR	1492	2.00

Tabla 3.5: Porcentaje de las características no disponibles de Visviri, Chile

En primer lugar, es de relevancia proponer algún mecanismo que esté detrás de la pérdida de datos. Estos mecanismos pueden clasificarse como:

- Datos no disponibles completamente aleatorios: La razón a la falta de datos es ajena a los datos mismos. La falta de datos no depende de la fecha o alguna otra característica.
- Datos no disponibles no aleatorios: La razón a la falta de datos depende de los datos mismos. Si existiera un umbral en el que no se registran datos.
- Datos no disponibles aleatorios: Punto intermedio entre los elementos anteriores. La razón de la falta de datos no depende de los mismos datos no disponibles, pero si puede depender de otras características dentro del conjunto de datos. Este es el caso de Visviri, porque para una fila hay otras columnas que tampoco tienen datos.

Para ello, se puede realizar un gráfico de barras que indique las horas en las que mayor falta de información hubo, tal como se puede observar en la figura 4 del Anexo. Algunas de las hipótesis propuestas que provocaron esta falta de información puede ser porque el sensor que registra los datos se desconectó del servidor, la persona que anotó los datos se equivocó, etc.

Para abordar el tratamiento de los valores no disponibles la estrategia debe ser más personalizada, pues si bien tampoco existe un criterio universal para imputar valores no disponibles, no es recomendable aplicar una estrategia genérica como si se puede usar en 3.3.1. De esta forma, se volvió a establecer un orden jerárquico para realizar el llenado.

En primer lugar, para imputar los valores no disponibles por más de una semana consecutiva en medio de un conjunto de valores (como el caso de la radiación para Visviri donde durante 5 semanas debido a un error desconocido no se midieron valores de radiación entre nov-2015 y dic-2015) se consideró el valor medio entre los datos del año anterior y posterior.

En segundo lugar, utilizando el valor del día anterior a la misma hora se imputaron casi todas las características excepto aquellas que la persistencia puede hacer que el conjunto de datos pierda calidad al mantener un valor que no es sostenible, por ejemplo, una lluvia esporádica transformarla en un día extra de precipitaciones.

Y, en tercer lugar, para completar la imputación de valores no disponibles utilicé la imputación de datos con los k-vecinos más cercanos. Esta forma de imputar datos es bastante útil cuando no se puede garantizar que existe correlación entre los datos (Troyanskaya y cols., 2001).

En la figura 3.2 se observa 4 alternativas utilizando esta técnica donde heurísticamente escogí la que seguía manteniendo la forma de los datos.

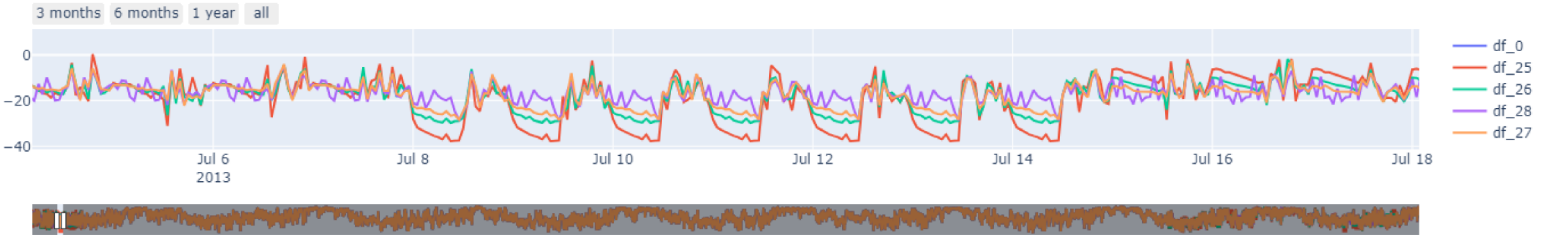


Figura 3.2: Diferentes formas de imputar datos para Visviri, Chile

3.3.4 Estandarización del conjunto de datos

El objetivo de esta sección será establecer un criterio para normalizar los datos pues es fundamental evitar el sobre-ajuste. De forma heurística se fue evaluando las características del conjunto de datos creado y optar por aplicar una estandarización estándar a las características en que sus valores extremos no superen 3 veces su desviación estándar y las demás con una estandarización min-max. La idea perseguida fue garantizar que todas las características estuvieran en una escala similar, ya que muchos algoritmos de aprendizaje automático son sensibles a la escala de las características de entrada. De esta forma para la localidad de Visviri, las estandarizaciones fueron las que observa en la tabla 3.6.

Característica	Tipo de estandarización
RR	minmax
QFE	minmax
RadGInst	minmax
ff	minmax
NS, EO	minmax
sin_hour, cos_hour	standard
HR	standard
Ts	standard
sin_dd, cos_dd	standard
sin_week, cos_week	standard
autumn, winter, spring, summer	standard

Tabla 3.6: Estandarizaciones aplicadas

3.4 Análisis de series temporales

Una vez obtenido un conjunto de datos completo, el primer paso en cualquier investigación de series temporales implica un examen de los datos registrados a lo largo del tiempo. Este escrutinio suele sugerir el método de análisis, así como las estadísticas que serán útiles para resumir la información de los datos. Antes de examinar más detenidamente los métodos estadísticos concretos, conviene mencionar que existen dos enfoques distintos, aunque no necesariamente excluyentes. Del análisis de series temporales, comúnmente identificados como enfoque del *dominio del tiempo* y el enfoque del *dominio de la frecuencia*. El enfoque temporal considera que la investigación de las relaciones desfasadas es lo más importante (por ejemplo, cómo afecta lo que ha ocurrido hoy a lo que ocurrirá mañana), mientras que el enfoque frecuencial considera que la investigación de los ciclos es lo más importante (por ejemplo, cuál es el periodo de tiempo a través de los cuales se observa una periodización de los datos).

Se dice que los datos de series temporales son estacionarios cuando las propiedades estadísticas como la media, la desviación estándar son constantes y no hay estacionalidad.

En otras palabras, las propiedades estadísticas de los datos de la serie temporal no deben ser función del tiempo. Para probar que los datos son estacionarios, lo más sencillo es observar el gráfico y buscar cualquier tendencia o estacionalidad obvia. Pero un método más sofisticado es realizar la prueba de Dickey-Fuller aumentada[10] para comprobar la estacionariedad de los datos.

Sin entrar en tanto detalle sobre el funcionamiento de esta prueba. Se realizó la interpretación del resultado para determinar la estacionariedad de la serie. Esta prueba devuelve valores de salida *p-value* y una *estadística de prueba*. Cuanto más negativa sea la estadística de prueba, más probable es que la serie sea estacionaria. Además, este valor debe ser inferior a los valores críticos (1 %, 5 %, 10 %). Por ejemplo, si el estadístico de prueba es inferior a los valores críticos del 5 %, se puede afirmar con un 95 % de confianza que se trata de una serie estacionaria. Y en cuanto al valor de p, si el valor de $p > 0.05$, la serie no es estacionaria, y si el valor de $p \leq 0.05$, la serie es estacionaria. En la tabla 3.7 puede ver los resultados de aplicar la prueba de Dickey-Fuller aumentada al conjunto de datos rellenado de Visviri.

	Estadística ADF	**1 %**	**5 %**	**10 %**	**p-value**
RR	-23.642294	-3.430	-2.862	-2.567	0
HR	-14.144363	-3.430	-2.862	-2.567	0
QFE	-18.183444	-3.430	-2.862	-2.567	0
RadGInst	-19.712638	-3.430	-2.862	-2.567	0
Ts	-11.308407	-3.430	-2.862	-2.567	0
cos dd	-20.092940	-3.430	-2.862	-2.567	0
sin dd	-25.569798	-3.430	-2.862	-2.567	0
ff	-20.238601	-3.430	-2.862	-2.567	0
NS	-25.323025	-3.430	-2.862	-2.567	0
EO	-21.460487	-3.430	-2.862	-2.567	0
Strokes	-19.614082	-3.430	-2.862	-2.567	0

Tabla 3.7: Resultados de la prueba de Dickey-Fuller aumentada aplicada a las características de Visviri, Chile

Para finalizar esta etapa, se realizaron los gráficos de autocorrelación para determinar la cantidad de rezagos que hacer con el propósito de estudiar si estos consiguen mejorar el desempeño del modelo. En la figura 5 del Anexo puede encontrar las gráficas de la función de autocorrelación del conjunto de datos.

3.5 Dividir el conjunto de datos

A diferencia de los conjuntos de datos tradicionales donde cada ejemplo es independiente del anterior, en series temporales el comportamiento del pasado viene a ser una fuente de información, por lo que la distribución de los conjuntos de prueba y de entrenamiento no se recomienda que sea aleatoria. De esta forma, y debido a la cantidad de datos descargados, el último año fue destinado para probar el modelo y todos los años restantes destinados a crear el modelo de aprendizaje. Para el caso de Visviri, las características finales son 19[11], resultando 67084 ejemplos para crear el modelo y 8760 para probar el modelo[12]

[10] La prueba de Dickey-Fuller aumentada es una prueba de raíz unitaria (característica de los procesos que evolucionan a través del tiempo y que puede causar problemas en inferencia estadística en modelos de series de tiempo) para una muestra de una serie de tiempo.

[11] sin_hour, cos_hour, sin_week, cos_week, autumn, winter, spring, summer, RR, HR, QFE, RadGInst, Ts, sin_dd, cos_dd, ff, NS, EO y Strokes.

[12] Aunque este número posteriormente será reducido a 365 para hacerlos más interpretables dado el objetivo perseguido es encontrar un modelo de predicción para los días de tormenta.

La estrategia consistió en probar principalmente 2 bibliotecas: Scikit-learn y TensorFlow. Luego, escoger la mejor máquina mediante validación cruzada k-fold estratificada para diferentes cantidades de rezagos. A cada una de las máquinas las denominé *modelos base*, esto con el propósito de posteriormente realizar un ajuste de hiper-parámetros.

La función de pérdida escogida fue la utilizada en la literatura revisada, esta es, el error cuadrático medio. La función de pérdida nos dirá qué tan cerca se aproxima el modelo a los datos.

3.6 Selección del modelo

Sin entrar en tanto detalle, las máquinas de aprendizaje automático de Scikit-learn que se consultaron fueron las siguientes:

- AdaBoostClassifier
- BaggingClassifier
- ExtraTreesClassifier
- GradientBoostingClassifier
- HistGradientBoostingClassifier
- RandomForestClassifier
- LogisticRegression
- LogisticRegressionCV
- PassiveAggressiveClassifier
- Perceptron
- SGDClassifier
- BernoulliNB
- ComplementNB
- GaussianNB
- MultinomialNB
- LinearSVC
- NuSVC
- SVC
- KNeighborsClassifier
- NearestCentroid
- RadiusNeighborsClassifier
- DecisionTreeClassifier
- ExtraTreeClassifier
- MLPClassifier
- XGBClassifier

Y las máquinas de aprendizaje automático de TensorFlow que se consultaron fueron las siguientes:

- Conv1D
- Conv1DTranspose
- DepthwiseConv1D
- Dense
- GRU
- LSTM
- LocallyConnected1D
- SeparableConv1D
- SimpleRNN

Se entrenaron cada uno de los modelos y se registraron los errores tipo I y II, la puntuación F1 y el tiempo de total de entrenamientos para luego seleccionar los que mejor desempeño tuvieron. El método de selección fue la media de las puntuaciones F1 asociadas al K-Fold estratificado. El K-Fold estratificado es útil cuando los datos utilizados presentan un desequilibrio en la distribución de la clase objetivo (Widodo, Brawijaya, y Samudi, 2022). En cuanto a los modelos de TensorFlow, se definió un numero de 100 épocas de entrenamiento como máximo pues se utilizó un EarlyStoppping como callback. El EarlyStoppping se refiere al proceso de modificar un modelo para evitar el sobreajuste al imponer algún tipo de restricción (Girosi, Jones, y Poggio, 1995).

3.7 Optimización del mejor modelo base

Para acercarse a la capacidad de generalización. Luego de desarrollar el modelo, se realizó una búsqueda de los hiper-parámetros que mejoran el éxito del algoritmo mediante la métrica F1 mediante Greed search dentro de un conjunto muy acotado de hiper-parámetros, pues se evidenciaron las limitaciones de hardware y tiempo disponible para realizar búsquedas más profundas.

Posteriormente se realizó un *ablation* o análisis de sensibilidad para realizar un ajuste fino para crear un mejor modelo.

Para validar el modelo se utilizó el año 2021. Además, dado que el objetivo es encontrar los días de tormenta, se realizó un post-procesamiento de la información. Esta consistió en volver a muestrear el conjunto de prueba cada 24 horas, entregando el valor máximo y evaluando el modelo en la matriz de confusión y mediante la puntuación F1.

4 Resultados y análisis

Índice general

4.1 Selección de la ubicación

Los resultados obtenidos sobre la localidad con mayor actividad eléctrica y atmosférica apuntaron a las coordenadas geográficas -17.594999, -69.477499 y -17.594721, -69.475277 dentro de Visviri[1]. En la figura 4.1 se puede ver una vista satélite con las ubicaciones de las estaciones. Las EMA de “Visviri Tenencia” y “Visviri INIA” se encuentran separadas 243 m entre sí y se están ubicadas a unas altura de 4084 y 4122 msnm, respectivamente.

Figura 4.1: Ubicación de las estaciones meteorológicas seleccionadas

[1] Visviri es la capital de la comuna de General Lagos, en la provincia de Parinacota, región de Arica y Parinacota. Con 154 personas, es el pueblo más septentrional de Chile.

La intersección de ambos conjuntos de datos creó un conjunto de 75.844 ejemplo y 19 características[2] para el pueblo de Visviri, Chile. Los datos fueron registrados desde inicios de mayo de 2013 hasta finales de diciembre de 2021. El conjunto de datos se encuentra desequilibrado, pues existe un 3 % de descargas registradas durante todo el tiempo en que fueron recopilados los datos. Y, en promedio, se perdieron alrededor de un 15 % de los datos.

4.2 Análisis exploratorio de datos

4.2.1 Análisis descriptivo para las características temporales

Para determinar las épocas de alta o baja cantidad de horas de tormenta, se procedió con el análisis descriptivo de las características asociadas a la descomposición de las horas y las semanas del conjunto de datos. Se puede ver en las tablas 4.1, 4.2, 4.3 y 4.4 que, en promedio, las tormentas ocurren durante el último cuarto del día mientras que los cielos despejados ocurren a cualquier hora del día y en cualquier semana del año. Esto tiene sentido pues las tormentas son eventos esporádicos que suelen ocurrir durante las tardes de verano. Durante el verano, las tormentas ocurren un 10.01 % del tiempo. El resto de porcentajes son 1.98 %, 0.47 y 2.56 % durante otoño, invierno y primavera, respectivamente.

Al recopilar los datos de los registros de tormenta en función de la hora del día para cada una de las estaciones del año, puede observar que la mayor cantidad de descargas ocurren durante el verano, tal como puede ver en la figura 4.2 para la localidad de Visviri, Chile. Los datos se encuentran en hora GMT.

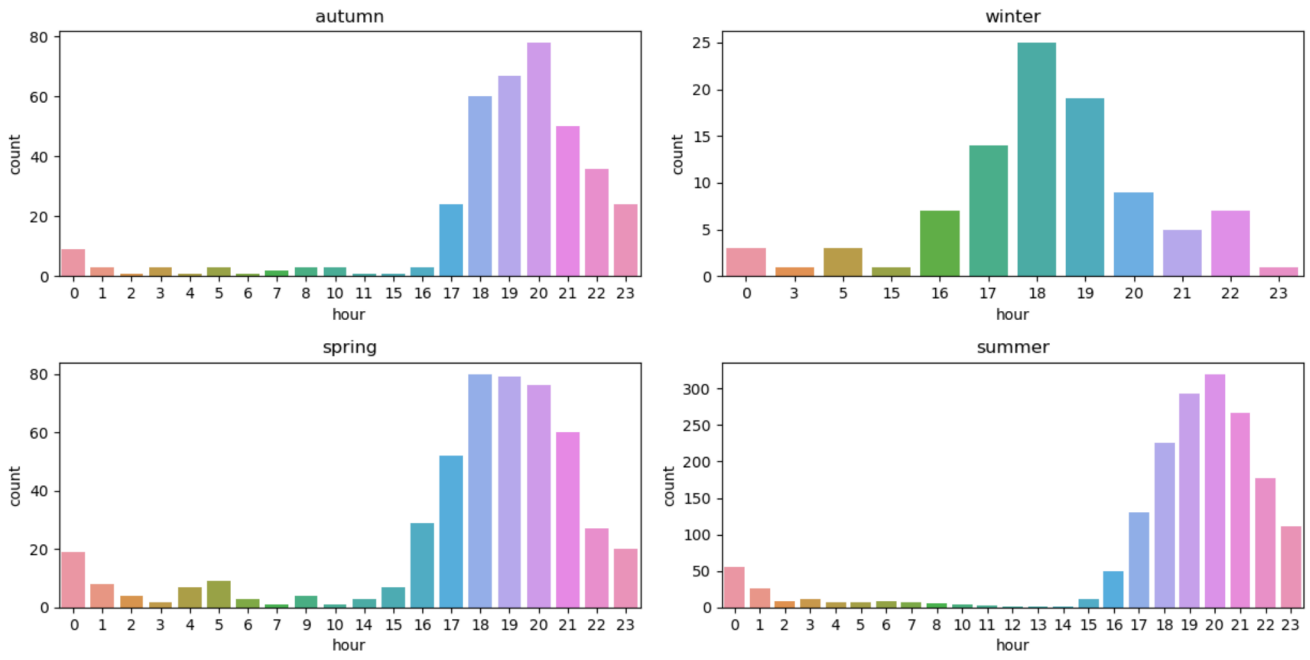


Figura 4.2: Distribución horaria (en hora GMT) sobre la ocurrencia de las tormentas para cada estación del año.

En las tablas 4.1, 4.2, 4.3 y 4.4 se puede ver las estadísticas descriptivas que resumen la cantidad, tendencia central y variabilidad de los datos para tormentas y cielos despejados durante los días de otoño, invierno, primavera y verano para la localidad de Visviri, respectivamente. Donde durante el otoño, el invierno, la primavera y el verano, en promedio, las tormentas también ocurren durante el último cuarto del día, y los cielos despejados también a cualquier hora del día.

[2] sin_hour, cos_hour, sin_week, cos_week, autumn, winter, spring, summer, RR, HR, QFE, RadGInst, Ts, sin_dd, cos_dd, ff, NS, EO y Descarga.

Tormentas eléctricas de otoño

	cantidad	media	desviación	mínimo	25%	50%	75%	máximo
sin_hour	373	-0.71	0.42	-1.00	-0.97	-0.87	-0.71	1.00
cos_hour	373	0.40	0.40	-0.97	0.00	0.50	0.71	1.00
sin_week	373	0.92	0.16	0.12	0.89	0.99	0.99	1.00
cos_week	373	-0.22	0.29	-0.99	-0.46	-0.12	0.00	0.24
autumn	373	1.00	0.00	1	1	1	1	1
winter	373	0.00	0.00	0	0	0	0	0
spring	373	0.00	0.00	0	0	0	0	0
summer	373	0.00	0.00	0	0	0	0	0
RR	373	0.03	0.13	0.0	0.0	0.0	0.0	1.7
HR	373	49.76	18.42	9.0	34.5	48.0	63.2	91.0
QFE	373	626.72	1.09	623.6	626.0	626.8	627.5	630.2
RadGInst	373	262.35	345.76	0.0	23.6	108.8	397.8	1,277.6
Ts	373	10.97	4.10	-6.5	8.8	11.4	13.7	18.9
sin_dd	373	-0.21	0.67	-1.00	-0.74	-0.48	0.42	1.00
cos_dd	373	-0.20	0.68	-1.00	-0.82	-0.47	0.50	1.00
ff	373	8.60	4.57	1	5	8	11	24
NS	373	-2.36	5.55	-11.59	-6.99	-3.33	2.30	15.13
EO	373	-2.65	7.10	-21.75	-7.08	-2.54	2.27	15.38
Descarga	373	1.00	0.00	1	1	1	1	1

Cielos despejados de otoño

	cantidad	media	desviación	mínimo	25%	50%	75%	máximo
sin_hour	18500	0.01	0.70	-1.00	-0.71	0.00	0.71	1.00
cos_hour	18500	-0.01	0.71	-1.00	-0.71	0.00	0.71	1.00
sin_week	18500	0.69	0.28	0.12	0.46	0.75	0.97	1.00
cos_week	18500	-0.57	0.35	-0.99	-0.89	-0.66	-0.24	0.24
autumn	18500	1.00	0.00	1	1	1	1	1
winter	18500	0.00	0.00	0	0	0	0	0
spring	18500	0.00	0.00	0	0	0	0	0
summer	18500	0.00	0.00	0	0	0	0	0
RR	18500	0.01	0.10	0.0	0.0	0.0	0.0	1.9
HR	18500	37.79	23.82	2.0	18.1	32.0	53.9	97.9
QFE	18500	627.85	1.36	621.6	626.9	627.9	628.8	632.7
RadGInst	18500	249.31	348.73	0.0	0.0	0.2	504.3	1,634.0
Ts	18500	4.74	7.20	-16.3	-0.5	4.6	10.8	21.7
sin_dd	18500	-0.38	0.63	-1.00	-0.91	-0.62	0.00	1.00
cos_dd	18500	-0.01	0.67	-1.00	-0.67	0.00	0.60	1.00
ff	18500	4.98	4.53	0	2	4	7	26
NS	18500	-2.37	3.90	-22.0	-4.9	-1.9	0.0	12.7
EO	18500	-1.61	4.63	-25.0	-2.7	0.0	0.8	17.9
Descarga	18500	-1.00	0.00	-1	-1	-1	-1	-1

Tabla 4.1: Estadísticas descriptivas durante otoño para Visviri, Chile

Tormentas de invierno

	cantidad	media	desviación	mínimo	25%	50%	75%	máximo
sin_hour	95	-0.79	0.42	-1.00	-1.00	-0.97	-0.87	0.97
cos_hour	95	0.18	0.41	-0.71	0.00	0.26	0.50	1.00
sin_week	95	-0.80	0.29	-1.00	-0.97	-0.94	-0.75	0.00
cos_week	95	-0.44	0.30	-1.00	-0.66	-0.35	-0.24	0.00
autumn	95	0.00	0.00	0	0	0	0	0
winter	95	1.00	0.00	1	1	1	1	1
spring	95	0.00	0.00	0	0	0	0	0
summer	95	0.00	0.00	0	0	0	0	0
RR	95	0.01	0.05	0.0	0.0	0.0	0.0	0.4
HR	95	39.60	18.20	11.4	25.6	37.8	48.0	91.0
QFE	95	627.05	1.14	623.3	626.2	627.3	627.8	629.5
RadGInst	95	330.81	369.58	0.0	47.9	167.4	486.9	1,247.0
Ts	95	10.05	3.84	-2.4	8.0	10.8	12.5	18.6
sin_dd	95	-0.14	0.69	-1.00	-0.73	-0.36	0.51	1.00
cos_dd	95	0.10	0.69	-1.00	-0.52	0.03	0.81	1.00
ff	95	7.82	3.93	0	5	7	10	18
NS	95	-1.91	5.82	-17.99	-5.81	-1.91	2.47	12.90
EO	95	0.14	6.09	-15.74	-3.55	0.07	4.82	12.50
Descarga	95	1.00	0.00	1	1	1	1	1

Cielos despejados de invierno

	cantidad	media	desviación	mínimo	25%	50%	75%	máximo
sin_hour	20133	0.00	0.71	-1.00	-0.71	0.00	0.71	1.00
cos_hour	20133	0.00	0.71	-1.00	-0.71	0.00	0.71	1.00
sin_week	20133	-0.57	0.34	-1.00	-0.89	-0.66	-0.24	0.12
cos_week	20133	-0.69	0.29	-1.00	-0.97	-0.75	-0.46	0.00
autumn	20133	0.00	0.00	0	0	0	0	0
winter	20133	1.00	0.00	1	1	1	1	1
spring	20133	0.00	0.00	0	0	0	0	0
summer	20133	0.00	0.00	0	0	0	0	0
RR	20133	0.00	0.04	0.0	0.0	0.0	0.0	1.3
HR	20133	26.49	18.40	1.9	12.5	21.9	35.0	97.6
QFE	20133	627.64	1.42	622.2	626.7	627.7	628.7	632.6
RadGInst	20133	260.51	353.41	0.0	0.0	0.2	561.3	1,600.0
Ts	20133	3.55	7.87	-19.1	-2.6	3.3	10.5	21.5
sin_dd	20133	-0.40	0.63	-1.00	-0.92	-0.66	0.00	1.00
cos_dd	20133	0.00	0.65	-1.00	-0.63	0.02	0.59	1.00
ff	20133	5.26	4.59	0	2	4	7	28
NS	20133	-2.69	4.21	-25.00	-5.16	-1.97	0.00	14.42
EO	20133	-1.43	4.58	-27.66	-2.70	0.00	0.97	21.55
Descarga	20133	-1.00	0.00	-1	-1	-1	-1	-1

Tabla 4.2: Estadísticas descriptivas durante invierno para Visviri, Chile

Tormentas de primavera

	cantidad	media	desviación	mínimo	25%	50%	75%	máximo
sin_hour	491	-0.69	0.48	-1.00	-0.97	-0.87	-0.71	1.00
cos_hour	491	0.29	0.46	-0.87	0.00	0.26	0.71	1.00
sin_week	491	-0.58	0.30	-1.00	-0.89	-0.57	-0.35	0.00
cos_week	491	0.70	0.30	-0.12	0.46	0.82	0.94	1.00
autumn	491	0.00	0.00	0	0	0	0	0
winter	491	0.00	0.00	0	0	0	0	0
spring	491	1.00	0.00	1	1	1	1	1
summer	491	0.00	0.00	0	0	0	0	0
RR	491	0.02	0.09	0	0	0	0	0.6
HR	491	38.47	20.57	2.0	22.1	33.0	51.9	90.8
QFE	491	625.91	1.36	622.8	625.0	625.9	626.8	630.3
RadGInst	491	375.69	408.07	0.0	53.4	221.2	601.6	1,586.0
Ts	491	12.34	4.81	-8.9	9.3	12.9	15.8	22.1
sin_dd	491	-0.23	0.67	-1.00	-0.78	-0.48	0.40	1.00
cos_dd	491	-0.23	0.66	-1.00	-0.84	-0.42	0.33	1.00
ff	491	10.23	5.68	0	6	9	14	28
NS	491	-2.85	6.58	-15.96	-8.12	-3.94	2.67	15.59
EO	491	-3.94	8.37	-25.58	-9.12	-2.83	1.78	18.54
Descarga	491	1.00	0.00	1	1	1	1	1

Cielos despejados de primavera

	cantidad	media	desviación	mínimo	25%	50%	75%	máximo
sin_hour	18917	0.02	0.70	-1.00	-0.71	0.00	0.71	1.00
cos_hour	18917	-0.01	0.71	-1.00	-0.71	0.00	0.71	1.00
sin_week	18917	-0.69	0.28	-1.00	-0.94	-0.75	-0.46	0.00
cos_week	18917	0.58	0.33	-0.12	0.35	0.66	0.89	1.00
autumn	18917	0.00	0.00	0	0	0	0	0
winter	18917	0.00	0.00	0	0	0	0	0
spring	18917	1.00	0.00	1	1	1	1	1
summer	18917	0.00	0.00	0	0	0	0	0
RR	18917	0.01	0.11	0	0	0	0	2.183333
HR	18917	30.14	21.62	1.0	13.7	24.0	40.0	99.0
QFE	18917	626.58	1.49	617.3	625.6	626.7	627.6	632.0
RadGInst	18917	335.50	435.06	0.0	0.0	8.2	704.6	1,618.5
Ts	18917	7.56	7.01	-13.3	2.3	6.8	13.7	23.2
sin_dd	18917	-0.36	0.65	-1.00	-0.90	-0.60	0.00	1.00
cos_dd	18917	-0.19	0.64	-1.00	-0.78	-0.33	0.29	1.00
ff	18917	6.88	6.00	0	3	5	10	29
NS	18917	-3.03	4.95	-19.00	-6.64	-2.62	0.00	21.28
EO	18917	-3.25	6.25	-27.73	-5.08	-1.17	0.44	21.99
Descarga	18917	-1.00	0.00	-1	-1	-1	-1	-1

Tabla 4.3: Estadísticas descriptivas durante primavera para Visviri, Chile

	Tormentas de verano							
	cantidad	**media**	**desviación**	**mínimo**	**25%**	**50%**	**75%**	**máximo**
sin_hour	1735	-0.70	0.40	-1.00	-0.97	-0.87	-0.50	1.00
cos_hour	1735	0.41	0.43	-1.00	0.00	0.50	0.71	1.00
sin_week	1735	0.55	0.33	-0.12	0.24	0.66	0.82	0.99
cos_week	1735	0.72	0.26	0.12	0.57	0.75	0.97	1.00
autumn	1735	0.00	0.00	0	0	0	0	0
winter	1735	0.00	0.00	0	0	0	0	0
spring	1735	0.00	0.00	0	0	0	0	0
summer	1735	1.00	0.00	1	1	1	1	1
RR	1735	0.03	0.12	0.0	0.0	0.0	0.0	1.3
HR	1735	54.30	18.75	4.0	39.0	54.0	69.0	94.4
QFE	1735	625.94	1.26	622.5	625.1	625.9	626.8	630.3
RadGInst	1735	359.74	414.40	0.0	37.9	181.3	568.5	1,602.5
Ts	1735	11.90	3.63	-1.7	9.2	12.0	14.7	21.5
sin_dd	1735	-0.14	0.68	-1.00	-0.68	-0.41	0.57	1.00
cos_dd	1735	-0.29	0.66	-1.00	-0.85	-0.56	0.25	1.00
ff	1735	9.89	5.24	0	6	9	13	27
NS	1735	-2.02	6.45	-16.62	-7.48	-3.15	3.43	24.81
EO	1735	-4.25	7.85	-24.54	-9.95	-3.60	1.25	17.00
Descarga	1735	1.00	0.00	1	1	1	1	1

	Cielos despejados de verano							
	cantidad	**media**	**desviación**	**mínimo**	**25%**	**50%**	**75%**	**máximo**
sin_hour	15600	0.08	0.69	-1.00	-0.50	0.26	0.71	1.00
cos_hour	15600	-0.04	0.72	-1.00	-0.71	0.00	0.71	1.00
sin_week	15600	0.55	0.34	-0.12	0.24	0.57	0.89	0.99
cos_week	15600	0.72	0.27	0.12	0.46	0.82	0.97	1.00
autumn	15600	0.00	0.00	0	0	0	0	0
winter	15600	0.00	0.00	0	0	0	0	0
spring	15600	0.00	0.00	0	0	0	0	0
summer	15600	1.00	0.00	1	1	1	1	1
RR	15600	0.10	0.33	0.0	0.0	0.0	0.0	3.7
HR	15600	59.69	24.64	3.0	39.0	63.2	82.0	97.3
QFE	15600	626.94	1.38	621.0	626.0	627.0	627.9	631.4
RadGInst	15600	281.92	413.40	0.0	0.0	0.4	514.9	1,631.0
Ts	15600	7.80	4.94	-7.0	4.4	6.7	11.2	22.5
sin_dd	15600	-0.24	0.71	-1.00	-0.87	-0.52	0.47	1.00
cos_dd	15600	-0.11	0.65	-1.00	-0.71	-0.21	0.44	1.00
ff	15600	5.65	4.92	0	2	4	7	28
NS	15600	-1.71	4.68	-17.02	-4.81	-1.62	1.17	20.92
EO	15600	-2.02	5.21	-25.21	-3.21	-0.65	0.70	21.25
Descarga	15600	-1.00	0.00	-1	-1	-1	-1	-1

Tabla 4.4: Estadísticas descriptivas durante verano para Visviri, Chile

4.2.2 Análisis descriptivo para las características meteorológicas

El objetivo de esta sección es describir cada una de las características meteorológicas registradas que favorecen la creación de tormentas. Una condición necesaria pero no suficiente para la formación de tormentas es la presencia de nubes, por lo que cada análisis se enfocó en resaltar esta característica.

Respecto de la precipitación, en promedio, fue de 0.03 mm cuando hay tormenta y también cuando hay cielos despejados. Esto es relevante, pues existe el sentido de que cuando precipita, ocurre tormenta, pero las medias indican que no existe una relación del agua precipitada respecto de las condiciones que favorezcan las tormentas o los cielos despejados. En la tabla 4.5 se pueden ver que para los valores medios de precipitación durante otoño, invierno y primavera ocurre lo esperado y las tormentas se encuentran más relacionadas con una mayor cantidad de precipitación caída.

	Tormenta (mm)	**Cielo despejado (mm)**
Otoño	0.03	0.01
Invierno	0.01	0.00
Primavera	0.02	0.01
Verano	0.03	0.10

Tabla 4.5: Valores medios de Precipitación para cada estación del año

Respecto de la humedad relativa del aire, en promedio, fue de 50.27 % cuando hay tormenta, mientras que es 37.37 % cuando hay cielos despejados. Esto es de esperar, pues la humedad es un factor relevante para la formación de nubes. En la tabla 4.6 se pueden ver que para los valores medios de humedad durante otoño, invierno y primavera ocurre lo esperado y las tormentas se encuentran más relacionadas con un mayor porcentaje de humedad.

	Tormenta (%)	**Cielo despejado (%)**
Otoño	49.76 %	37.79 %
Invierno	39.60 %	26.49 %
Primavera	28.47 %	30.14 %
Verano	54.30 %	59.69 %

Tabla 4.6: Valores medios de Humedad relativa para cada estación del año

Respecto de la presión, en promedio, fue de 626.08 hPa cuando hay tormenta, mientras que es 627.27 hPa cuando hay cielos despejados. También es esperable, pues la presión disminuye debido a la presencia de nubes, lo que es un factor para la presencia de tormentas. En la tabla 4.7 se pueden ver que para los valores medios de presión durante otoño, invierno, primavera y verano se repite esta misma tendencia.

	Tormenta (hPa)	**Cielo despejado (hPa)**
Otoño	626.72	627.85
Invierno	627.05	627.64
Primavera	625.91	626.58
Verano	625.94	626.94

Tabla 4.7: Valores medios de Presión a nivel de estación para cada estación del año

Respecto de la radiación, esta característica debe estudiarse con mayor cuidado pues, en promedio, cuando hay tormenta es 348.14 W/m^2 mientras que cuando hay cielos despejados es 281.64 W/m^2. A primera vista parecen resultados contradictorios con la formación de nubes, pero al apartar los horarios en los que es de noche, podemos evitar los valores cero

de las noches de invierno que enmascaran los resultados. En promedio, cuando hay tormenta es 388.20 W/m^2 mientras que cuando hay cielos despejados es 522.49 W/m^2. En la tabla 4.8 se pueden ver que para los valores medios de radiación evitando las horas de noche durante otoño, invierno, primavera y verano repiten esta misma tendencia.

	Tormenta (W/m^2)	**Cielo despejado (W/m^2)**
Otoño	302.02	478.59
Invierno	353.11	494.43
Primavera	423.09	583.34
Verano	398.30	529.56

Tabla 4.8: Valores medios de Radiación Solar Instantánea evitando las horas de noche para cada estación del año

Respecto de la temperatura, ocurre algo similar pero esta vez en el sentido contrario, pues el enmascaramiento del promedio hace que los valores durante las horas nocturnas reduzcan el promedio, haciendo que los valores sean mucho mejor de lo esperado pues, en promedio, cuando hay tormenta es 11.78 °C mientras que cuando hay cielos despejados es 5.79 °C. Al apartar los horarios en los que es de noche, en promedio, cuando hay tormenta es 12.42 °C mientras que cuando hay cielos despejados es 9.61 °C. En la tabla 4.9 se pueden ver que para los valores medios de temperatura evitando las horas de noche durante otoño, invierno, primavera y verano se repite esta misma tendencia.

	Tormenta (°C)	**Cielo despejado (°C)**
Otoño	11.83	8.67
Invierno	10.51	8.02
Primavera	13.26	11.23
Verano	12.41	10.62

Tabla 4.9: Valores medios de Temperatura de aire seco evitando las horas de noche para cada estación del año

Respecto del viento, en promedio, fue de 9.70 kt con una dirección suroeste mientras que cuando hay cielos despejados fue de 5.69 kt con una dirección oeste. Esto es esperable pues el viento es un factor, y ya si bien la dirección no cambia tanto, si lo hace la intensidad. En la tabla 4.10 se pueden ver los valores medios de viento para cada estación del año donde durante otoño, invierno, primavera y verano puede ver como el viento del oeste es un factor influyente en las descargas eléctricas.

	Tormenta (kt)	**Cielo despejado (kt)**
Otoño	8.60 kt dirección suroeste	4.98 kt dirección oeste
Invierno	7.82 kt dirección noroeste	5.26 kt dirección oeste
Primavera	10.23 kt dirección suroeste	6.88 kt dirección suroeste
Verano	9.89 kt dirección suroeste	5.65 kt dirección suroeste

Tabla 4.10: Valores medios de magnitud y dirección del Viento evitando las horas de noche para cada estación del año

4.2.3 Análisis del grado de simetría de los datos

Respecto de la simetría, la estadística descriptiva también incluye el grado de simetría de los datos respecto de su medida central y la concentración de los datos alrededor de dicho valor. Sin entrar en tanto detalle, mientras más negativa sea la simetría, más a la derecha se encuentran los datos (media ≤ mediana ≤ moda), y mientras más positiva sea, más a la izquierda

están los datos (media ≥ mediana ≥ moda). La curtosis indica el grado de concentración de los datos en torno al valor central. Si $c > 3$ la distribución es leptocúrtica (con forma de punta), si $c = 3$ es mesocúrtica (con forma de campana) y si $c < 3$ es platicúrtica (con forma plana). En este caso, en algún grado, todas las características se encuentran concentradas hacia la izquierda a excepción de la temperatura, mientras que todas las características son mesocúrticas, a excepción de la precipitación que es leptocúrtica. Estos valores reflejan algunas condiciones, como que durante las noches no hay radiación por lo que se espera que su asimetría se incline hacia la izquierda, al igual que las esporádicas lluvias que no son tan frecuentes en el norte.

4.2.4 Relaciones causa-efecto entre las características

En la figura 4.3 se puede ver una cuadrícula con las relaciones entre las características para cuando ocurre tormenta y cuando hay cielos despejados.

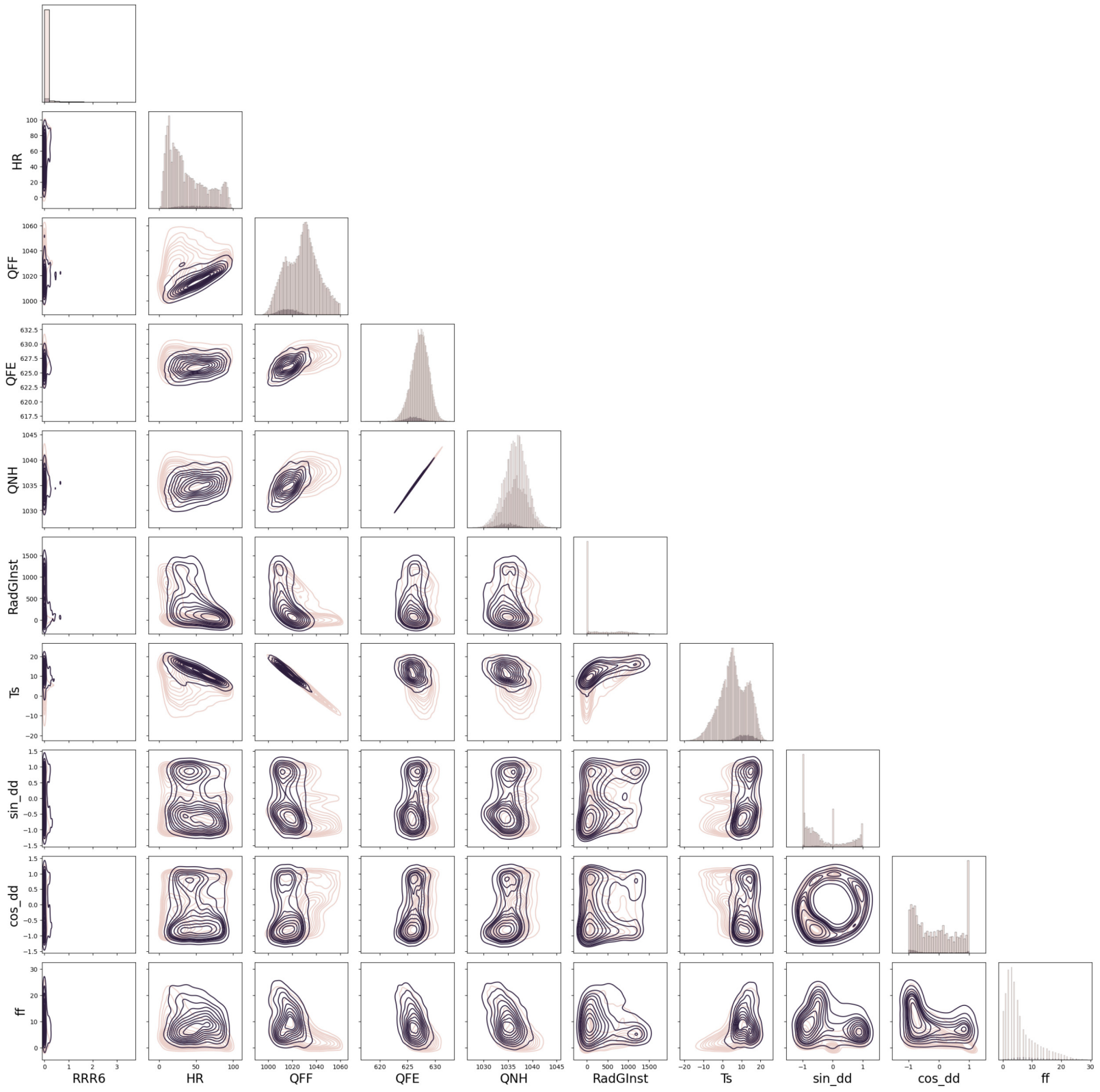


Figura 4.3: Relaciones entre el conjunto de datos cuando ocurre tormenta y cuando hay cielos despejados

Varias de las mejores correlaciones no resultaron informativas, pues evidenciaron las relaciones propias de la descomposición de la dirección del viento con las componentes Norte-Sur y Este-Oeste. Por lo que para evitar aprender dos veces sobre las mismas características,

retiro la descomposición de la dirección. En la figura 4.4 puede ver las correlaciones entre cada una de las características para Visviri, Chile.

Tras realizar este ajuste en la ingeniería de las características, se puede ver en la figura 4.5 el nuevo mapa de correlaciones, donde las mejores correlaciones ocurren entre sin_hour y Temperatura, sin_hour y Radiación y Temperatura y Radiación, pues se refleja el incremento de la Temperatura cuando aparece la luz del Sol a cierta hora del día. Las mejores respecto de los rayos fueron el verano, sin_dd, la temperatura y la presión.

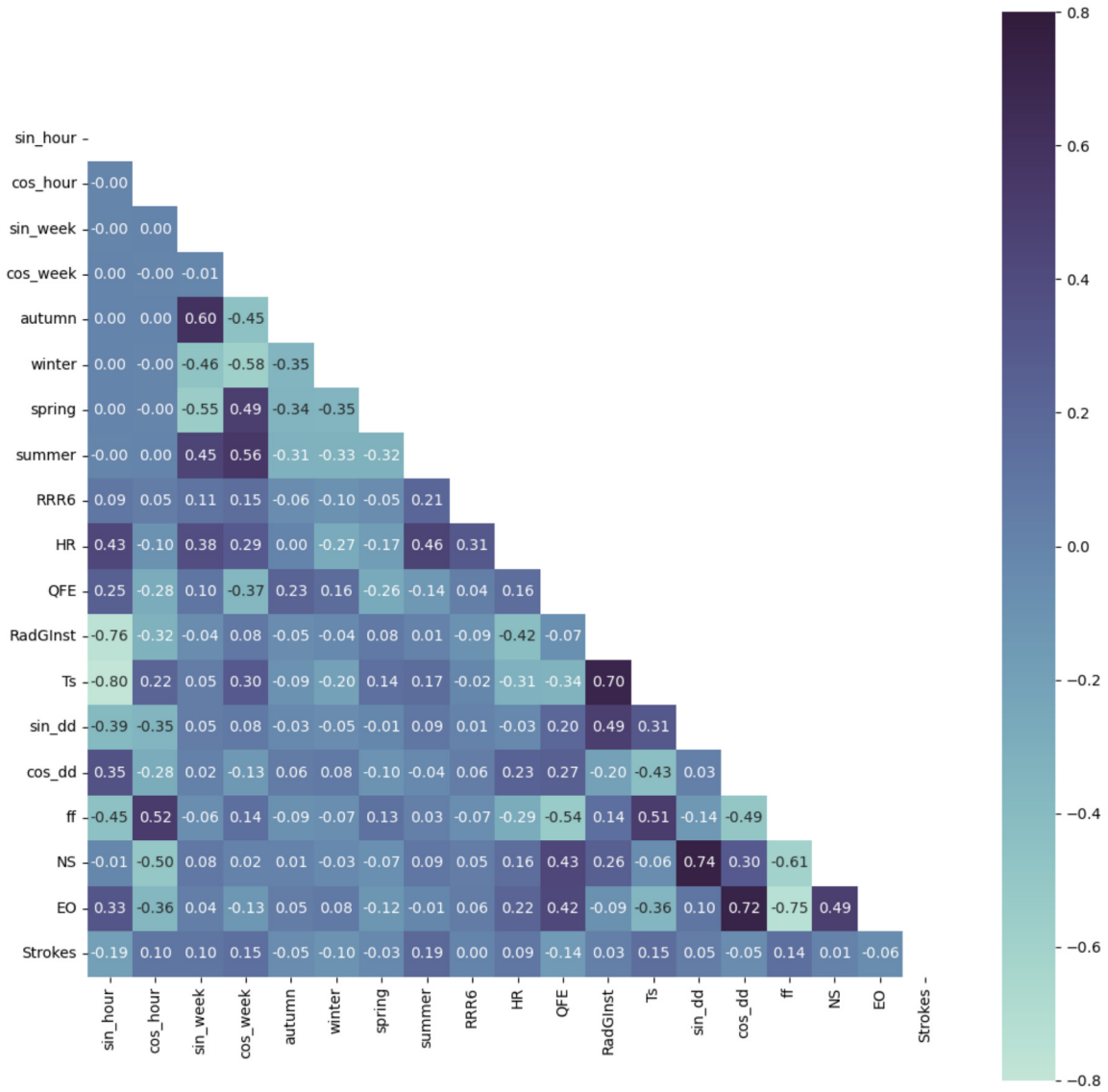


Figura 4.4: Correlaciones para el conjunto de datos

4.3 Tratamiento previo del conjunto de datos

Tras realizar la ingeniería de características, se observó que la mayoría de características presenta porcentajes cercanos al 5% de datos no disponibles, mientras que la presión y temperatura cuentan con un porcentaje cercano al 25%. Se puede encontrar una representación de la falta de información en la figura 3 del Anexo. Esta cantidad de datos no disponibles no representó mayor impacto al comparar las distribuciones de cada una de las características antes y después de aplicar las técnicas de imputación de datos. La falta de información sobre presión y temperatura ocurre principalmente para los primeros ejemplos, por lo que si bien una opción hubiera podido ser descartar todos esos ejemplos, esto significa tener un conjunto menor de entrenamiento lo que conllevaría a crear un modelo con menor historia, aumentando las probabilidades de sobre-ajuste. En las tablas 4.11 y 4.12 puede ver las estadísticas

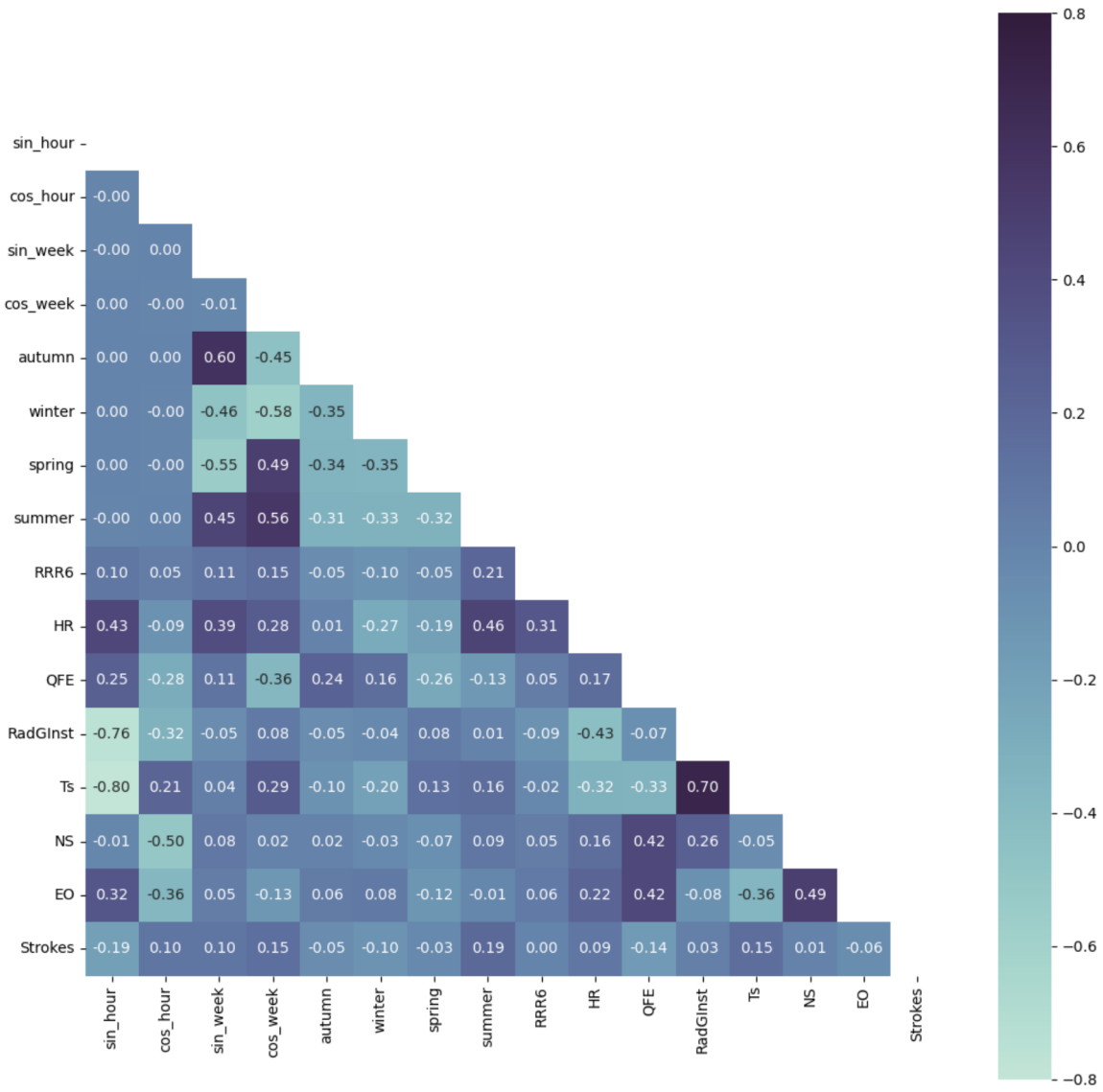


Figura 4.5: Correlaciones finales para el conjunto de datos

descriptivas que resumen la cantidad, la tendencia central y la variabilidad del conjunto de datos creado antes y después de aplicar las técnicas de imputación de datos respectivamente.

La mayor pérdida de datos ocurrió entre las 5 y 6 de la mañana hora GMT, es decir, a las 2 y 3 de la mañana en hora chilena. No están claros los motivos de esta falta de información. En la figura 4 del Anexo se pueden observar gráficos de barras de los datos no disponibles para cada característica en función de las horas.

4.4 Análisis de series temporales

Los resultados realizados mediante la prueba de Dickey-Fuller aumentada entregaron que todas las características del conjunto de datos son al menos un 99 % estacionarias, lo que favorece el uso de las redes que aprenden del pasado.

Respecto de los análisis de autocorrelación para determinar la cantidad y desplazamientos (rezagos) a aplicar, las mejores correlaciones ocurren cada 24 horas[3] y luego cada aproximadamente 8760 horas (1 año). En la figura 5 del Anexo puede ver estos gráficos. De esta forma, cada una de las características es capaz de establecer un múltiplo de desplazamientos de 24 horas antes de que la siguiente mejor correlación sea la que ocurre cerca de las 8760 horas. Si bien usar los desplazamientos que ocurren tras un año permite hacer un entrenamiento

[3] Aunque una presión presentó altas correlaciones cada 12 horas, no son tan grandes como las ocurridas cada 24 horas.

	cantidad	**media**	**desviación**	**mínimo**	**25%**	**50%**	**75%**	**máximo**
sin_hour	75,844	-0.00	0.71	-1.00	-0.71	0.00	0.71	1.00
cos_hour	75,844	0.00	0.71	-1.00	-0.71	0.00	0.71	1.00
sin_week	75,844	-0.03	0.70	-1.00	-0.75	-0.00	0.66	1.00
cos_week	75,844	-0.01	0.71	-1.00	-0.75	-0.00	0.66	1.00
autumn	75,844	0.25	0.43	0	0	0	0	1
winter	75,844	0.27	0.44	0	0	0	1	1
spring	75,844	0.26	0.44	0	0	0	1	1
summer	75,844	0.23	0.42	0	0	0	0	1
RR	74,352	0.03	0.17	0.00	0.00	0.00	0.00	3.68
HR	71,925	38.18	25.49	1.00	17.0	31.0	56.6	99.0
QFE	71,926	627.21	1.53	617.3	626.2	627.3	628.3	632.7
RadGInst	73,417	286.66	391.04	0.0	0.0	5.5	574.0	1,634.0
Ts	71,924	6.18	7.12	-19.1	1.5	6.0	11.9	23.2
NS	71,927	-2.47	4.61	-25.00	-5.60	-1.96	0.00	24.81
EO	71,927	-2.18	5.45	-27.73	-3.53	-0.52	0.78	21.99
Strokes	75,844	-0.93	0.37	-1	-1	-1	-1	1

Tabla 4.11: Estadísticas descriptivas antes de imputar los datos no disponibles

	cantidad	**media**	**desviación**	**mínimo**	**25%**	**50%**	**75%**	**máximo**
sin_hour	75,844	0.00	0.71	-1.00	-0.71	0.00	0.71	1.00
cos_hour	75,844	0.00	0.71	-1.00	-0.71	0.00	0.71	1.00
sin_week	75,844	-0.03	0.70	-1.00	-0.75	0.00	0.66	1.00
cos_week	75,844	-0.01	0.71	-1.00	-0.75	0.00	0.66	1.00
autumn	75,844	0.25	0.43	0	0	0	0	1
winter	75,844	0.27	0.44	0	0	0	1	1
spring	75,844	0.26	0.44	0	0	0	1	1
summer	75,844	0.23	0.42	0	0	0	0	1
RR	75,844	0.03	0.17	0.00	0.00	0.00	0.00	3.68
HR	75,844	37.83	25.23	1.00	17.0	30.8	55.5	99.0
QFE	75,844	627.22	1.52	617.3	626.2	627.3	628.3	632.7
RadGInst	75,844	284.00	390.16	0.0	0.0	3.8	566.9	1,634.0
Ts	75,844	6.01	7.17	-19.1	1.2	5.9	11.8	23.2
NS	75,844	-2.48	4.55	-25.00	-5.48	-1.96	0.00	24.81
EO	75,844	-2.14	5.38	-27.73	-3.44	-0.48	0.78	21.99
Strokes	75,844	-0.93	0.37	-1	-1	-1	-1	1

Tabla 4.12: Estadísticas descriptivas después de imputar los datos no disponibles

considerando la anualidad de los datos, la dimensión del conjunto de datos eleva el tiempo de entrenamiento en cada modelo. Los límites antes de que la siguiente mejor correlación sea la que ocurre aproximadamente al año para cada una de las características fueron los siguientes: RR no más de 4 días, HR no más de 25 días, QFE no más de 60 días, RadGInst podría alcanzar los 365 días, Ts no más de 66 días, NS no más de 60 días y EO no más de 60 días.

4.5 Resultados modelos base

Respecto de las máquinas de aprendizaje automático de Scikit-learn y TensorFlow, en las figuras 4.6 y 4.7 respectivamente, se pueden ver los diagramas de cajas que representan los valores de la puntuación F1 entrenando diferentes clasificadores mediante validación cruzada K-Fold estratificada para diferentes cantidades de rezagos. Cada etiqueta representa el valor acumulado de rezagos hechos, es decir, un "Lag 48"significa que considera las características del conjunto original de datos, más los rezagos de 24 y 48 horas. Solo las características atmosféricas reciben rezagos, esto para estudiar el comportamiento histórico con los días anteriores a la misma hora. Puede ver también que hay máquinas que empeoran y otras que mejoran su desempeño al agregar una mayor cantidad de regazos. Las máquinas que mejoran su desempeño al agregar una mayor cantidad de rezagos no alcanzaron a mostrar valores dado el alcance estudiado. En la tabla 4.13 puede encontrar los promedios de las 5 mejores máquinas cada 24 horas de retraso al considerar tanto las bibliotecas de Scikit-learn como las de TensorFlow.

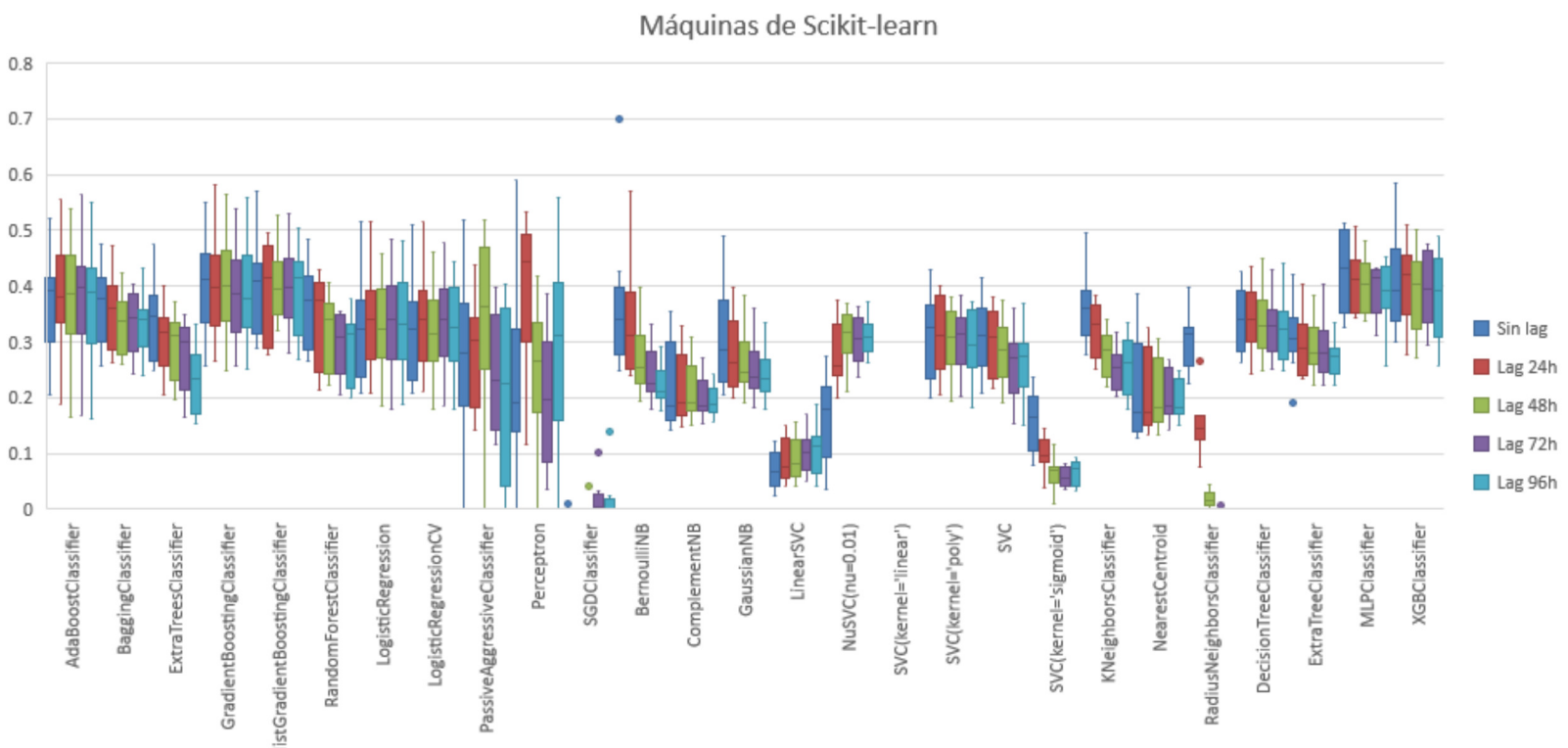


Figura 4.6: Puntuación F1 de Scikit-learn mediante validación cruzada K-Fold estratificada

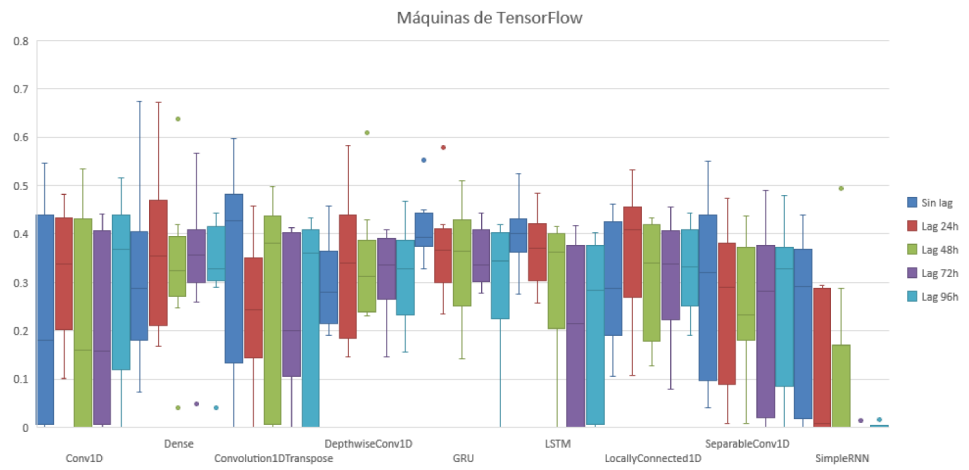


Figura 4.7: Puntuación F1 de TensorFlow mediante validación cruzada K-Fold estratificada

Sin lag	**F1**	**Lag 24h**	**F1**
MLPClassifier	**0.4254**	MLPClassifier	0.4104
GRU	0.4094	XGBClassifier	0.4052
XGBClassifier	0.4092	GradientBoostingClassifier	0.3998
HistGradientBoostingClassifier	0.4012	HistGradientBoostingClassifier	0.3952
GradientBoostingClassifier	0.4005	Perceptron	0.3917

Lag 48h	**F1**	**Lag 72h**	**F1**
MLPClassifier	0.4020	HistGradientBoostingClassifier	0.3980
HistGradientBoostingClassifier	0.4007	XGBClassifier	0.3958
GradientBoostingClassifier	0.3987	MLPClassifier	0.3909
XGBClassifier	0.3938	GradientBoostingClassifier	0.3866
AdaBoostClassifier	0.3743	AdaBoostClassifier	0.3790

Lag 96h	**F1**
HistGradientBoostingClassifier	0.3966
MLPClassifier	0.3884
GradientBoostingClassifier	0.3867
XGBClassifier	0.3771
AdaBoostClassifier	0.3676

Tabla 4.13: Valores medios de las cinco mejores máquinas cada 24 horas de retraso

4.5.1 Selección del modelo y ajuste de hiper-parámetros

Tras revisar los resultados de los modelos base de la tabla 4.13, el modelo que mejor desempeño obtuvo, en promedio, fue una clasificador multicapa sin retraso de 24h en sus características climáticas. Por lo que la siguiente etapa consistirá en realizar un ajuste de hiper-parámetros para este modelo. En la tabla 4.14 puede ver la matriz para la búsqueda de hiper-parámetros para el modelo seleccionado.

neuronas por capa	1, 5, 10, 50, 100, 500
capas ocultas	1, 2, 3, 4, 5
función de activación	“identity”, “logistic”, “tanh”, “relu”

Tabla 4.14: Malla de búsqueda para el ajuste de hiper-parámetros

Tras realizar el ajuste de hiper-parámetros, los resultados indicaron que el modelo con el menor error de entrenamiento tiene 2 capas ocultas, activación tangente hiperbólica y 500 neuronas.

Si bien el modelo no ofrece un mayor ajuste como la adición de dropout, en la tabla 4.15 puede ver los resultados del análisis de sensibilidad realizado para las capas y neuronas.

Para mejorar el rendimiento de este modelo, se le añadió una capa de dropout y se volvió a realizar este mismo ajuste. Las capas de dropout eliminan aleatoriamente las neuronas de las capas, esto ayuda a evitar el sobreajuste (Liang y cols., 2021). Si bien esto significó dejar de usar la librería de Scikit-learn para usar las capas densas de TensorFlow, a final de cuentas los resultados resultaron mejores en el análisis de sensibilidad. Los resultados entregaron que el modelo debiera consistir en la combinación de 5 capas densas con dropout al 10 %. Cada capa debe tener 500 neuronas y una función de activación ReLu. Los resultados de la matriz de confusión, entregó una puntuación F1 de 0.6218.

Finalmente, los resultados tras realizar un nuevo análisis de sensibilidad fueron los que

Combinaciones de neuronas en capas ocultas	F1
200 y 100	0.3803
200 y 200	0.3648
200 y 300	0.3903
500 y 300	0.4022
500 y 500	0.4596
500 y 700	0.4090
700 y 500	0.3872
700 y 700	0.3713
700 y 1000	0.4396

Optimizador	F1
adam	0.4596
sgd	0.4354
lbfgs	0.2927

Tabla 4.15: Ajuste de sensibilidad

se muestran en la tabla 4.16 para los datos post-procesados[4].

Lo que deja como resultado un modelo compuesto por 30 combinaciones de capa densa y dropout al 20 %, cada capa con 100 neuronas y función de activación ReLu para puntuar un F1 de 0.8046.

4.5.2 Prueba del modelo

Este modelo creado necesita realizar un post-procesamiento de los resultados una vez esté entrenado, pues como el objetivo es encontrar días de tormenta, basta con que durante 1 hora se haya predicho una tormenta para que se considere todo el día como tormenta, esto significa volver a muestrear los valores predichos cada 24 horas y considerar el valor máximo. Tras realizar esto, el resultado de la matriz de confusión respecto del conjunto de prueba lo puede ver en la figura 4.8.

		Predicted values		
		Negative	**Positive**	**Total**
Actual values	**Negative**	TN 172	FP 62	N′
	Positive	FN 16	TP 115	P′
	Total	N	P	

Figura 4.8: Matriz de confusión modelo post-procesado

[4] El post-procesamiento se explica en la sección 4.5.2.

Función de activación	**F1**
Elu	0.6471
Exponencial	0.6131
GELU	0.6537
Hard sigmoid	0.6537
Lineal	0.6363
ReLu	0.7009
SELU	0.6404
Sigmoidea	0.6051
Softmax	0.6091
Softplus	0.6794
Softsing	0.5699
Swish	0.6131
Tanh	0.6470

Número de neuronas	**F1**
5	0.5561
10	0.6635
20	0.6667
30	0.6667
40	0.6827
50	0.6794
100	0.7042
200	0.6468
300	0.6857
400	0.6952
500	0.6794

Número de capas ocultas	**F1**
1	0.6948
2	0.6667
3	0.6634
4	0.7163
5	0.7097
10	0.6977
20	0.7619
30	0.7801
40	0.7521
50	0.7692

Dropout	**F1**
0.01	0.7103
0.02	0.7103
0.05	0.6977
0.1	0.7215
0.2	0.8046
0.3	0.7417
0.4	0
0.5	0

Tabla 4.16: Ajuste de sensibilidad final

En resumen, esta red recibió variables de entrada temporales y meteorológicas, 30 combinaciones de capa densa con 100 neuronas y función de activación ReLu más una capa dropout al 20 % y una salida con función sigmoidea que indica la presencia de rayos. El mejor modelo completamente entrenado obtuvo un desempeño del 41.99 % en la puntuación F1 con los datos sin procesar y un 80.46 % con los datos post-procesados sobre el conjunto de validación. Para el conjunto de prueba los resultados fueron del 33.47 % en la puntuación F1 con los datos sin procesar y un 74.68 % con los datos post-procesados.

5 Conclusiones y recomendaciones

Índice general

5.1 Principales conclusiones

En este trabajo de memoria se presentó un modelo para la predicción de días de tormenta dentro del territorio chileno mediante el uso de técnicas de IA para lugar preestablecido dentro del territorio chileno. Este lugar fue Visviri, dado sus particulares condiciones de tormenta eléctrica y de la posibilidad de obtener la suficiente cantidad de ejemplos y características para entrenar un modelo con la suficiente cantidad de datos. Sin embargo, pese a estos esfuerzos, la cantidad de descargas registradas solo correspondió al 3 %, lo que significó abordar un problema de clases desequilibradas.

Respecto del análisis exploratorio de datos, la mayor cantidad de descargas ocurre durante el último cuarto del día del primer cuarto del año, lo que coincide con las tardes de verano. Pero existen comportamientos diferentes para cada estación del año, obteniendo diferentes conclusiones según la estación que se observe.

La precipitación, en promedio, para las estaciones de otoño, invierno y primavera es mayor cuando hay tormenta respecto de los cielos despejados, mientras que la precipitación es menor en verano cuando hay tormenta. Para las estaciones de otoño e invierno, la humedad relativa es mayor cuando hay tormenta, y es menor en primavera y verano. Para todas las estaciones del año, la presión a nivel de estación es menor cuando hay tormenta. Para todas las estaciones del año, la radiación solar instantánea es menor cuando hay tormenta. Para todas las estaciones del año, la temperatura del aire seco es mayor cuando hay tormenta. Para todas las estaciones del año, la magnitud del viento es mayor cuando hay tormenta. Y para las estaciones de otoño, primavera y verano, la dirección del viento se dirige más al noreste cuando hay tormenta, mientras que se dirige más al suroeste en invierno. Todas las características se encuentran concentradas hacia la izquierda a excepción de la temperatura, mientras que todas las características son mesocúrticas, a exepción de la precipitación que es leptocúrtica.

Las mejores correlaciones ocurrieron entre la triada sin_hour[1], radiación y temperatura, pues la formación de nubes de tormenta ocurre más frecuentemente durante las tardes de poca radiación solar instantánea y alta temperatura del aire seco.

[1] Componente vertical de la descomposición cartesiana que representa a un reloj de 24 horas.

Respecto de las máquinas de aprendizaje, los mejores resultados recayeron en los modelos basados en las redes neuronales y los árboles de clasificación. Pero para escoger el modelo a optimizar y acorde a la metodología, los resultados de la validación cruzada k-fold estratificada respecto de la puntuación F1 mostraron que el modelo que mejor se acopló al conjunto de datos es una red multicapa en Scikit-learn.

Este modelo fue optimizado y una vez completamente entrenado, obtuvo un desempeño del 74.68 % en la puntuación F1. Un adecuado modelo de predicción de tormentas eléctricas vendría a ser una herramienta extremadamente útil que permitiría activar alertas preventivas, evitando la pérdida de vidas humanas y equipos eléctricos que reducen el normal funcionamiento de cualquier actividad o instalación. También da pie a reportar avances en distintos ámbitos sociales (como por ejemplo una integración al SAE), económicos (actividades ligadas a faenas mineras, actividades deportivas, industria aeronáutica, etc.), ampliar la frontera del saber (permitiendo su integración a pronósticos atmosféricos bajo simulaciones del modelo Weather Research and Forecasting (WRF)) y por último, dar pie a futuros desarrollos o investigaciones.

5.2 Soluciones frente a las problemáticas ocurridas

Una de las principales problemáticas afrontadas en este trabajo fue abordar la pérdida de los datos en cada una de las características. Esta problemática si bien fue solucionada usando la imputación con los k-vecinos más cercanos, la elección de el mejor número de vecinos cercanos fue seleccionado heurísticamente. Esto trajo algunas consideraciones para las primeras posiciones de datos, pues no todas las características iniciaron su registro en la misma fecha. Esta consideración consistió en que se evidenció la misma secuencia de valores cada 24 horas para los primeros días en los que no hubo datos.

Otra de las problemáticas tuvo estrecha relación con la anterior, pues en 2015 y durante 5 semanas, la radiación no fue registrada, para lo cual la estrategia consistió en usar la media entre los valores del año anterior y posterior. Esto logró imputar considerablemente los datos no disponibles para esas semanas.

Finalmente, el problema de las clases desequilibradas fue abordado usando una validación estratificada en la que la clase mayoritaria se fracciona y la clase minoritaria (la clase asociada a la tormeta) se usa en cada uno de estos nuevos conjuntos. Esto se tradujo en un aumento en el tiempo asociado a la búsqueda de los mejores hiper-parámetros.

5.3 Limitaciones de la investigación y trabajo futuro

Este estudio tiene ciertas limitaciones y trabajos bajo supuestos que se pasará a mencionar a continuación.

La primer limitación existente es que el valor de la función objetivo de la clase minoritaria, influye fuertemente en los valores de sus respectivas características (valores de temperatura, presión, radiación, etc. para las cuales ocurre tormenta), pues cada vez que se realice la validación cruzada, estos valores meteorológicos estarán sobre-representados.

El pronóstico del valor de la hora siguiente no fue abordado, pero la forma de atacar estos problemas puede ser usando alternativas como Online Learning, Transfer Learning o modelos matemáticos (model-based) como SARIMA u otro.

Si bien la red que registra las tormentas también puede entregar un valor de energía, esto no permite determinar el tipo de rayo ocurrido. Este valor de energía define la intensidad de la señal que se registró en las antenas. Y otra característica que tampoco fue abordada fue la multiplicidad de veces que un rayo impactó en un lugar determinado durante una tormenta.

El cuadrado de lado 30 km fue una aproximación en la que un observador ubicado a 15 km puede ver un rayo. Una mejor aproximación del conjunto de datos es programar un círculo

de radio 15 km a la redonda desde el punto medio entre la o las EMA que cuenten con la mayor información meteorológica.

La falta de poder registrar otras características relevantes, como el valor del campo eléctrico. Registrar estas características sería un gran aporte en la calidad del conjunto de datos. Tener la posibilidad de medir e incluir esta característica puede aumentar el rendimiento del modelo de predicción.

Sería bastante conveniente poder establecer un criterio más particular para los valores atípicos, pues en este caso, el último valor aislado correspondió a una presión decenas de desviaciones estándar desplazadas del valor medio, lo que no dice si es que puede existir una mejor forma de abordar este problema.

La forma de imputar los valores no disponibles fue mediante técnicas tradicionales, también se puede abordar esto desde la mirada del aprendizaje automático y que la salida del modelo puedan ser los valores no disponibles. Luego con el conjunto de datos completo, crear el modelo de predicción de tormentas.

Existen limitaciones en la cantidad de rezagos aplicados, pues hubo modelos que tras los rezagos de 96 horas, mejoraron el valor de su puntuación F1. Sería interesante estudiar rezagos más profundo en estos modelos, al menos considerando la autocorrelación de las características como límite. Así como también evaluar otra cantidad de rezagos para todos los modelos, diferentes a los múltiplos de 24 horas utilizado.

Resolver la limitaciones mencionadas encaminan el trabajo futuro, pues el potencial que tienen estas tecnologías de aprendizaje automático viene a ser una gran herramienta que apoye la toma de decisiones.

Referencias

Abdullah, N. H., Adnan, R., Samad, A. M., y Ahmat Ruslan, F. (2018). Lightning forecasting modelling using artificial neural network (ann): Case study sultan abdul aziz shah airport or skypark subang. En *2018 ieee conference on systems, process and control (icspc)* (p. 1-4). doi: 10.1109/SPC.2018.8704147

Aggarwal, C. C. (2014). Data classification: Algorithms and applications..

Anaconda. (s.f.). https://www.anaconda.com//. (Accessed: 2023-05-25)

Arnold, L. (1974). *Stochastic differential equations: Theory and applications.* Wiley Interscience. doi: 10.1142/6453

Athanasopoulos, R. J., George; Hyndman. (2018). *Forecasting: Principles and practice* (2nd ed.). Descargado de http://gen.lib.rus.ec/book/index.php?md5=1a297f9de179e8b0d3884224401e9372

Bala, K., Choubey, D. K., y Paul, S. (2017). Soft computing and data mining techniques for thunderstorms and lightning prediction: A survey. En *2017 international conference of electronics, communication and aerospace technology (iceca)* (Vol. 1, p. 42-46). doi: 10.1109/ICECA.2017.8203729

Carey, L. D., Rutledge, S. A., y Petersen, W. A. (2003). The relationship between severe storm reports and cloud-to-ground lightning polarity in the contiguous united states from 1989 to 1998. *Monthly Weather Review, 131*, 1211-1228.

Chaudhuri, S. (2011, 01). A probe for consistency in cape and cine during the prevalence of severe thunderstorms:statistical – fuzzy coupled approach. *Atmospheric and Climate Sciences, 4*, 197 - 205. doi: 10.4236/acs.2011.14022

Cooper, M. A., y Holle, R. (2019, 01). Lightning detection. En (p. 139-149). doi: 10.1007/978-3-319-77563-0_14

Dirección meteorológica de chile. (s.f.). http://www.meteochile.gob.cl/. (Accessed: 2023-04-29)

Girosi, F., Jones, M., y Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation, 7*(2), 219-269. doi: 10.1162/neco.1995.7.2.219

Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep learning.* The MIT Press.

Jayendra, G., Lucas, R., Kumarawadu, S., Neelawala, L., Jeevantha, C., y Dharmapriya, P. (2007). Intelligent lightning warning system. En *2007 third international conference on information and automation for sustainability* (p. 19-24). doi: 10.1109/ICIAFS.2007.4544774

Johari, D., Rahman, T. K. A., y Musirin, I. (2007). Artificial neural network based technique for lightning prediction. En *2007 5th student conference on research and development* (p. 1-5). doi: 10.1109/SCORED.2007.4451448

Jupyterlab documentation. (s.f.). https://jupyterlab.readthedocs.io/en/stable/. (Accessed: 2023-05-25)

Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., y Steinbrecher, M. (2016). Introduction to neural networks..

Leo, B. (1996). Bagging predictors. En *Machine learning* (Vol. 24, p. pages123–140). doi: 10.1023/A:1018054314350

Liang, X., Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., ... Liu, T.-Y. (2021). *R-drop: Regularized dropout for neural networks*.

McCulloch, W. S., y Pitts, W. H. (2021). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*, 99-115.

Mitchell, T. M. (1997). *Machine learning* (1.ª ed.). McGraw-Hill. Descargado de http://gen.lib.rus.ec/book/index.php?md5=f3aa83fb7adab9c8675871a717db6231

Mostajabi, A., Finney, D. L., Rubinstein, M., y Rachidi, F. (2019). Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *npj Clim Atmos Sci*, *2*, 41. Descargado de https://www.nature.com/articles/s41612-019-0098-0/ doi: https://doi.org/10.1038/s41612-019-0098-0

Orlanski, I. (1975). A rational subdivision of scales for atmospheric processes. En *Bulletin of the american meteorological society* (Vol. 56, p. 527–530).

Pakdaman, M., Naghab, S. S., Khazanedari, L., Malbousi, S., y Falamarzi, Y. (2020). Lightning prediction using an ensemble learning approach for northeast of iran. *Journal of Atmospheric and Solar-Terrestrial Physics*, *209*, 105417. Descargado de https://www.sciencedirect.com/science/article/pii/S1364682620302236 doi: https://doi.org/10.1016/j.jastp.2020.105417

Price, C. G., y Rind, D. (1992). A simple lightning parameterization for calculating global lightning distributions. *Journal of Geophysical Research*, *97*, 9919-9933.

Python. (s.f.). https://docs.python.org/3/license.html/. (Accessed: 2023-05-01)

Ramzi, M. M., Adnan, R., Samad, A. M., y Ruslan, F. A. (2018). Lightning prediction modelling using mlpnn structure. case study: Kuala lumpur international airport (klia). En *2018 ieee international conference on automatic control and intelligent systems (i2cacis)* (p. 63-66). doi: 10.1109/I2CACIS.2018.8603704

Robert H. Shumway, D. S. S. (2017). *Time series analysis and its applications with r examples* (4.ª ed.). Springer. Descargado de http://gen.lib.rus.ec/book/index.php?md5=ccd300e85d7455ef25b4c6fcea4e52e7

Romps, D. M., Seeley, J. T., Vollaro, D., y Molinari, J. E. (2014). Projected increase in lightning strikes in the united states due to global warming. *Science*, *346*, 851 - 854.

Scikit-learn. (s.f.). https://pypi.org/project/scikit-learn/. (Accessed: 2023-07-03)

Singh, D., Singh, R., Singh, A., Kulkarni, M., Gautam, A., y Singh, A. (2011). Solar activity, lightning and climate. *Surveys in Geophysics*, *32*, 659-703.

Tensorflow. (s.f.). https://www.tensorflow.org/about?hl=es-419/. (Accessed: 2023-07-03)

Tomas, G. R. (2004). Actualización del mapa isoceráunico de guatemala y su influencia en el diseño de líneas de transmisión..

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... Altman, R. B. (2001, 06). Missing value estimation methods for DNA microarrays . *Bioinformatics*, *17*(6), 520-525. Descargado de https://doi.org/10.1093/bioinformatics/17.6.520 doi: 10.1093/bioinformatics/17.6.520

Viale, M., y Garreaud, R. (2013). Summer precipitation events over the western slope of the subtropical andes. En *Monthly weather review* (Vol. 142, p. 1704–1092).

Wayback. (s.f.). https://web.archive.org/web/20200224120525/https://luca-d3.com/es/data-speaks/diccionario-tecnologico/python-lenguaje/. (Accessed: 2023-05-01)

Widodo, S., Brawijaya, H., y Samudi, S. (2022, Oct.). Stratified k-fold cross validation optimization on machine learning for prediction. *Sinkron : jurnal dan penelitian teknik infor-*

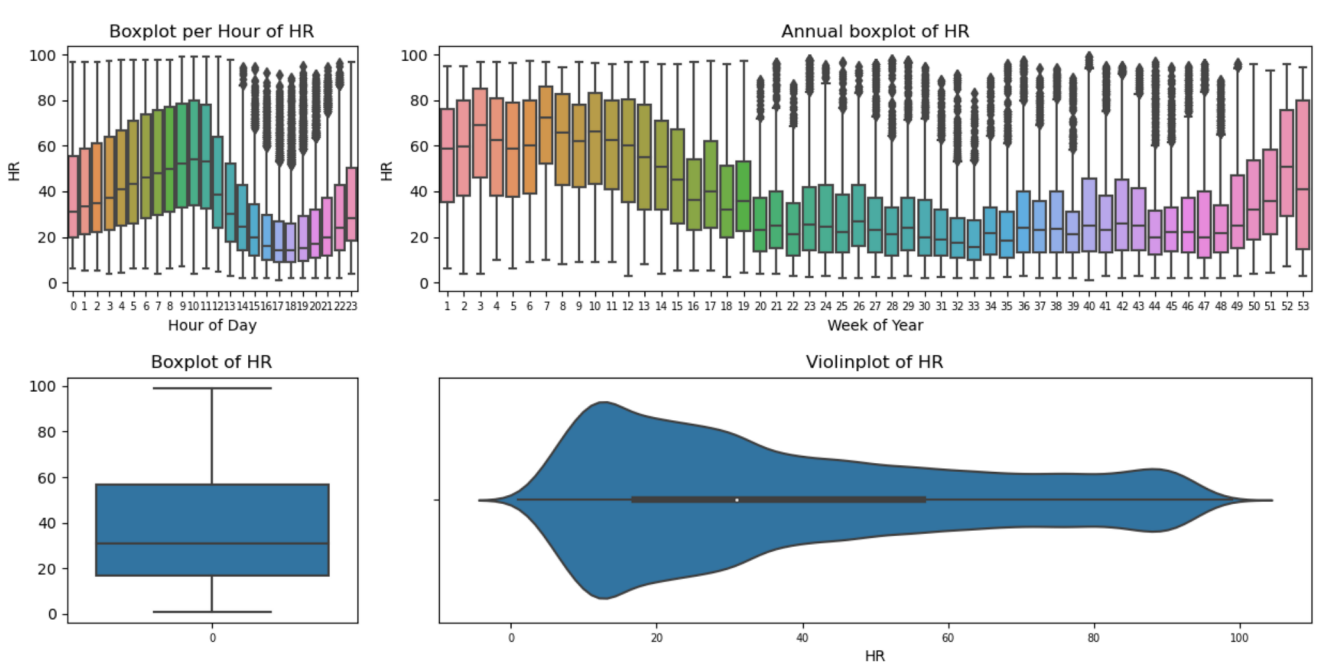
matika, *7*(4), 2407-2414. Descargado de https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11792 doi: 10.33395/sinkron.v7i4.11792

Wilks, D. S. (1995). Statistical methods in the atmospheric sciences: An introduction..

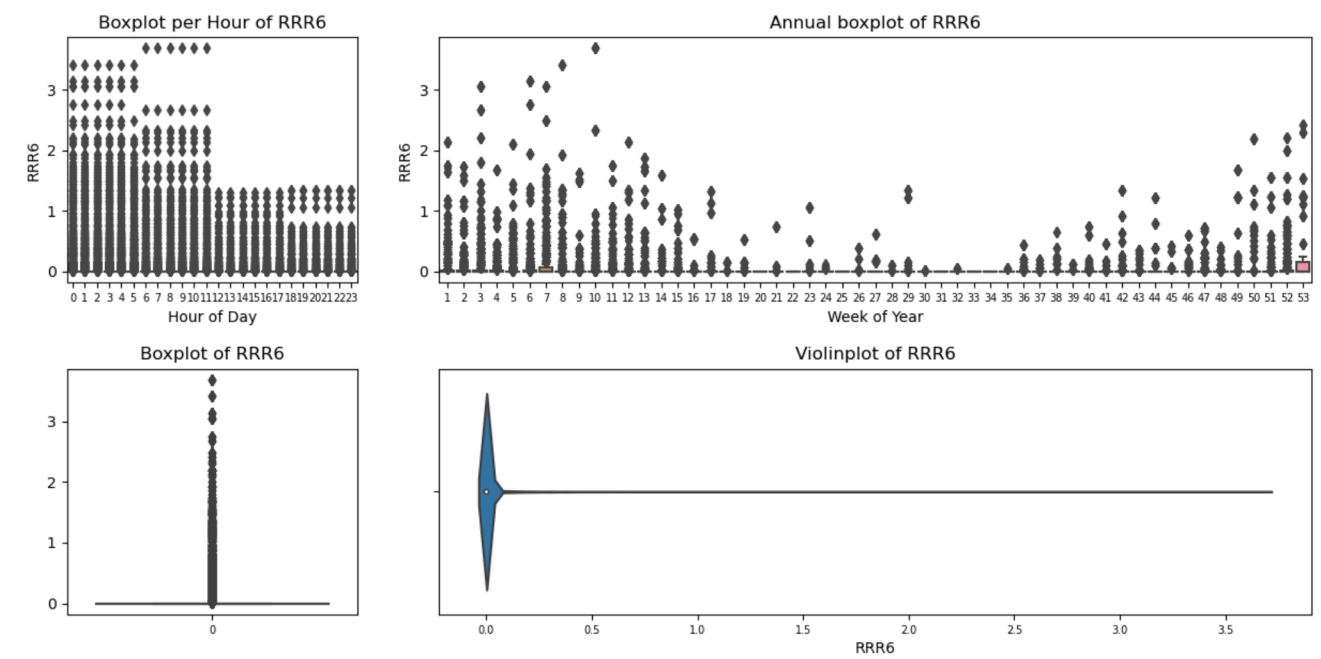
Willoughby, R. A. (1979). Solutions of ill-posed problems (a. n. tikhonov and v. y. arsenin). *SIAM Review*, *21*(2), 266-267. Descargado de https://doi.org/10.1137/1021044 doi: 10.1137/1021044

Wwlln. (s.f.). https://wwlln.net/. (Accessed: 2023-04-29)

Anexo: Figuras

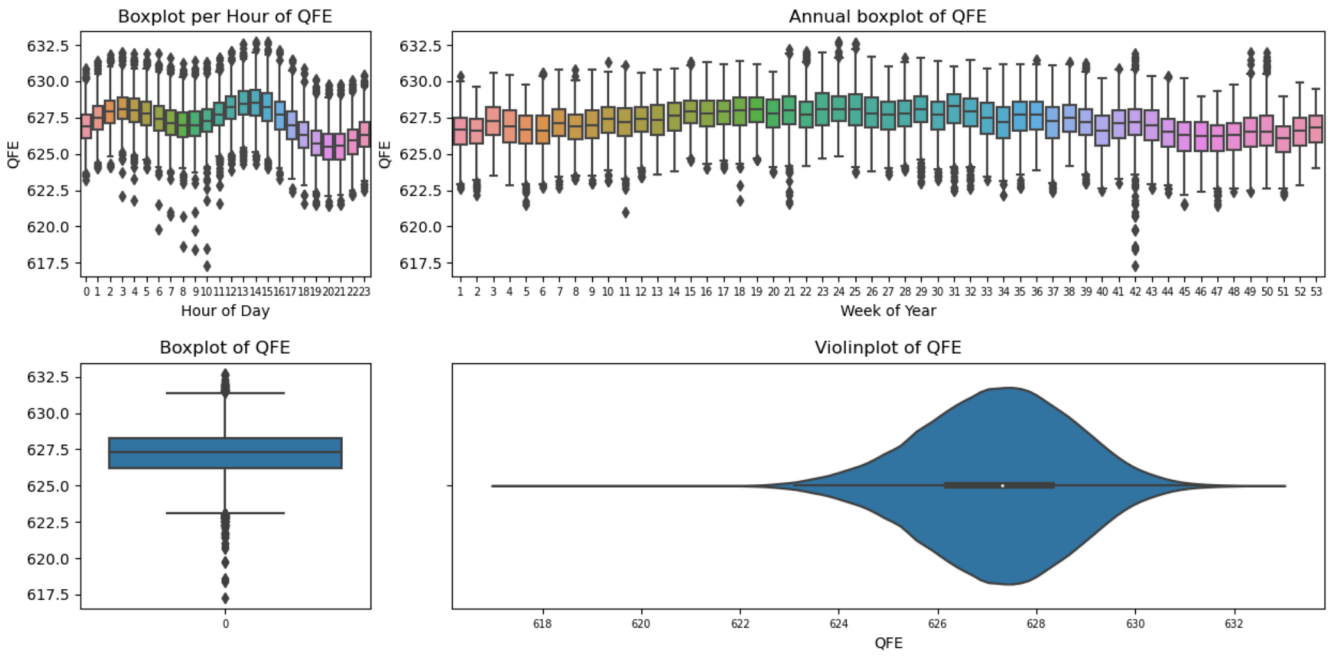


(a) Diagramas de caja y de violín para la Humedad relativa

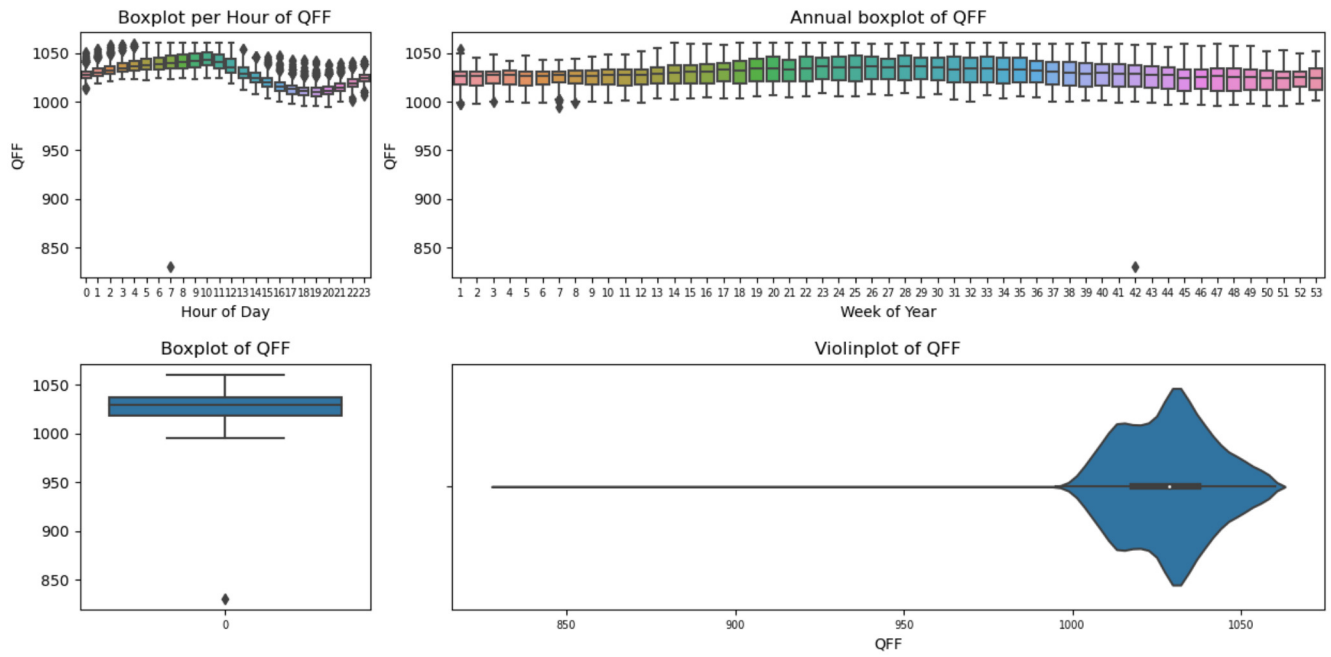


(b) Diagramas de caja y de violín para la Precipitación

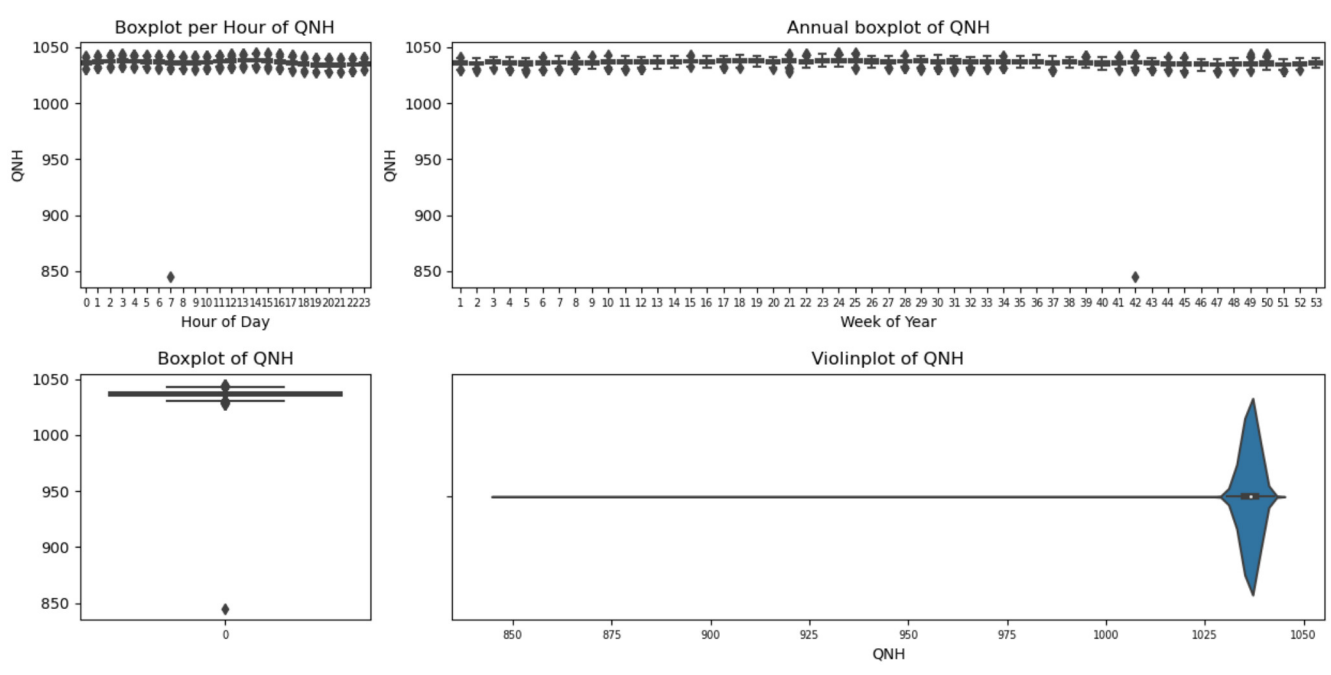
Figura 1: Diagramas de caja y de violín para las características registradas en Visviri, Chile



(c) Diagramas de caja y de violín para la Presión a nivel de estación

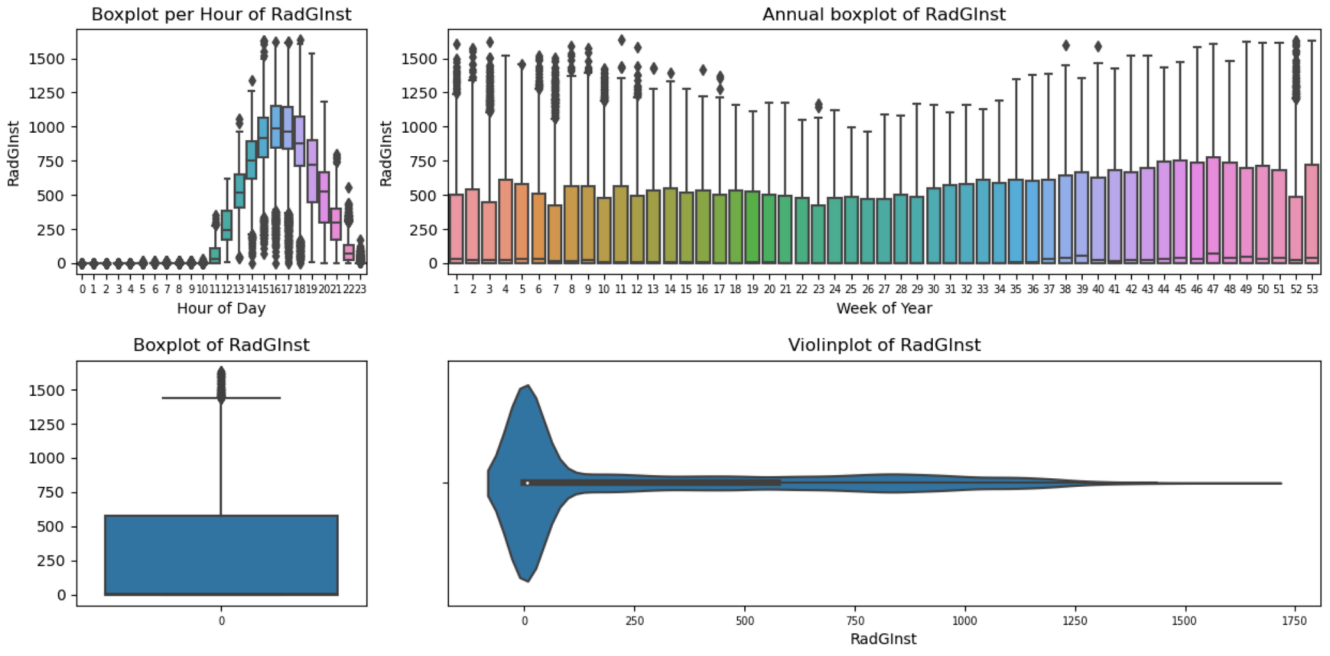


(d) Diagramas de caja y de violín para la Presión a nivel del mar

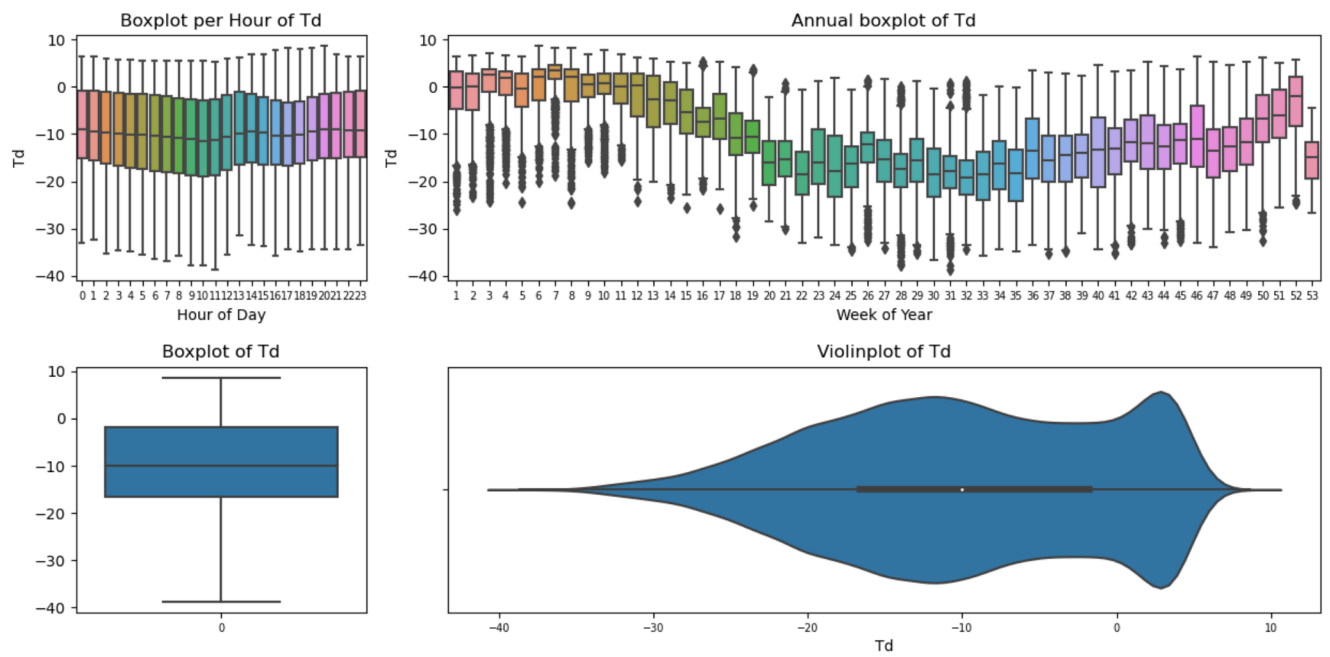


(e) Diagramas de caja y de violín para la Presión a nivel del mar mediante Atmósfera Estándar de la OACI

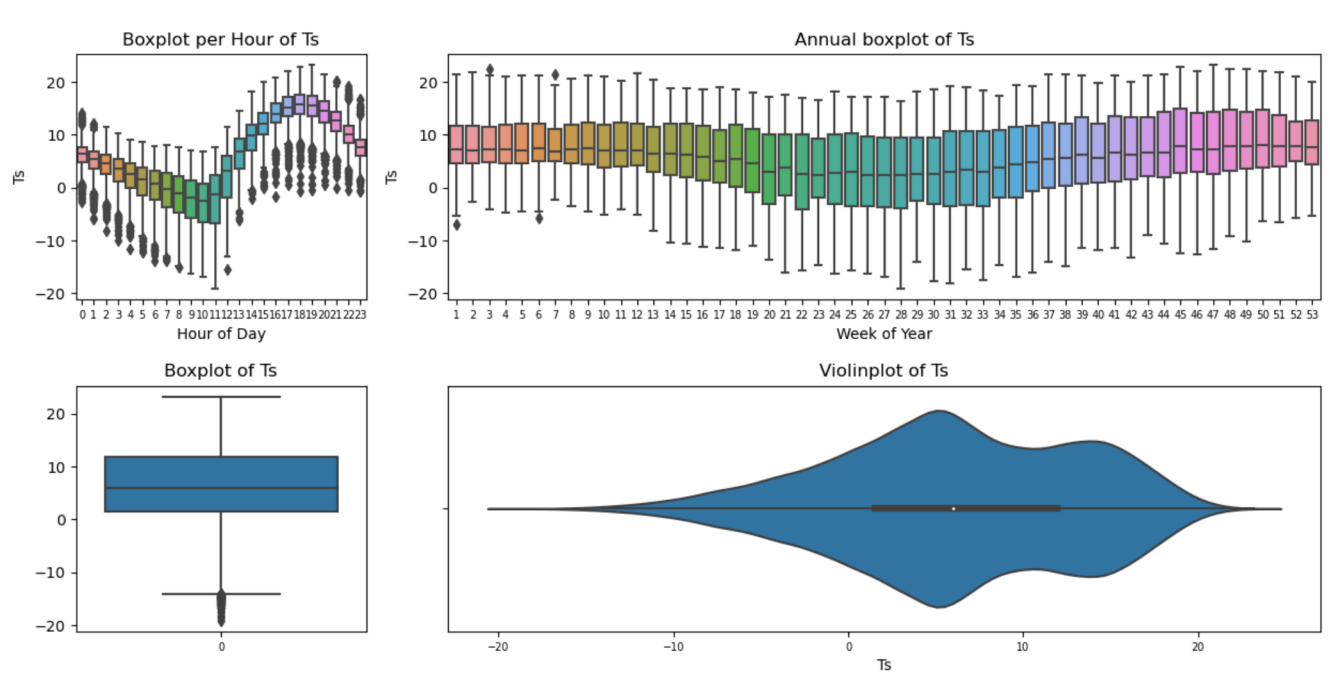
Figura 1: Diagramas de caja y de violín para las características registradas en Visviri, Chile



(f) Diagramas de caja y de violín para la Radiación Solar Instantánea



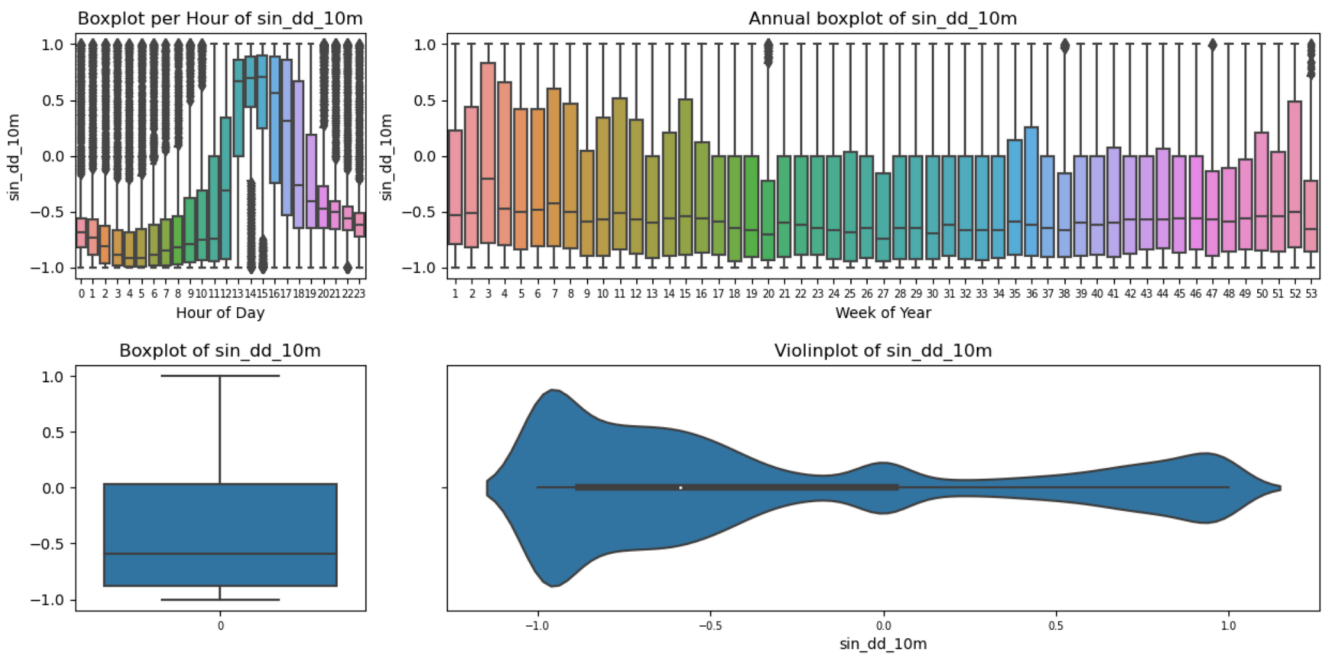
(g) Diagramas de caja y de violín para la Temperatura del punto de rocío



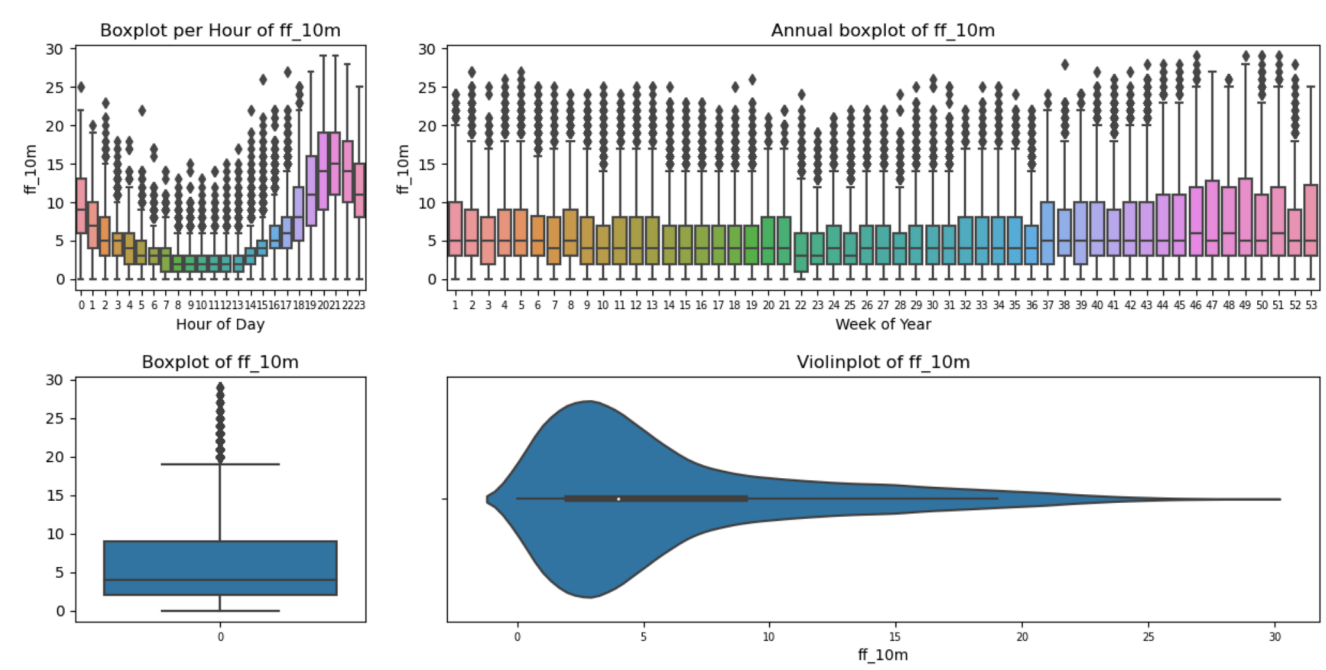
(h) Diagramas de caja y de violín para la Temperatura del aire seco

Figura 1: Diagramas de caja y de violín para las características registradas en Visviri, Chile

(i) Diagramas de caja y de violín para la Componente horizontal de la dirección del viento a 10 m de altura

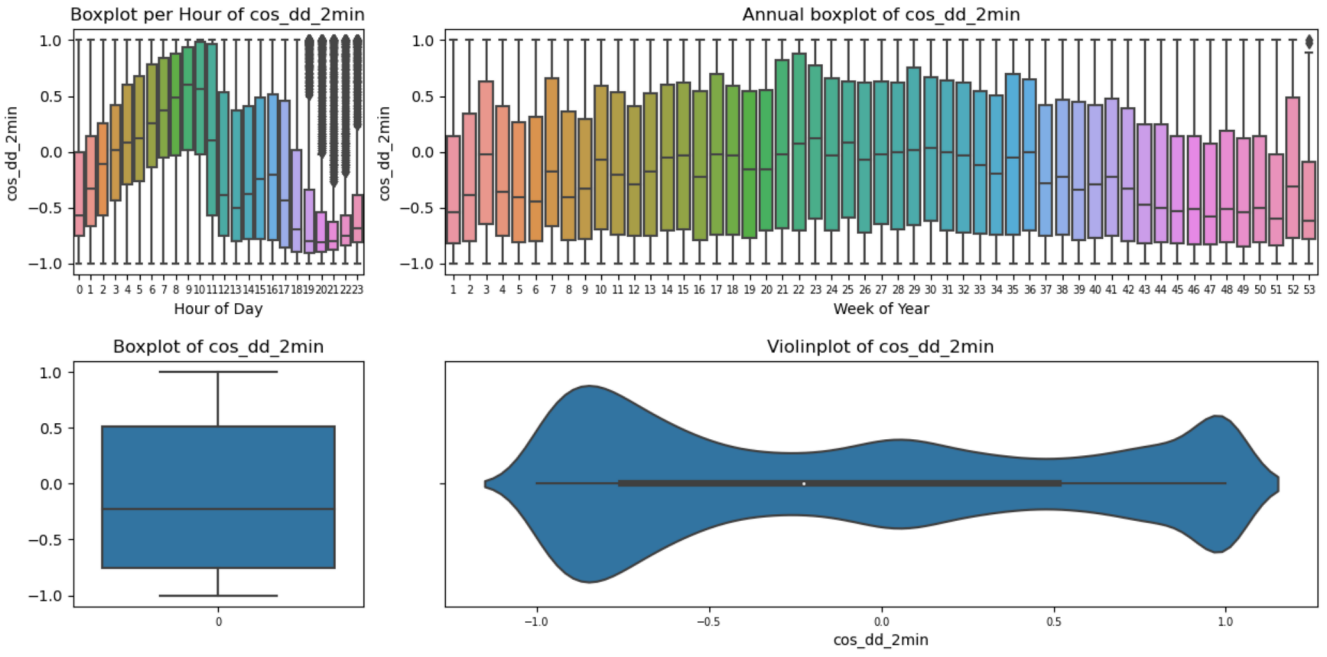


(j) Diagramas de caja y de violín para la Componente vertical de la dirección del viento a 10 m de altura

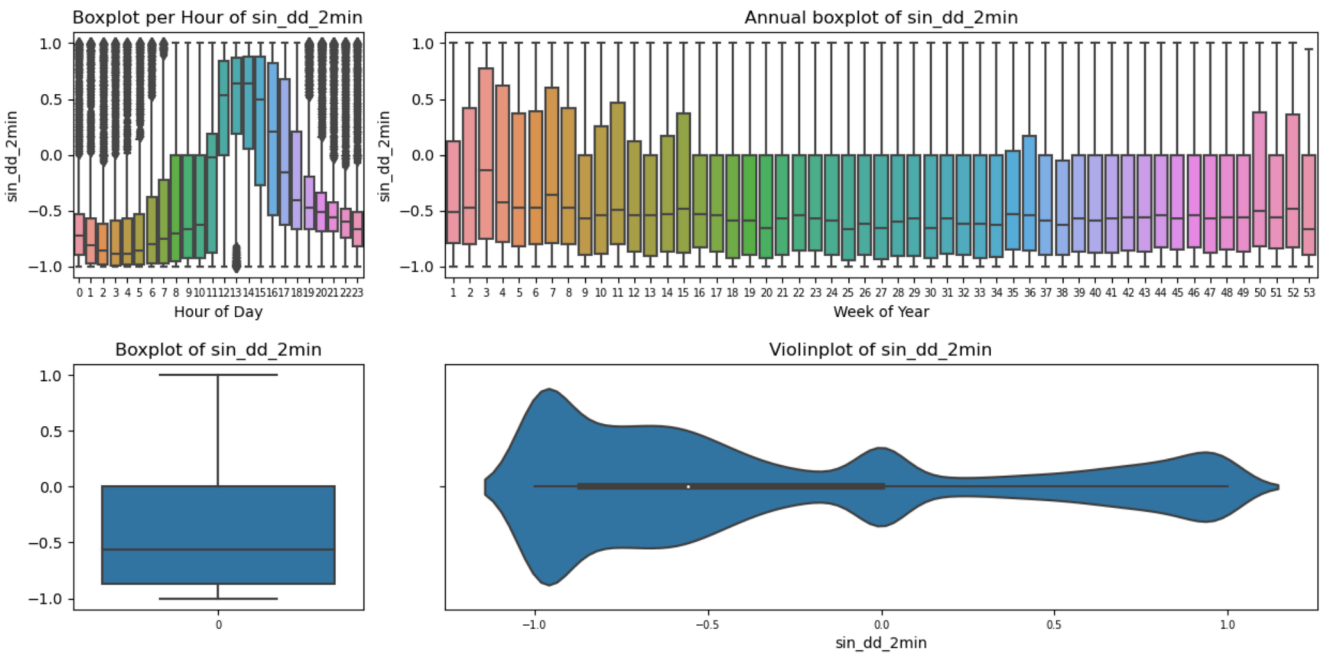


(k) Diagramas de caja y de violín para la Intensidad del viento a 10 m de altura

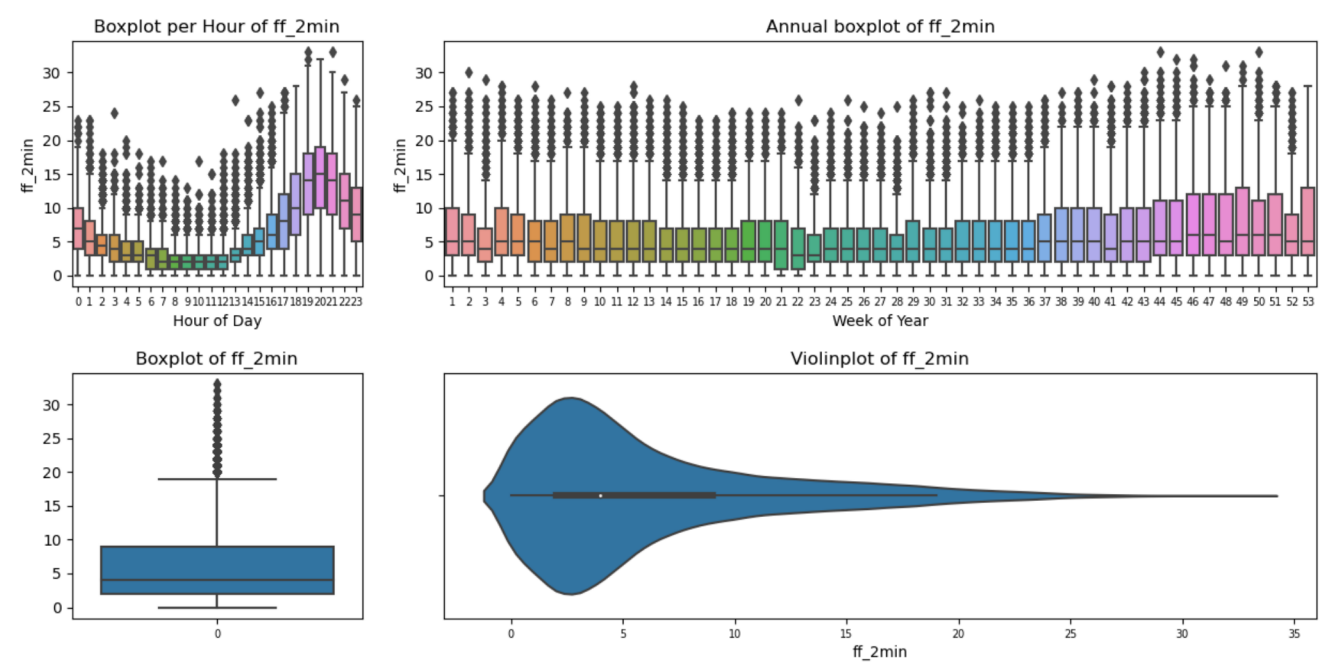
Figura 1: Diagramas de caja y de violín para las características registradas en Visviri, Chile



(l) Diagramas de caja y de violín para la Componente horizontal de la dirección del viento promedio cada 2 minutos



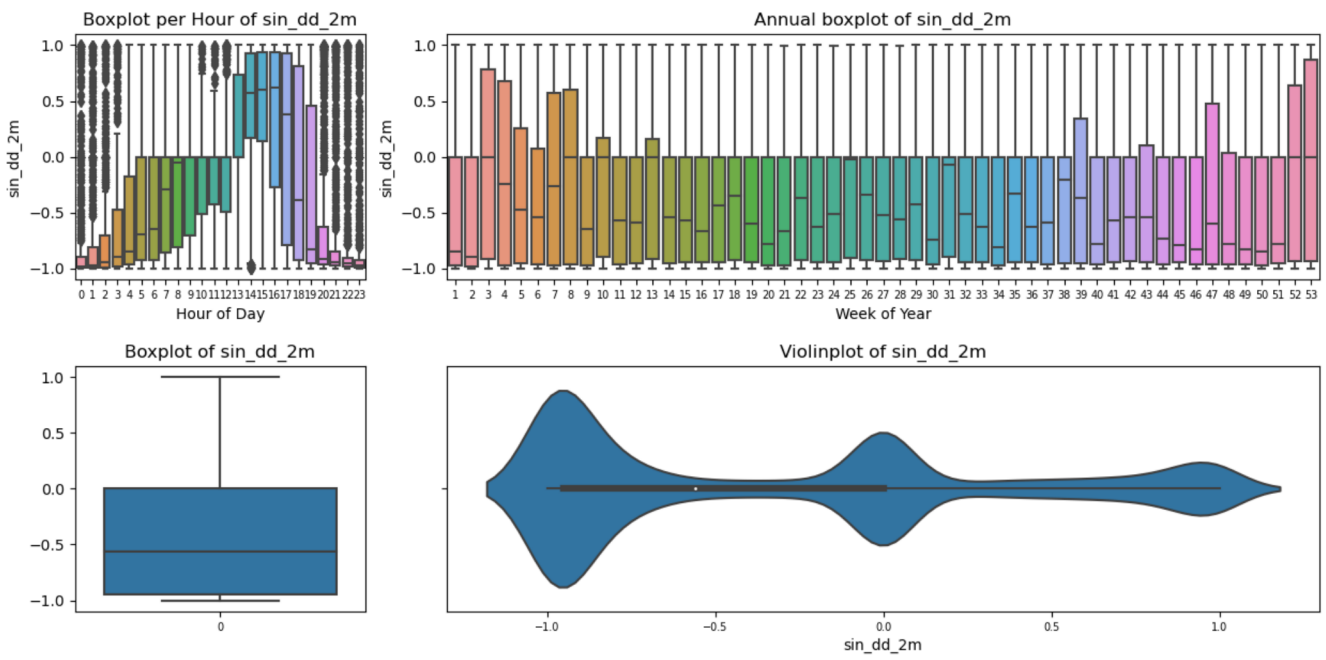
(m) Diagramas de caja y de violín para la Componente vertical de la dirección del viento promedio cada 2 minutos



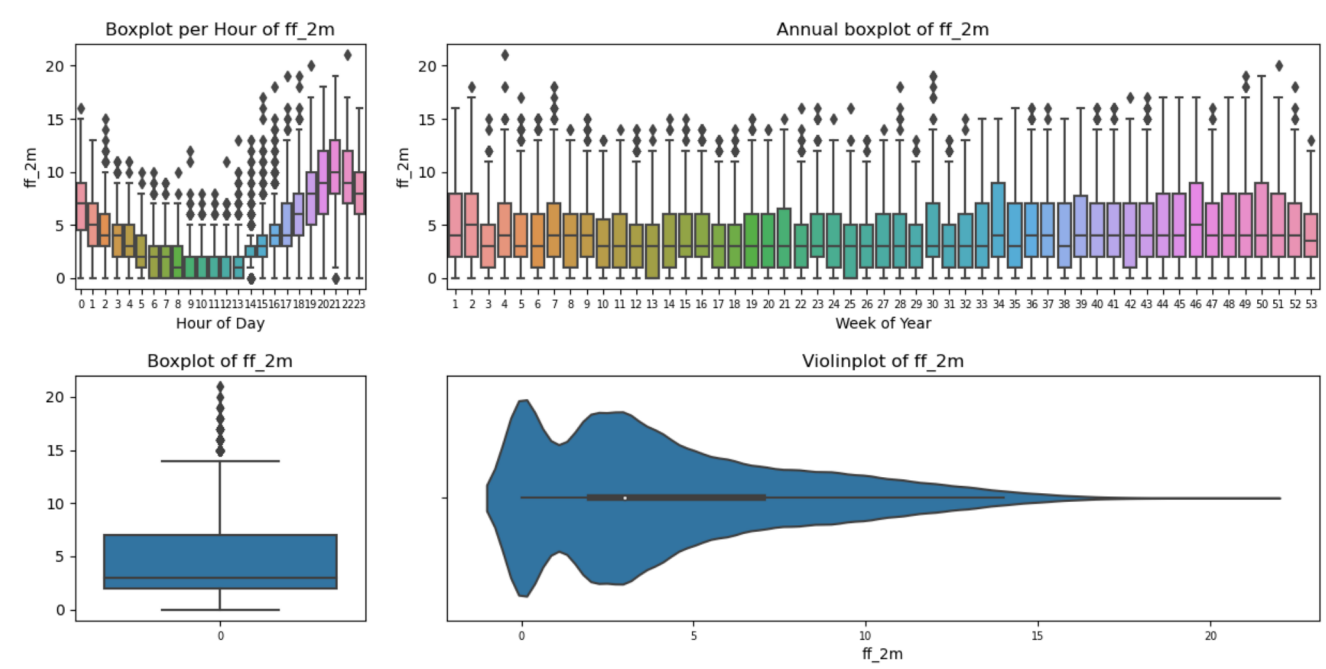
(n) Diagramas de caja y de violín para la Intensidad del viento promedio cada 2 minutos

Figura 1: Diagramas de caja y de violín para las características registradas en Visviri, Chile

(ñ) Diagramas de caja y de violín para la Componente horizontal de la dirección del viento a 2 m de altura

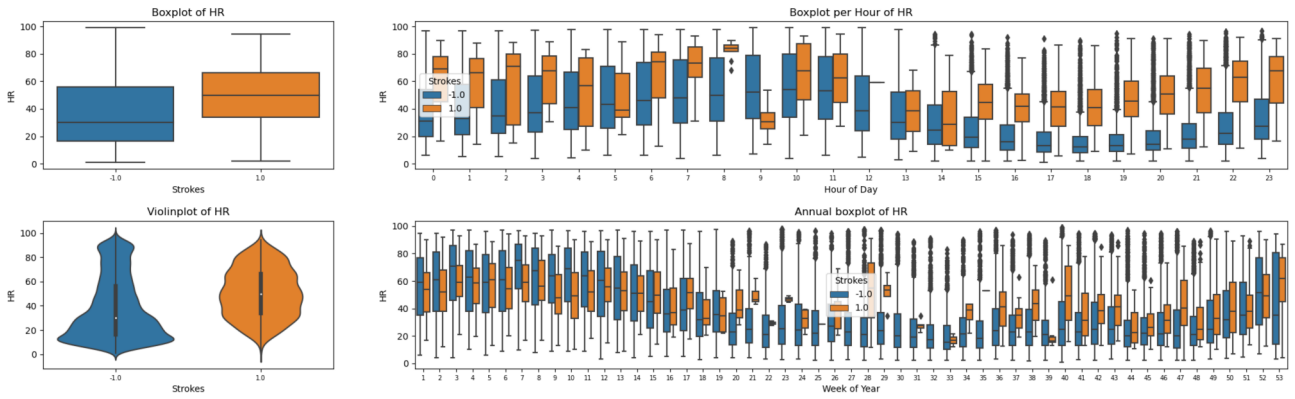


(o) Diagramas de caja y de violín para la Componente vertical de la dirección del viento a 2 m de altura

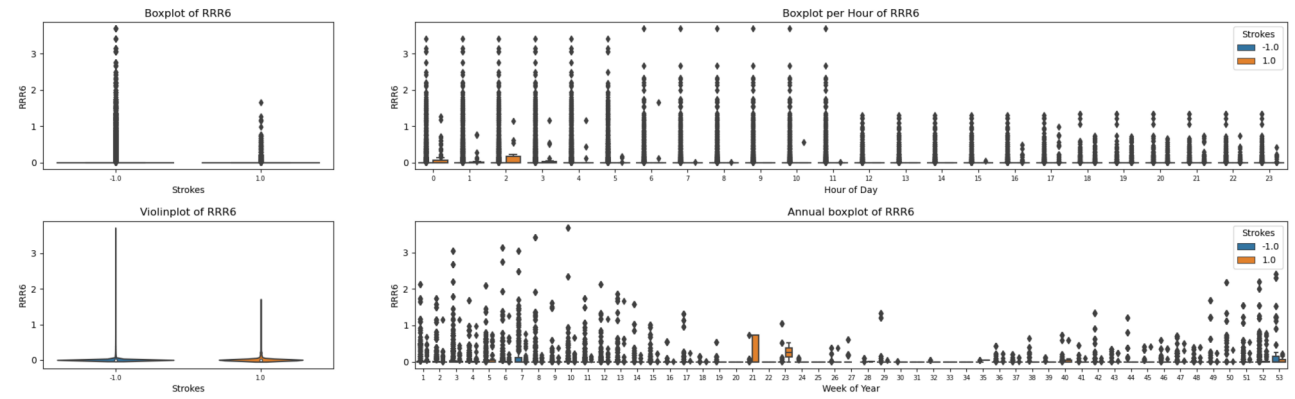


(p) Diagramas de caja y de violín para la Intensidad del viento a 2 m de altura

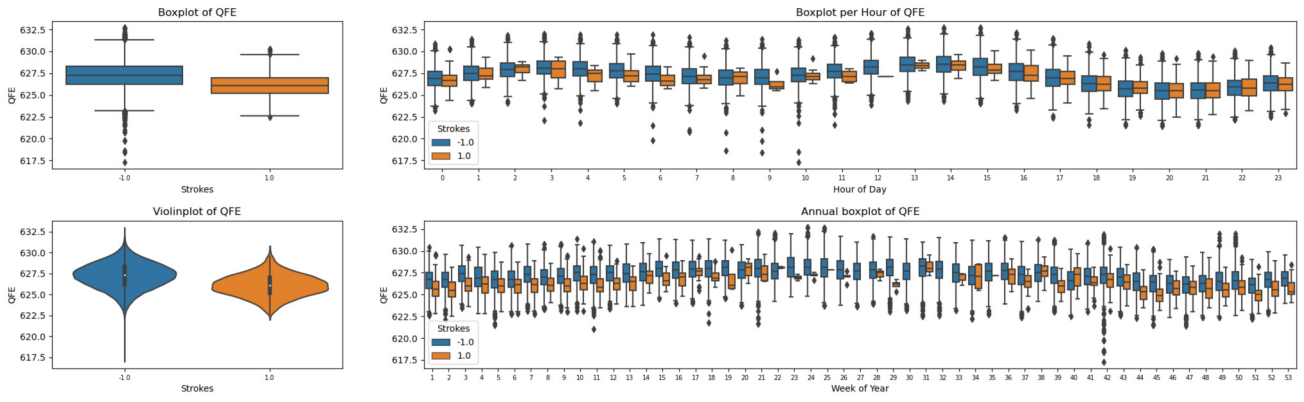
Figura 1: Diagramas de caja y de violín para las características registradas en Visviri, Chile



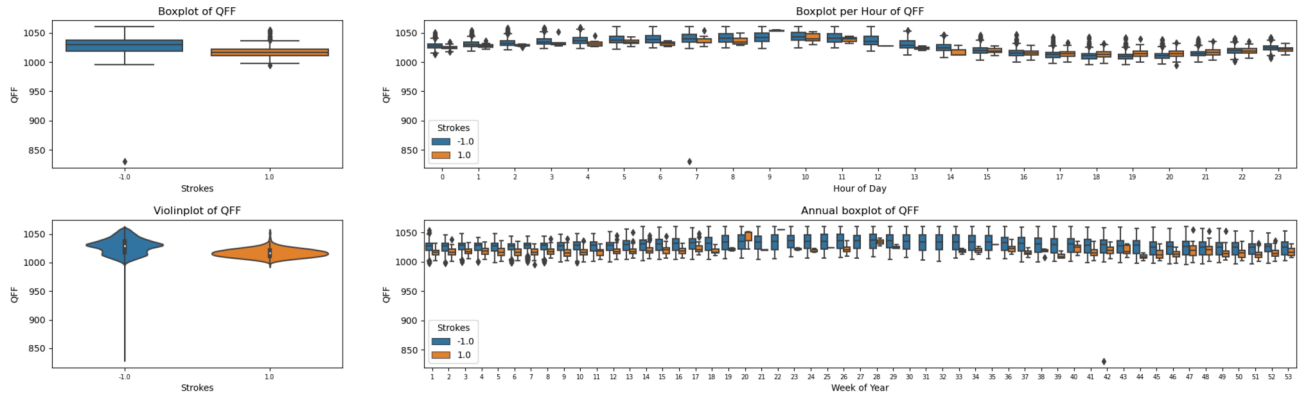
(a) Diagramas de caja y de violín para la Humedad relativa



(b) Diagramas de caja y de violín para la Precipitación

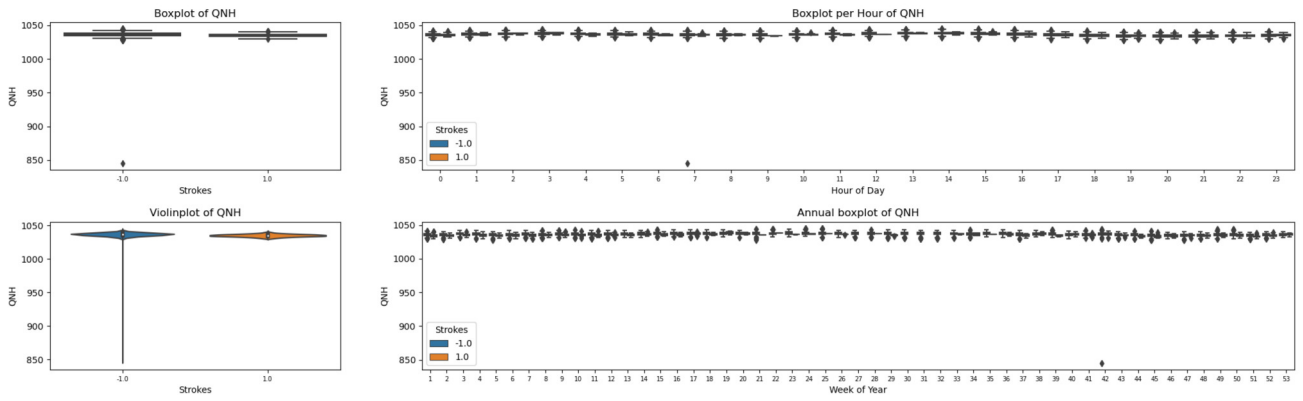


(c) Diagramas de caja y de violín para la Presión a nivel de estación

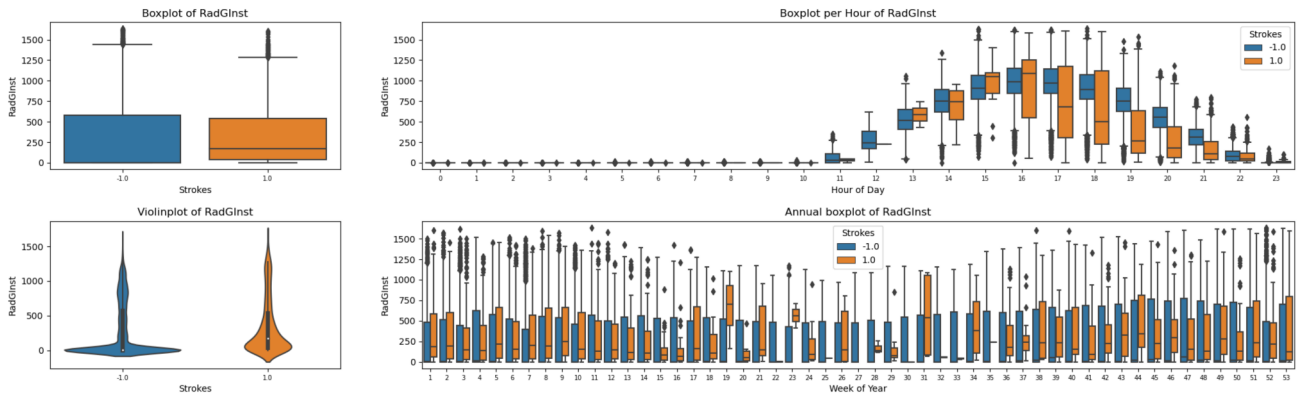


(d) Diagramas de caja y de violín para la Presión a nivel del mar

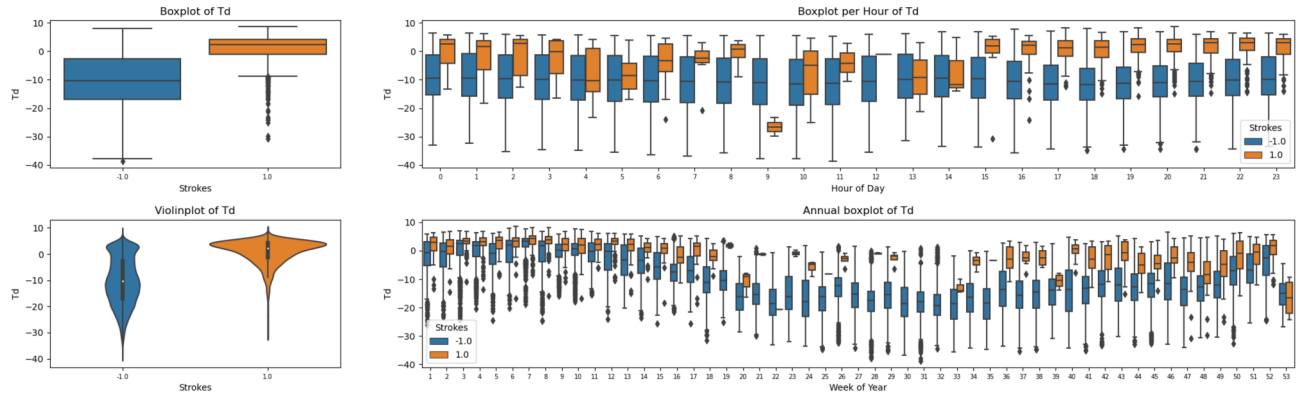
Figura 2: Diagramas de caja y de violín para las características registradas en Visviri cuando ocurre tormenta y cuando es un cielo despejado



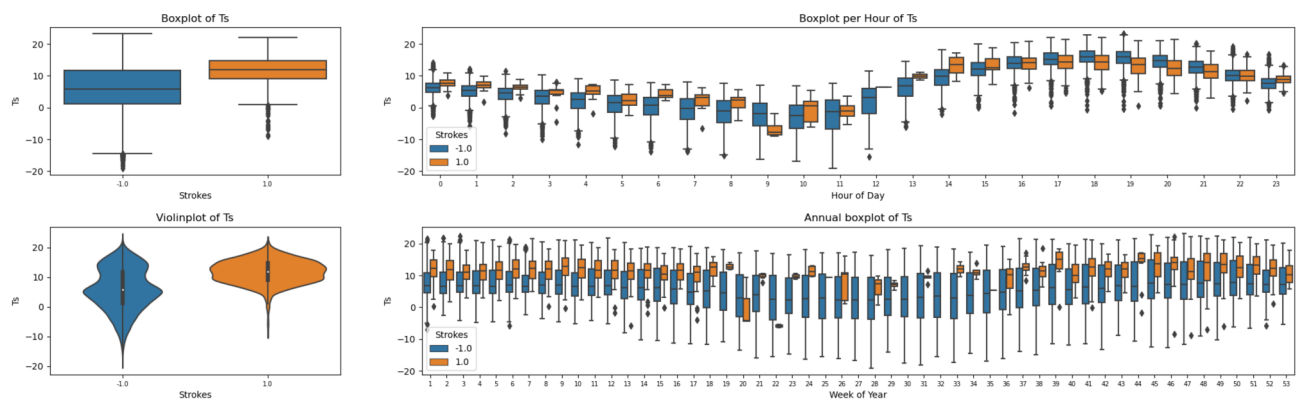
(e) Diagramas de caja y de violín para la Presión a nivel del mar mediante Atmósfera Estándar de la OACI



(f) Diagramas de caja y de violín para la Radiación Solar Instantánea

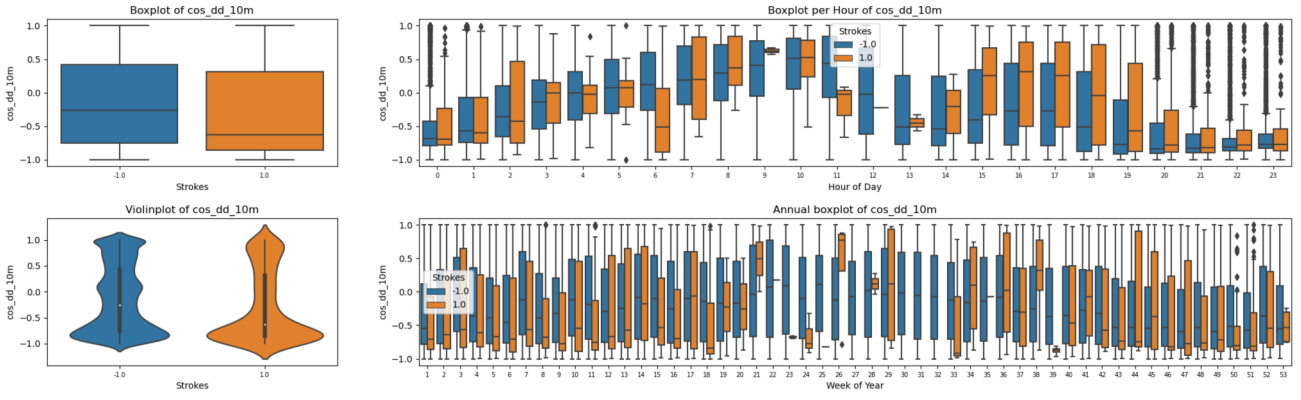


(g) Diagramas de caja y de violín para la Temperatura del punto de rocío

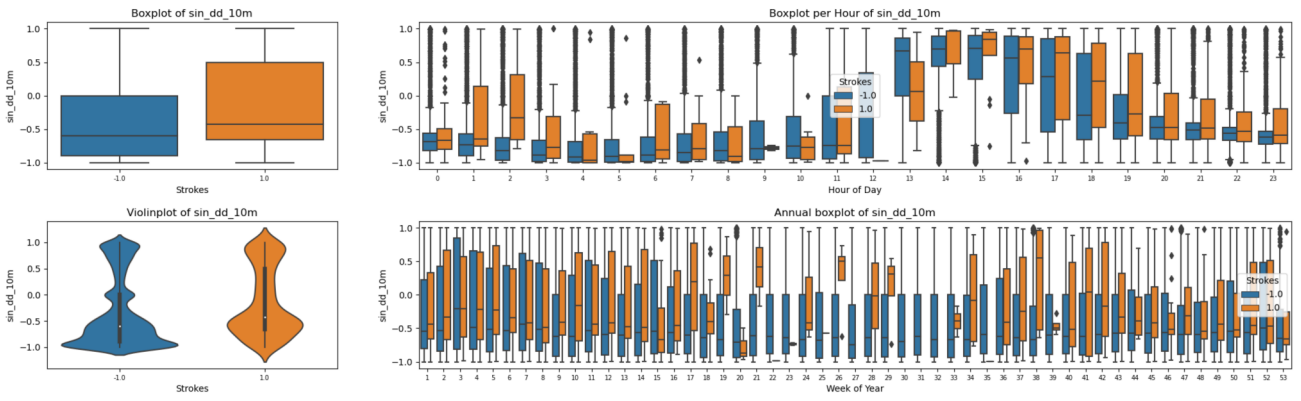


(h) Diagramas de caja y de violín para la Temperatura del aire seco

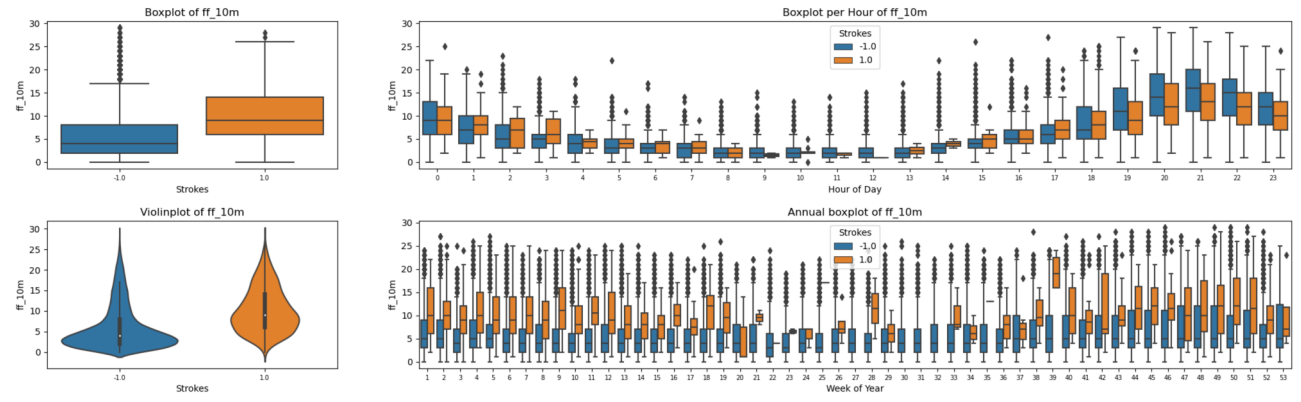
Figura 2: Diagramas de caja y de violín para las características registradas en Visviri cuando ocurre tormenta y cuando es un cielo despejado



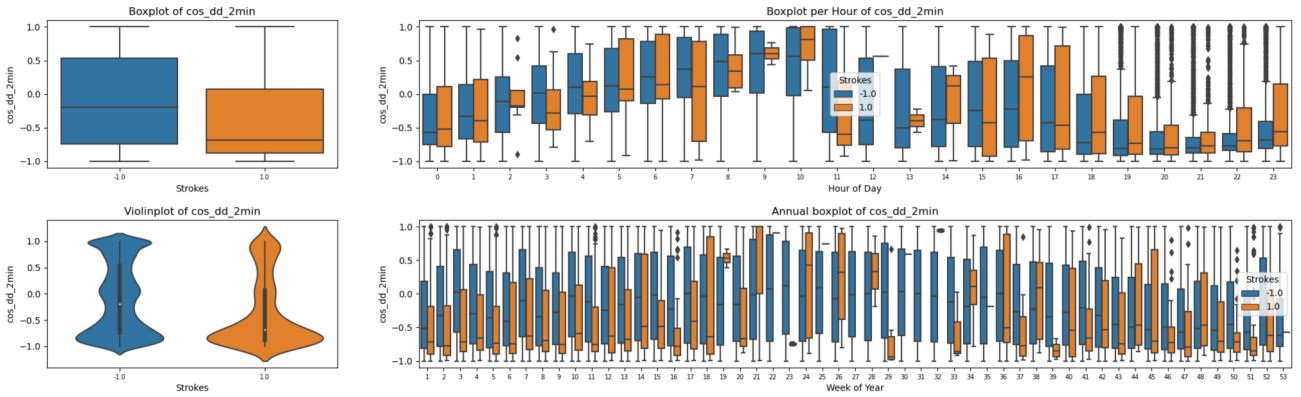
(i) Diagramas de caja y de violín para la Componente horizontal de la dirección del viento a 10 m de altura



(j) Diagramas de caja y de violín para la Componente vertical de la dirección del viento a 10 m de altura

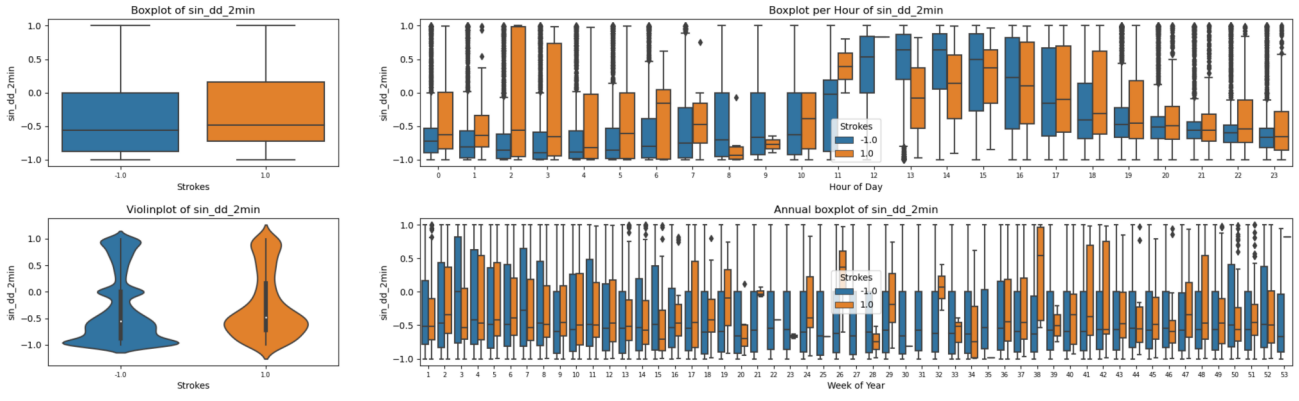


(k) Diagramas de caja y de violín para la Intensidad del viento a 10 m de altura

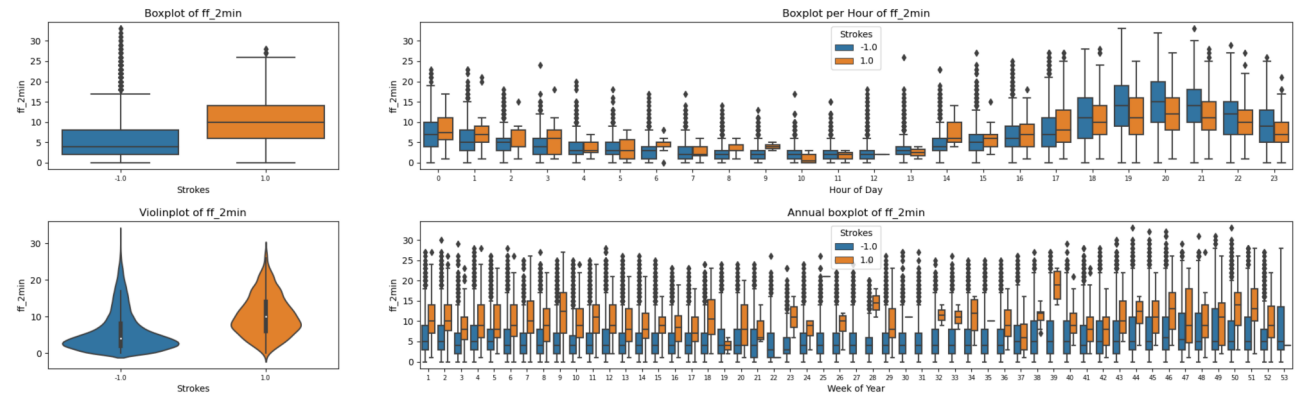


(l) Diagramas de caja y de violín para la Componente horizontal de la dirección del viento promedio cada 2 minutos

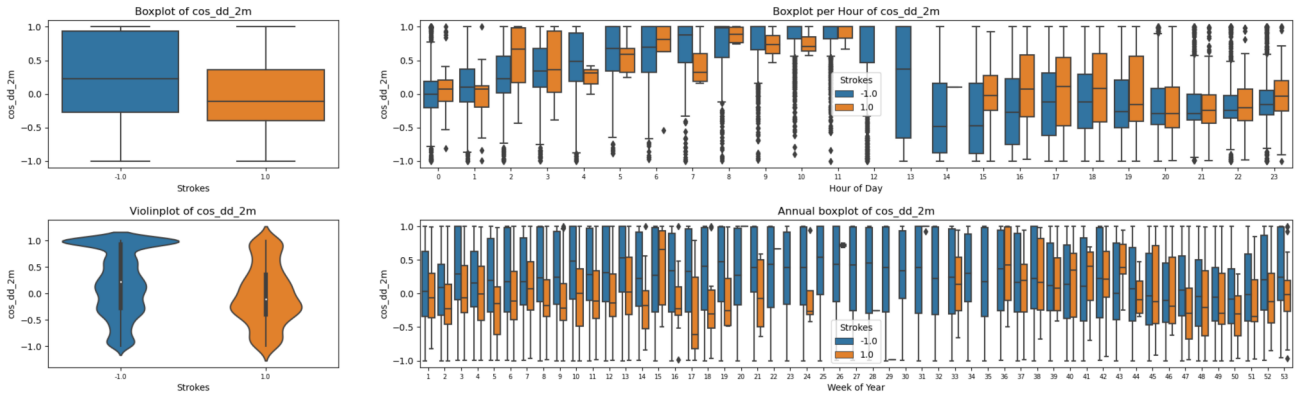
Figura 2: Diagramas de caja y de violín para las características registradas en Visviri cuando ocurre tormenta y cuando es un cielo despejado



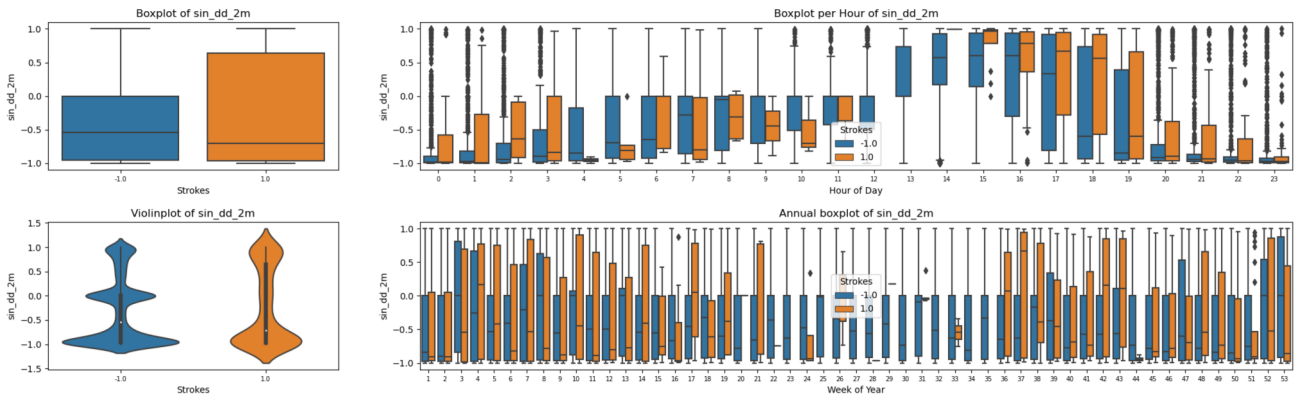
(m) Diagramas de caja y de violín para la Componente vertical de la dirección del viento promedio cada 2 minutos



(n) Diagramas de caja y de violín para la Intensidad del viento promedio cada 2 minutos

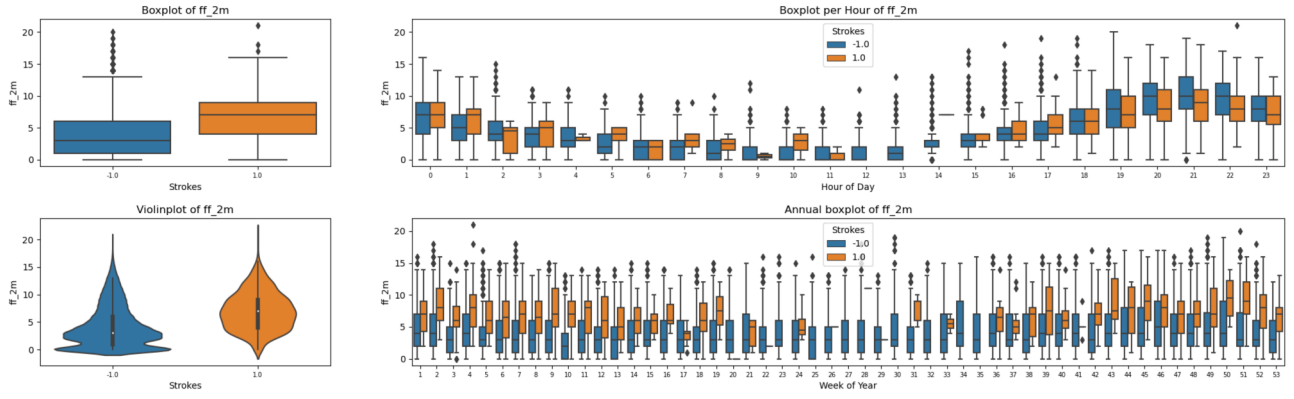


(ñ) Diagramas de caja y de violín para la Componente horizontal de la dirección del viento a 2 m de altura



(o) Diagramas de caja y de violín para la Componente vertical de la dirección del viento a 2 m de altura

Figura 2: Diagramas de caja y de violín para las características registradas en Visviri cuando ocurre tormenta y cuando es un cielo despejado



(p) Diagramas de caja y de violín para la Intensidad del viento a 2 m de altura

Figura 2: Diagramas de caja y de violín para las características registradas en Visviri cuando ocurre tormenta y cuando es un cielo despejado

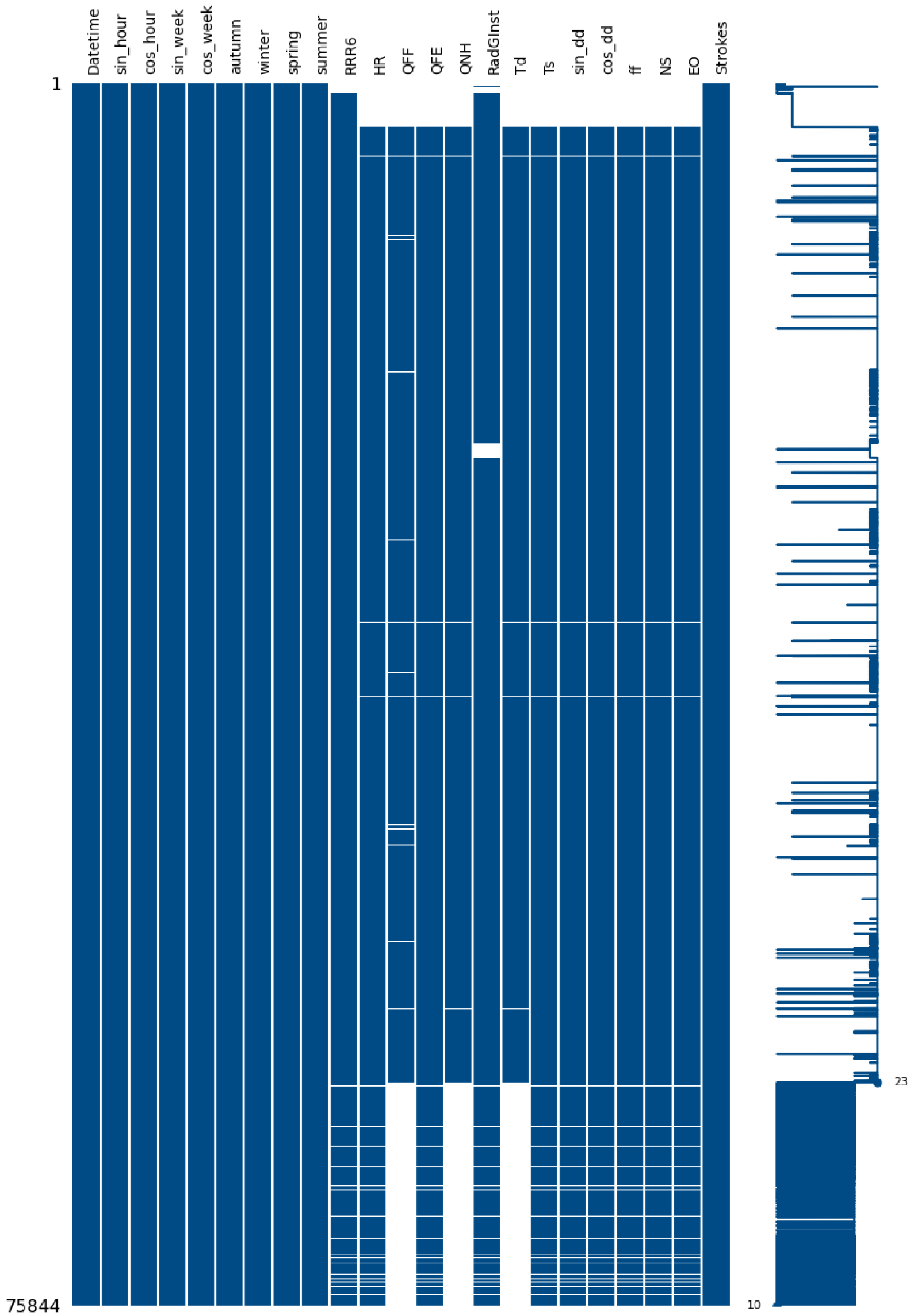
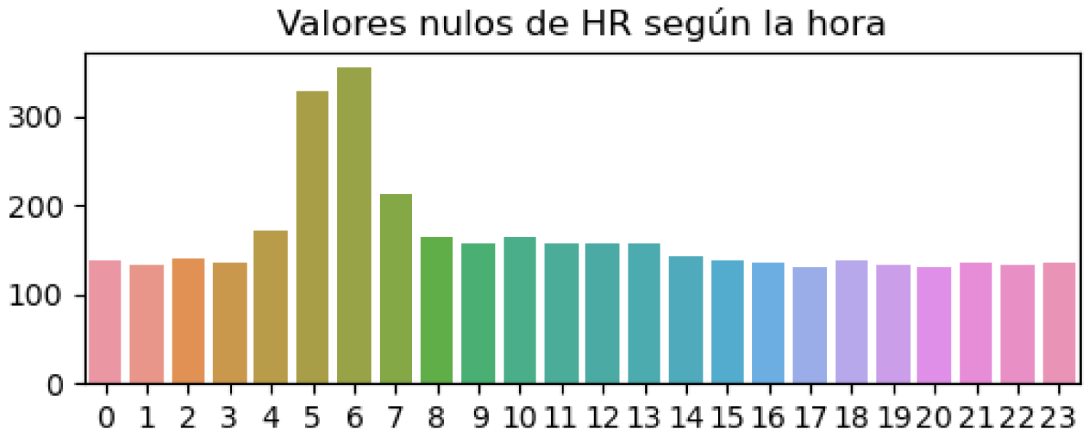
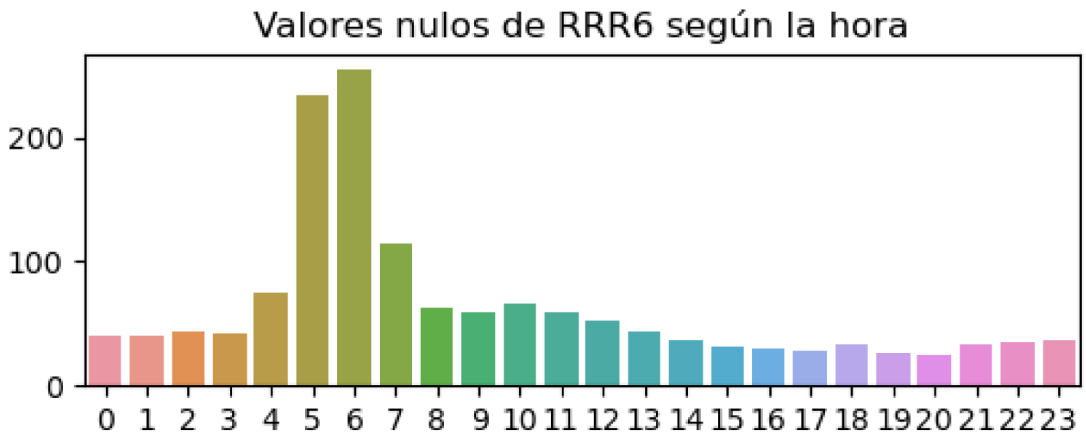


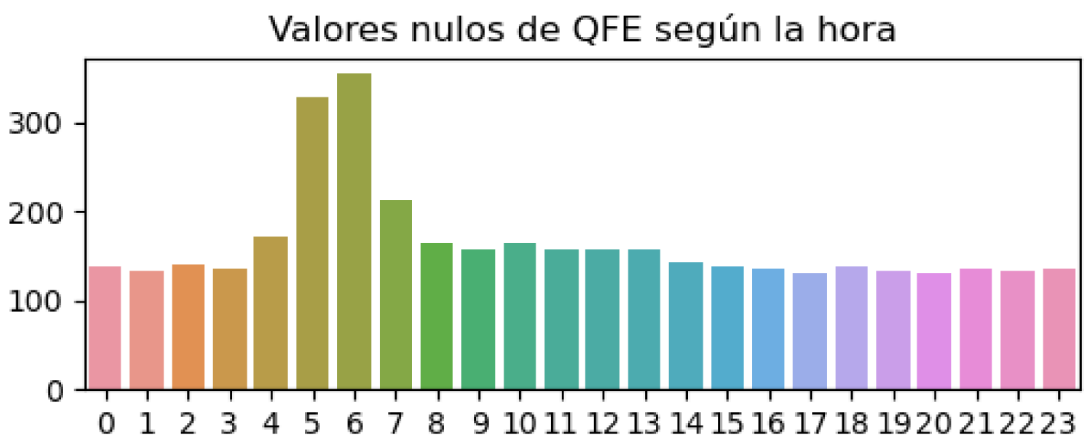
Figura 3: Valores disponibles una vez realizada la ingeniería de características para Visviri, Chile



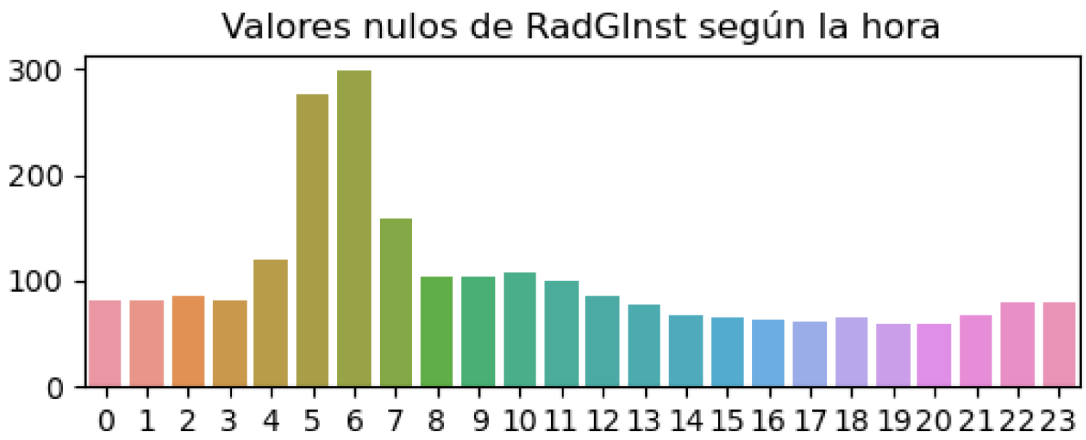
(a) Cantidad de datos no disponibles de Humedad relativa



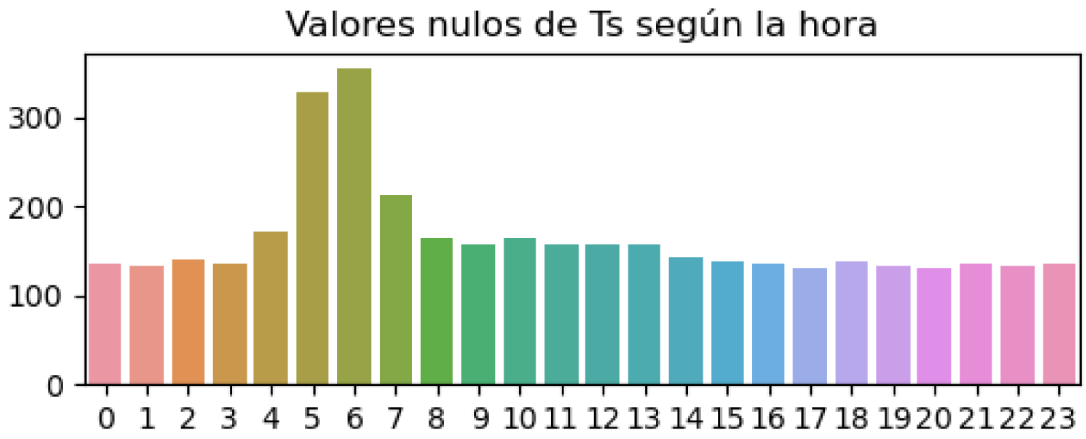
(b) Cantidad de datos no disponibles de Precipitación



(c) Cantidad de datos no disponibles de Presión a nivel de estación

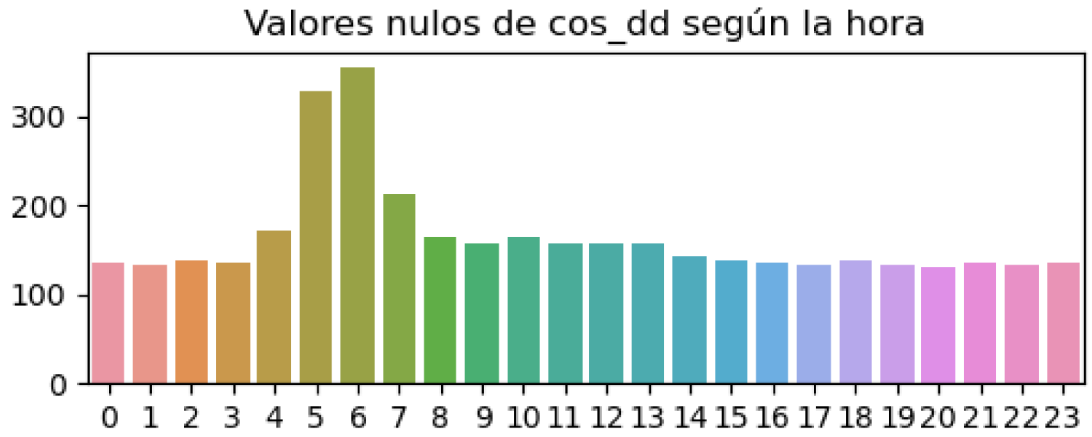


(d) Cantidad de datos no disponibles de Radiación Solar Instantánea

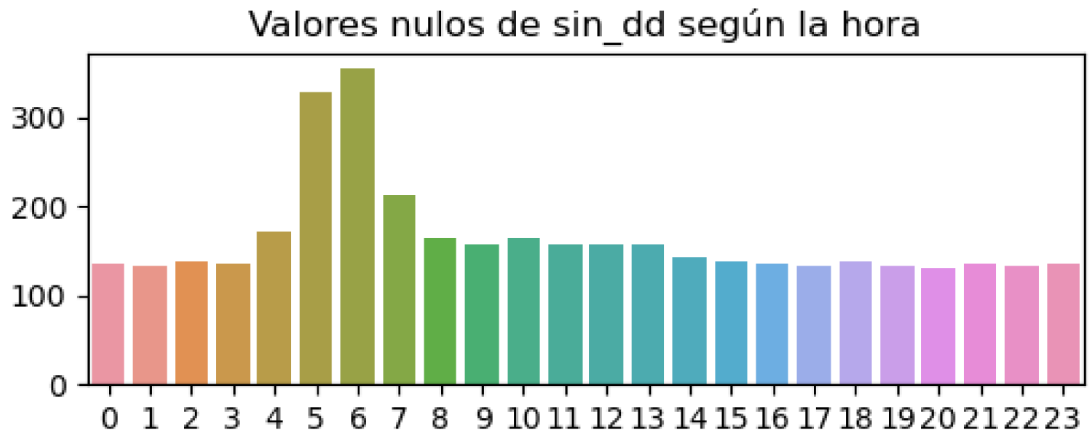


(e) Cantidad de datos no disponibles de Temperatura del aire seco

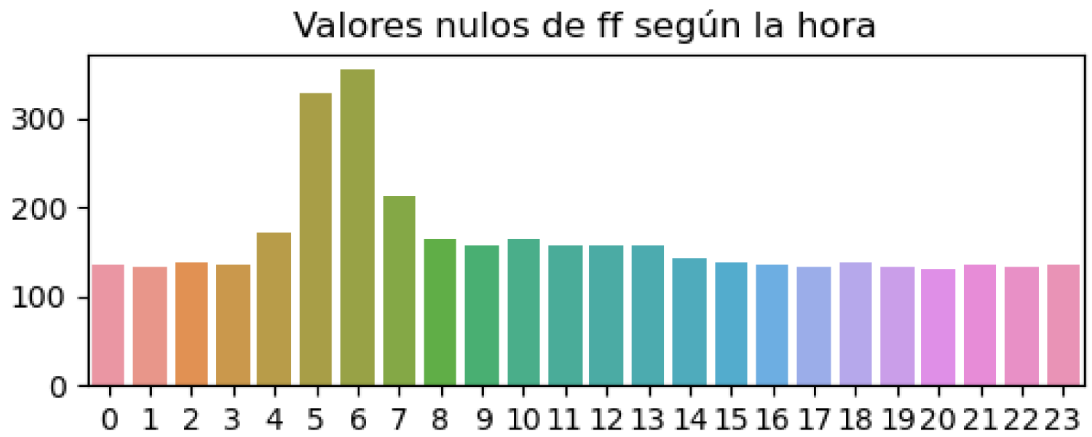
Figura 4: Cantidad de datos no disponibles de las características registradas en Visviri tras la ingeniería de características



(f) Cantidad de datos no disponibles de Componente horizontal de la dirección del viento

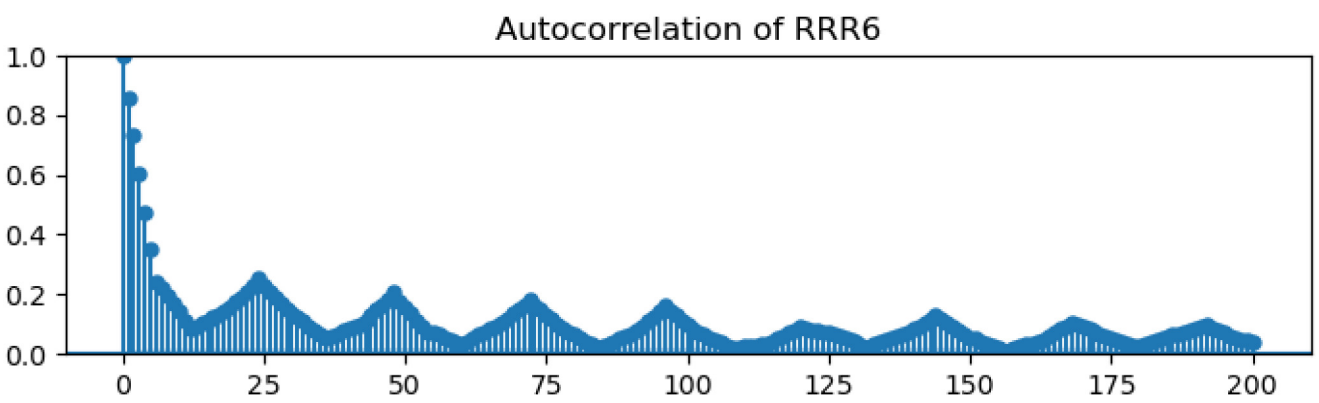


(g) Cantidad de datos no disponibles de Componente vertical de la dirección del viento

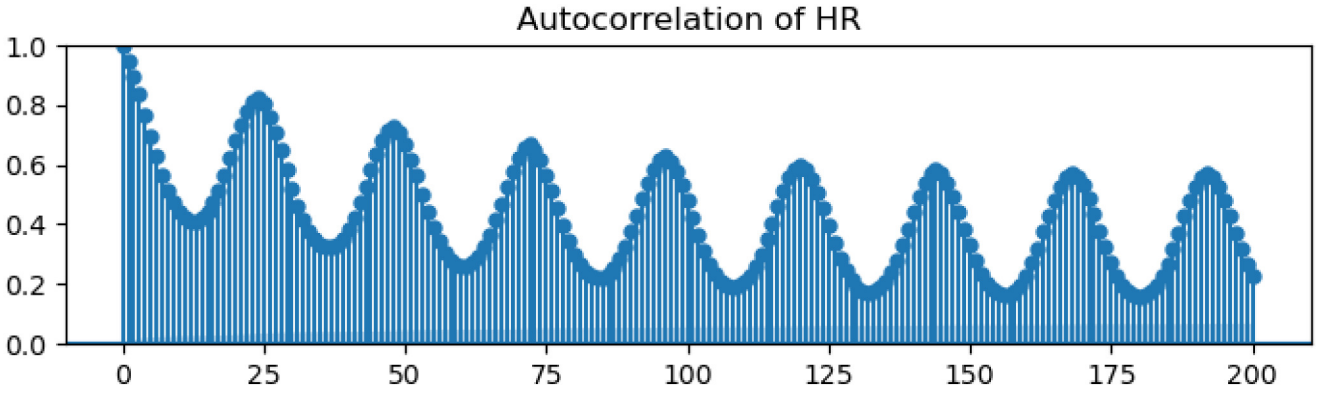


(h) Cantidad de datos no disponibles de Intensidad del viento

Figura 4: Cantidad de datos no disponibles de las características registradas en Visviri tras la ingeniería de características

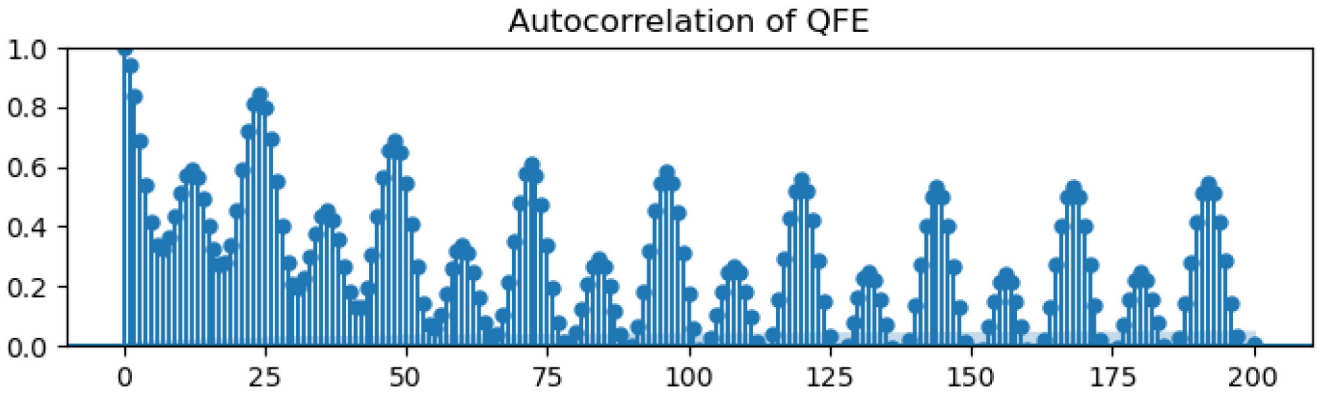


(a) Autocorrelación para la Precipitación

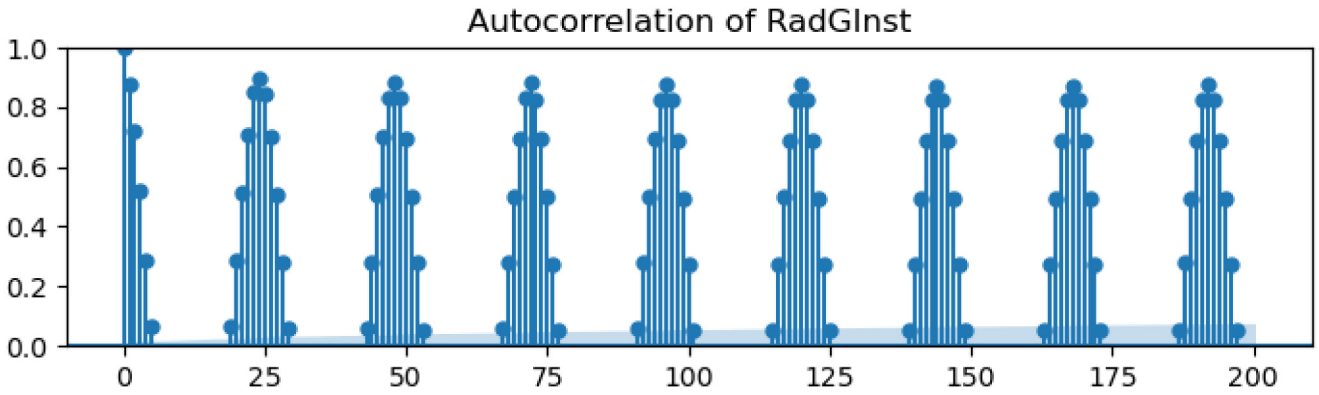


(b) Autocorrelación para la Humedad relativa

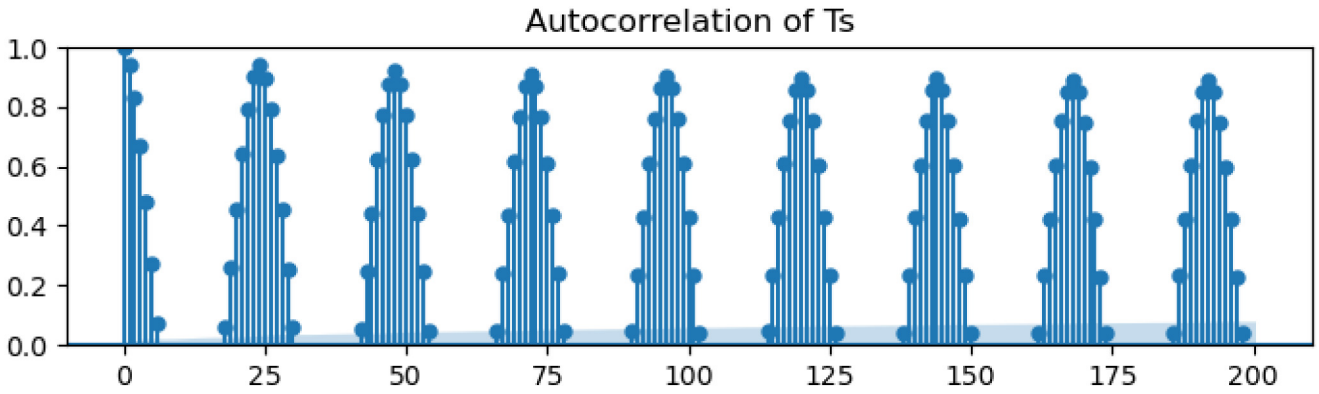
Figura 5: Autocorrelación para las características registradas en Visviri



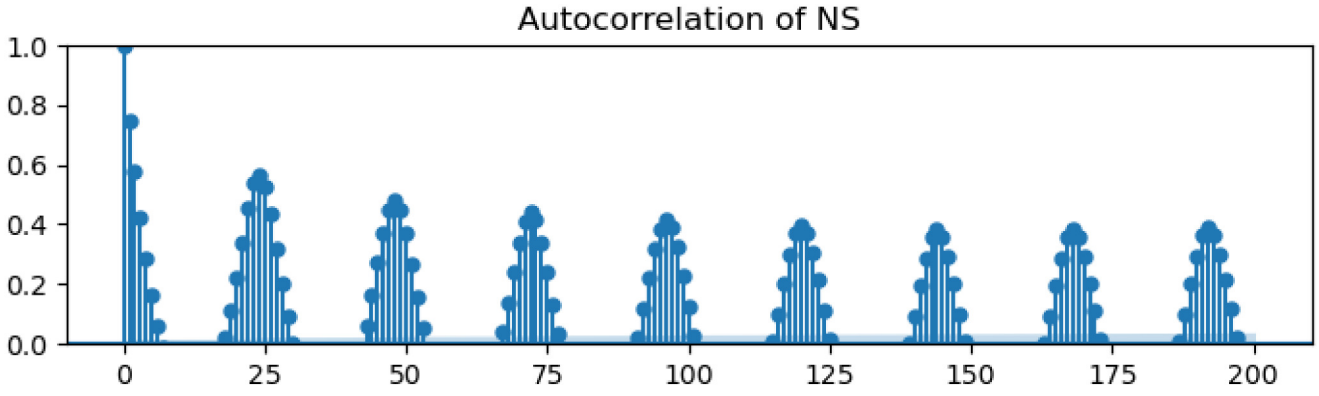
(c) Autocorrelación para la Presión a nivel de estación



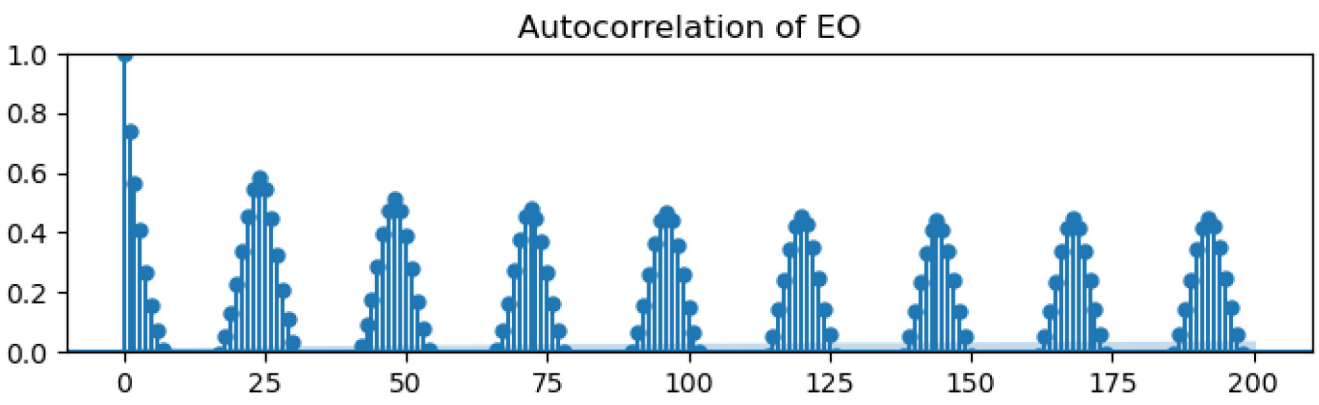
(d) Autocorrelación para la Radiación Solar Instantánea



(e) Autocorrelación para la Temperatura del aire seco

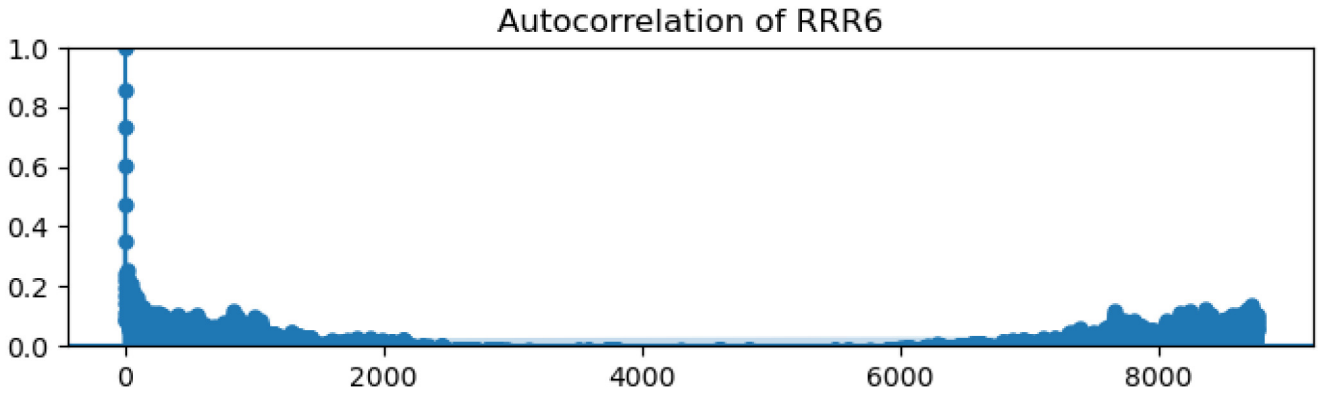


(f) Autocorrelación para la Componente Norte-Sur del viento

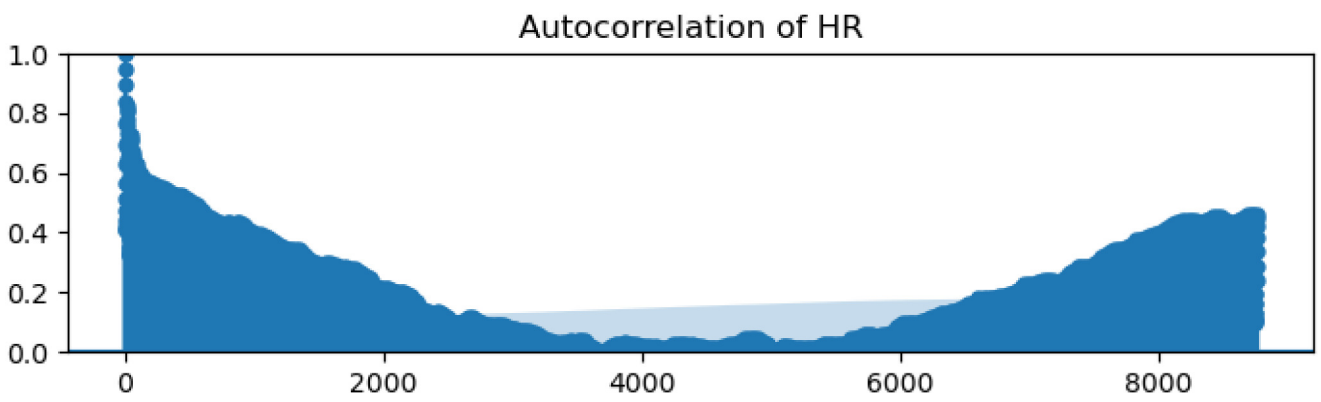


(g) Autocorrelación para la Componente Este-Oeste del viento

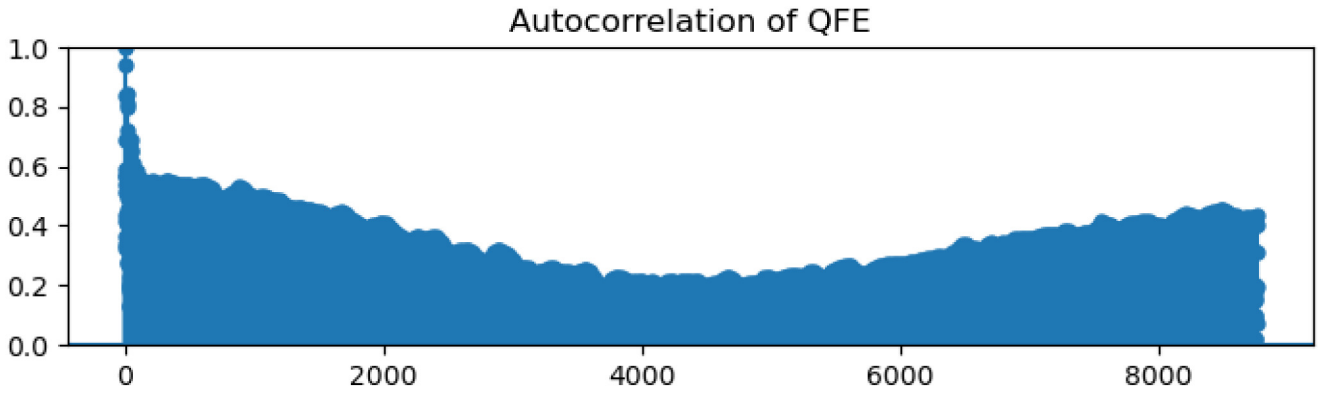
Figura 5: Autocorrelación para las características registradas en Visviri



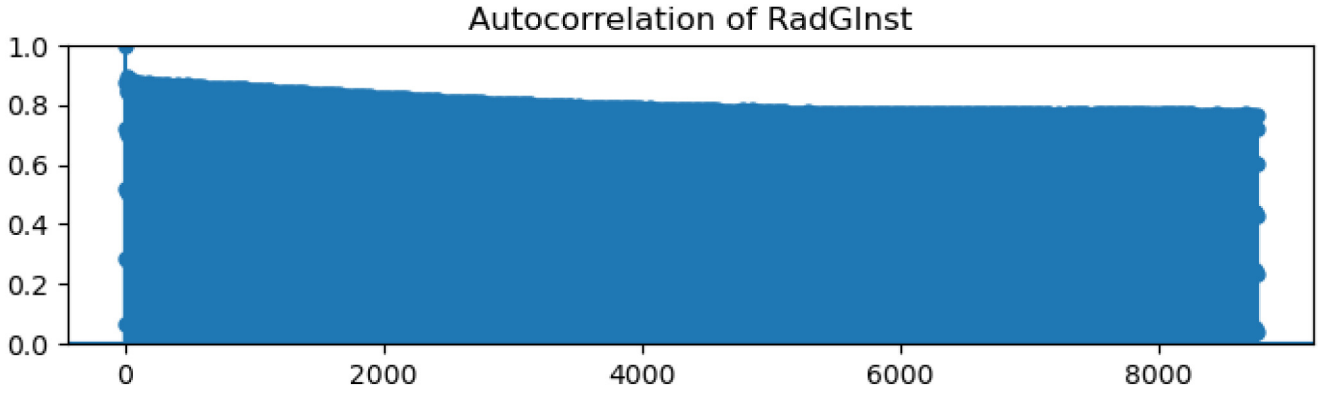
(h) Autocorrelación para la Precipitación, horizonte 1 año



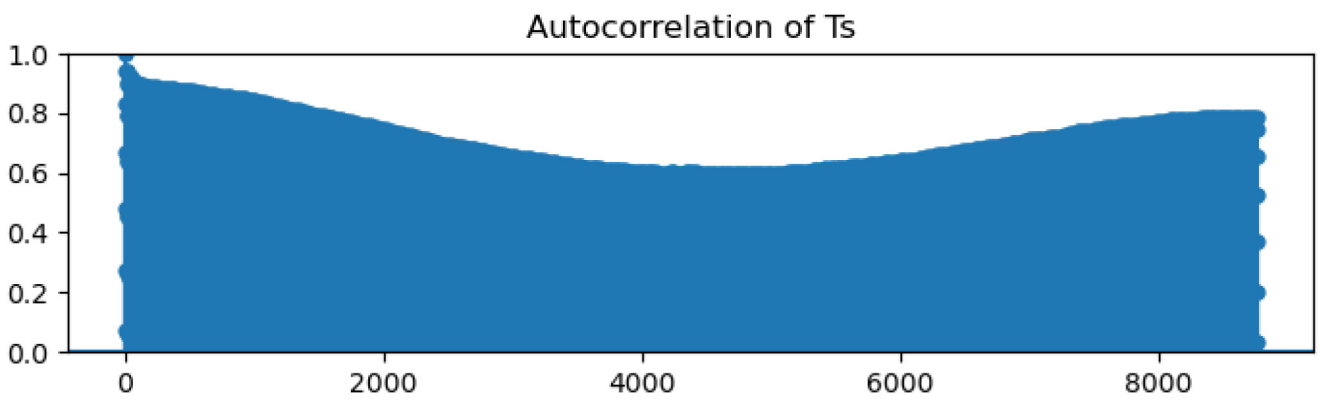
(i) Autocorrelación para la Humedad relativa, horizonte 1 año



(j) Autocorrelación para la Presión a nivel de estación, horizonte 1 año

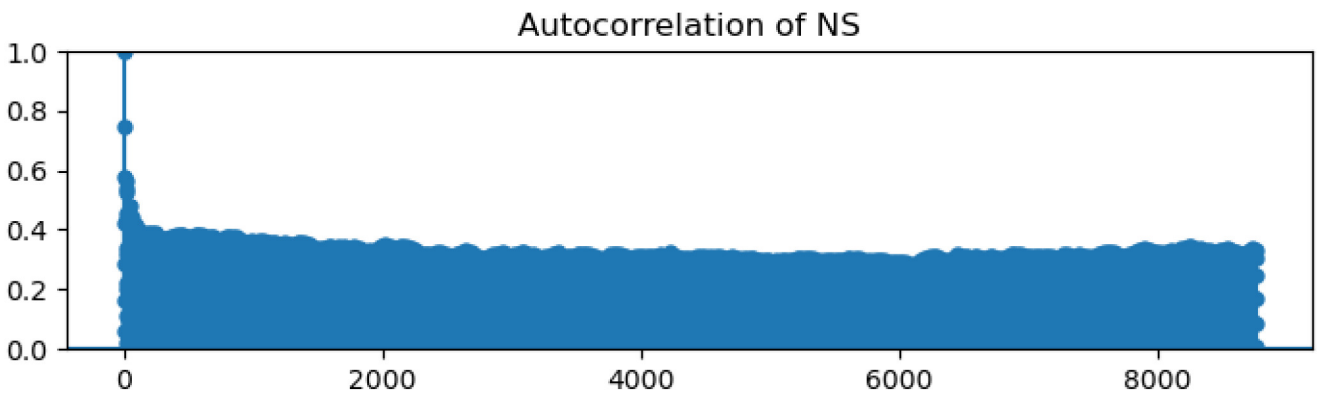


(k) Autocorrelación para la Radiación Solar Instantánea, horizonte 1 año

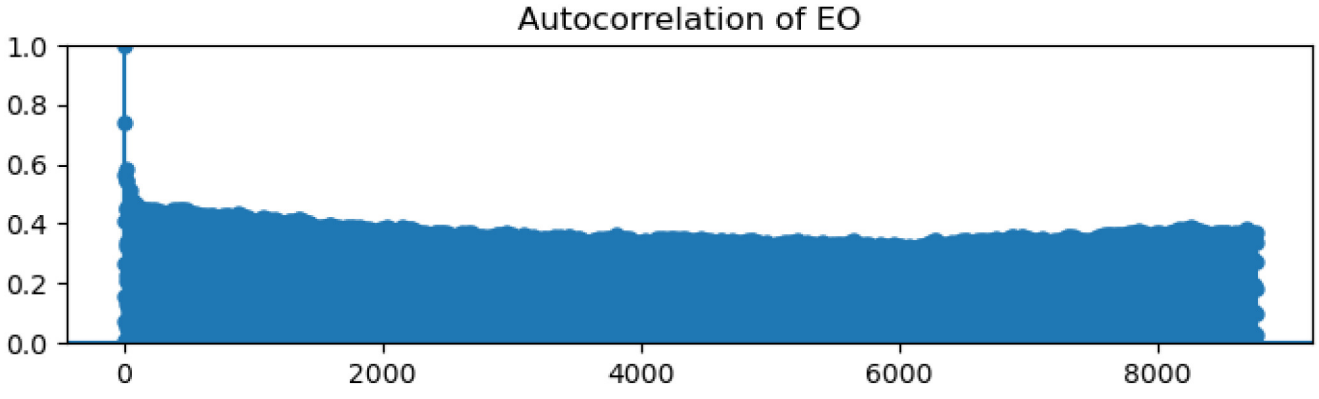


(l) Autocorrelación para la Temperatura del aire seco, horizonte 1 año

Figura 5: Autocorrelación para las características registradas en Visviri



(m) Autocorrelación para la Componente Norte-Sur del viento, horizonte 1 año



(n) Autocorrelación para la Componente Este-Oeste del viento, horizonte 1 año

Figura 5: Autocorrelación para las características registradas en Visviri