Universidad Nacional de San Juan

Facultad de Ciencias Exactas, Físicas y Naturales

Tesis de Maestría en Informática

Análisis de Fenómenos en Estaciones Agrometeorológicas mediante Ciencia de Datos

Autora: Lic. María Isabel Masanet Yañez

Director: Mg. Ing. Raúl Oscar Klenzi

Junio de 2023

San Juan, Argentina

Agradecimientos

A la Universidad Nacional de San Juan, y en particular a la Facultad de Ciencias Exactas, Físicas y Naturales por la formación académica y profesional brindada.

Al asesor, Mg. Ing. Raúl Klenzi, por su constante acompañamiento en este camino recorrido.

Al Dr. Ing. Flavio Capraro por su colaboración en este trabajo.

A la Mg. Lic. Sonia Silva, responsable del Servicio de Agrometeorología de la Estación Experimental INTA, por su predisposición en aclarar las dudas.

A mis amigos y compañeros de trabajo que hoy se alegran por este logro.

A mi familia, especialmente a mis hijas, por el apoyo incondicional que me brindan.

A mis padres por el esfuerzo que hicieron para que estudie y los ejemplos de vida que me dieron.

A mi pequeña nieta, porque su llegada me impulsó a terminar esta tesis.

A todos… Gracias!

Índice de contenidos

Capítulo 1

Introducción

1. Introducción

Las Tecnologías de Información y Comunicación (TIC) se han transformado en poderosas herramientas que permiten la organización y el acceso al conocimiento, lo que ha llevado a que paulatinamente se incorporen a distintos ámbitos. La agricultura forma parte de esta transformación, implementando el uso de nuevas tecnologías para facilitar y mejorar el desarrollo de sus actividades, como son los sistemas de riego por goteo, las maquinarias para la siembra y el laboreo, las cosechadoras mecanizadas, entre otros [1].

La actividad agrícola debe enfrentar la problemática que presentan distintos fenómenos meteorológicos adversos como el viento, el granizo, las heladas, entre otros. Estos factores no se pueden evitar y tienen gran relevancia para el productor y el desarrollo productivo de una región. La protección de las plantas contra los efectos letales de las bajas temperaturas es muy importante en la agricultura.

La helada se produce cuando la superficie terrestre y el aire que se siente sobre ella alcanza una temperatura por debajo de los 0°C [1] [2]. Generan daños significativos en la actividad agrícola, causando pérdidas de cosechas de todo un año y comprometiendo los ingresos del año siguiente [3].

En algunos casos se usan invernaderos para resguardar el cultivo, pero esto no es factible en plantas altas, como el almendro, una de las producciones más afectadas por las heladas tardías porque es la especie frutal de floración más temprana [4]. En estos casos los productores deben adoptar mecanismos de defensa en el preciso momento que sucede el fenómeno meteorológico para evitar que los cultivos sean dañados. Dichos mecanismos requieren de una previsión de recursos y de una preparación con antelación a la ocurrencia de la adversidad.

Es importante el pronóstico de la ocurrencia y la magnitud de las heladas, brindando a los productores la oportunidad de llevar a cabo acciones de protección que permitan mitigar el daño que el fenómeno provoca. Iniciar con los mecanismos de defensa en tiempo y forma es importante para evitar pérdidas en los cultivos que resultan de poner en marcha demasiado tarde, cuando ya está helando; por otra parte,

ahorra energía al reducir el tiempo de operación de los diversos métodos en cualquier momento durante la noche.

Siempre se espera un alto porcentaje de certeza en la predicción, un caso de falso positivo ocasiona pérdida de recursos en la preparación de la defensa. Pero lo más grave son las consecuencias de un falso negativo, significa no aplicar mecanismo cuando se produce el fenómeno meteorológico ocasionando daño en el cultivo. El momento crítico para comenzar cualquier actividad de protección contra heladas es una hora antes del momento en que se espera la temperatura de daño crítico [5]. Esto no es perfecto, pero es bastante preciso.

Los valores de los datos meteorológicos son particulares de una zona (fenómenos locales), presentando variaciones entre ellas a pesar de que se encuentren próximas, por ejemplo, los balances de radiación en una zona de la superficie terrestre dependen de la ubicación sobre la Tierra, porque la inclinación de los rayos solares que llegan a la zona influye en la cantidad de energía que ésta recibe [6]. Ante esto, resulta necesario conocer los datos meteorológicos específicos del lugar donde se encuentra el cultivo para pronosticar la ocurrencia de un factor meteorológico.

El productor puede valerse de la experiencia propia o de la opinión de un experto para pronosticar un factor meteorológico adverso. En cualquier caso, requiere conocer los valores de las variables meteorológicas de la zona donde se encuentra el cultivo. Esto puede ser constatado personalmente en el lugar o a través de una herramienta tecnológica que monitoree los datos meteorológicos para predecir la adversidad.

La verificación in situ de las variables suele transformarse en un inconveniente para la persona, debido a que exige una intensa dedicación de tiempo y esfuerzo, además en algunos casos implica el traslado de una distancia considerable para llegar al campo cultivado. En cambio, a través del uso de una herramienta tecnológica mejora la certeza de la predicción, se optimiza el uso de los recursos y el trabajo de los productores.

1.1. Impacto económico de las heladas

La helada ha dañado la belleza y la producción de cultivos en países como Argentina, Australia, Georgia, México, China, Corea del Sur, Irán, Estados Unidos, Italia, Alemania, Berlín y muchos otros [7].

La helada es un fenómeno localizado, puede ocasionar daño parcial en diferentes niveles en el mismo campo de cultivo. Puede destruir toda la producción en cuestión de horas, provocando pérdidas de la cosecha de un año entero y comprometiendo ingresos del siguiente, sobre todo para fruticultores y viticultores. Incluso si el daño no es visible inmediatamente después del evento, los efectos pueden surgir en el fin de temporada, reduciendo tanto la cantidad como la calidad de la cosecha [8].

En 2.013, en Chile, los productores de frutas y cultivos sufrieron importantes eventos de heladas. Los efectos de las heladas fueron devastadores para la agricultura en general. La pérdida económica fue entre US$ 600 a US$ 900 millones. El volumen exportado disminuyó entre un 15% a 20%. El impacto negativo sobre el empleo de temporeros agrícolas tuvo una baja cercana al 20% [3] [7]. Las pérdidas fueron estimadas entre el 40% y el 50% en Región del Libertador General Bernardo O'Higgins, Metropolitana, Maule y Biobío. En menor magnitud, aproximadamente un 15%, en las Regiones de Atacama, Coquimbo y Valparaíso el 15% [9].

Entre las diversas condiciones atmosféricas que pueden producir daños a la economía en Brasil, la helada puede considerarse entre aquellas con los mayores impactos. Afecta a la agroindustria y sectores relacionados. Durante los años con altas tasas de incidencia de heladas, hay una caída sustancial en la producción (posiblemente incluyendo los años siguientes) y, en consecuencia, un aumento de los precios, tanto en el mercado interno como en el externo. Un ejemplo clásico fue un episodio de escarcha en la noche del 16 de julio de 1.975, en el sur / sureste de Brasil, que causó un fuerte caída en la producción de café y un aumento de casi 200% en precio [10].

En Argentina, el trabajo de investigación realizado por el Instituto de Tecnología de Karlsruhe (KIT) advierte que los viñedos en Mendoza y San Juan representan las regiones de mayor riesgo en el mundo por condiciones climáticas extremas y peligros naturales. Esta realidad cuantifica uno de los aspectos que puede generar un evento de

heladas, pero las consecuencias socioeconómicas afectan no solo a los productores, sino también al transporte, el comercio y los servicios generales, que requieren largos períodos de recuperación [8].

Durante el año 2.020, en la provincia de Mendoza, las heladas tardías afectaron treinta mil hectáreas de vid y dieciséis mil hectáreas de frutales. En el caso del cultivo de vid, el daño promedio estimado fue del 24%, mientras que en el resto de las frutas fue del 84% [11]. El año siguiente, más precisamente el 4 de octubre de 2.021, se produjo una helada tardía que afectó a provincias de la Región de Cuyo. En algunas zonas de Mendoza la temperatura descendió hasta los -5°C. En San Juan la temperatura alcanzó durante la mañana los -1.8°C, comenzando el fenómeno alrededor de las 6:00 horas y prolongándose hasta las 7:30 horas aproximadamente, ocasionando daños en los parrales [12].

1.2. Problemática

Las heladas, dependiendo de la intensidad y duración de las mismas y el estado fenológico del cultivo, provocan daños de magnitudes variables. Pueden tener un efecto drástico para la planta entera o pueden afectar únicamente a una pequeña parte del tejido de la planta, lo cual reduce el rendimiento o deprecia la calidad del producto [13]. La protección de las plantas contra los efectos de las bajas temperaturas letales es esencial en la agricultura, especialmente en la producción de frutas y verduras de alto valor.

En San Juan, hacia fines de agosto y primera semana de septiembre, se producen heladas meteorológicas tardías que afectan los cultivos en floración y cuaje, provocando daños importantes en los frutos recién formados por ser los de mayor sensibilidad a las bajas temperaturas [14]. Además, los inviernos presentan días con temperaturas benignas que aumentan la sensibilidad a las heladas. En general, las heladas que ocurren en San Juan son del tipo mixtas [14].

En el Valle de Tulum, provincia de San Juan, se cultiva el almendro, producción que resulta muy afectada por las condiciones climáticas adversas de las heladas tardías y viento Zonda. Se trata de un cultivo muy sensible a las temperaturas frías de finales de

invierno y principios de primavera, que pueden dañar e incluso destruir completamente la cosecha de almendras [4].

El clima es un proceso multidimensional, dinámico y caótico, y estas propiedades hacen que el pronóstico sea un gran desafío debido a la naturaleza no lineal de los datos meteorológicos [9]. Por esto el estudio del pronóstico del tiempo moderno implica una combinación de modelos informáticos complejos, observaciones in situ y el conocimiento de las tendencias y patrones climáticos mediante una metodología contemporánea avanzada, que ha llamado la atención de investigadores y científicos en diversos campos y disciplinas.

Predecir cómo la temperatura puede cambiar durante la noche es útil para la protección contra heladas, ayudando a los agricultores a decidir si es necesaria la protección del cultivo y cuándo poner en marcha sus sistemas que la efectúan. La idea principal en el pronóstico de las temperaturas críticas es que no deberían considerarse como absolutamente correctas, sino como una directriz para la toma de decisiones sobre cuándo poner en marcha o detener los métodos activos de protección de los cultivos [13].

Primero hay que consultar los servicios meteorológicos locales para determinar si hay previsiones disponibles. Los servicios meteorológicos tienen acceso a más información y utilizan modelos sinópticos y/o de meso-escala para proporcionar pronósticos regionales. Normalmente, los pronósticos locales (micro-escala) están menos disponibles, a no ser que los proporcionen servicios de predicción privados[13].

Actualmente la problemática que afrontan los productores es que, prevén la ocurrencia de la helada en la localidad donde se encuentra el cultivo, a partir de la experiencia propia o de algún experto (productor o climatólogo). El productor no dispone de un medio o herramienta tecnológica que realice pronósticos locales de las heladas, que sirva de apoyo a la toma de decisiones para el despliegue de las medidas de mitigación del fenómeno, optimizando el uso de recursos para los mecanismos de defensa y evitando pérdidas en la producción. Lo que se traduce en un mayor desarrollo productivo y económico de la zona.

1.3. Objetivos

1.3.1. Objetivo General

El objetivo general de este trabajo es:

Desarrollar y analizar comparativamente modelos de predicción del fenómeno de la helada basados en algoritmos de ciencia de datos a partir de las variables meteorológicas registradas por estaciones agrometeorológicas instaladas en la provincia de San Juan.

1.3.2. Objetivos Específicos

Los objetivos específicos son:

1) Analizar las variables meteorológicas obtenidas a partir de las mediciones que realizan las estaciones agrometeorológicas.
2) Identificar el conjunto de valores adecuado para los principales parámetros de los modelos de predicción.
3) Desarrollar modelos de predicción de heladas para la zona geográfica donde se encuentra la estación agrometeorológica.
4) Evaluar y comparar modelos de predicción desarrollados.
5) Visualizar de forma gráfica los resultados obtenidos.

Capítulo 2

Antecedentes

2. Antecedentes

El pronóstico de fenómenos meteorológicos o de alguna variable en particular ha sido objeto de investigación en distintas partes del mundo.

Jorenoosh y Sepaskhah [5] realizaron un estudio con el objetivo de predecir las temperaturas mínimas diarias en las zonas semiáridas de Bajgah y Kooshkak, provincia de Fars, Irán. Desarrollaron un modelo de regresión simple que, a partir de parámetros meteorológicos diarios, predecía la temperatura mínima en horario de la madrugada del día siguiente. Los datos fueron obtenidos de 27 estaciones meteorológicas. En la evaluación de la validez utilizaron el análisis estadístico error cuadrático medio normal (NRMSE), error absoluto medio (MAE), error medio (ME), índice de acuerdo (d), coeficiente de Nash-Sutcliff (CNS) y R^2. El trabajo concluye que el punto de rocío y la humedad relativa son las variables que determinan la temperatura mínima de la madrugada del día siguiente. Y que, los parámetros climáticos diarios, como la velocidad del viento, la evaporación, las horas de sol y la lluvia, no muestran un efecto significativo sobre la temperatura mínima.

En Noruega, Abrahamsen et all. [15], obtuvieron datos meteorológicos (con una API Python) del Instituto Meteorológico Noruego a través del sitio web frost.met.no. Entrenaron y ajustaron distintas redes neuronales artificiales autorregresivas (AR-ANN) mediante la librería TensorFlow en Python. Los modelos predecían la temperatura con horizontes de 1, 3, 6 y 12 horas. Un primer experimento fue una red neuronal autorregresiva (AR-NN) en la que utilizaron solo los datos de temperatura como entrada. En un segundo experimento, los datos de temperatura y precipitación se introdujeron en la red, formando una red neuronal autorregresiva con entrada exógena (ARXNN); constituyendo en total ocho modelos (dos por cada horizonte de predicción). Concluyeron que la introducción de la precipitación como una entrada en el modelo ARX mejoraba ligeramente el rendimiento de la predicción.

Un estudio sobre la previsión de las heladas en huertos de manzano en la región de Tirol del Sur, en Italia, utilizó un modelo de promedio móvil integrado autorregresivo no estacional (Autoregressive Integrated Moving Average - ARIMA) y tres modelos lineales diferentes (LR) [16]. Los datos meteorológicos empleados como entrada a los modelos corresponden a 150 estaciones meteorológicas ubicadas en la región de

estudio, abarcando un periodo de 20 años. Lograron predecir la temperatura hasta 12 horas después de la puesta del sol. En cuanto a los modelos, el mejor resultado se obtuvo con el modelo ARIMA, con el valor óptimo de 1 para *recall* en caso de pronóstico de intervalos de confianza del 95%. A pesar de este resultado alentador, la tasa de falsos positivos tuvo una sensibilidad del 21%, una tasa demasiado alta que significa riesgo de frecuentes falsas alarmas. A raíz de esto concluyeron que es deseable realizar más investigaciones.

Los autores de [7], aplicaron el enfoque de aprendizaje automático para detectar eventos de heladas a través de una red neuronal convolucional (CNN). Usaron el algoritmo Random Forest. Los datos de entrada para el entrenamiento del modelo los obtuvieron de una red de sensores inalámbricos implementados en diferentes campos en Estados Unidos. Fue necesario un intenso preprocesamiento en los datos iniciales debido a que tenían mucho ruido, el cual debía eliminarse para que los resultados sean más eficientes. Recopilaron los datos del suelo y la temperatura desde el año 2.014 hasta 2.019. Predijeron eventos de heladas para el período comprendido desde febrero de 2.018 a enero de 2.019 y obtuvieron una precisión del 98,86% en el pronóstico.

Otro trabajo que destaca la eficiencia de ARIMA para pronosticar temperaturas diarias promedio, mínimas y máximas para predicciones a más largo plazo, como mensuales y anuales es [17]. Aunque, para predicciones a corto plazo, como es por hora y por día, usaron los algoritmos de aprendizaje automático basados en modelos de redes neuronales, como multicapa perceptrón (MLP), memoria a largo-corto plazo (LSTM) y red neuronal de convolución (CNN). El estudio analizó los datos meteorológicos de tres regiones de Corea del Sur con diferentes características climáticas, para el período de 2.009 a 2.018. La unidad de tiempo considerada fue por hora y las variables de entrada para los modelos fueron 15 (temperatura, cantidad de precipitación de agua, humedad, presión de vapor, temperatura del punto de rocío, presión atmosférica, presión al nivel del mar, horas de luz solar, cantidad de radiación solar, cantidad de nieve, nubosidad total, nubosidad de nivel medio y bajo, temperatura de la superficie del suelo, velocidad del viento y dirección del viento). Concluyeron que en la mayoría de los casos los datos de entrada por hora funcionaron mejor que los datos de entrada diarios. Además, en los

resultados experimentales, dependiendo de la región de destino, el modelo CNN mostró el mejor rendimiento.

La investigación desarrollada en el Departamento de Ciencias de la Computación de la Universidad Edge Hill, Inglaterra [18] compara el rendimiento de modelos de memoria a corto plazo (LSTM) y redes convolucionales temporales (TCN) para el pronóstico del clima con enfoques clásicos (Regresión estándar, ARIMA, Random Forest, entre otros). Del Sistema de Pronóstico Global (GFS) extrajeron un total de 12 parámetros meteorológicos del período comprendido entre enero de 2.018 y mayo del mismo año. Esto se utiliza como conjunto de datos de entrenamiento. Los datos de junio de 2.018 fueron utilizados para la prueba. Esto es para probar diferentes modelos profundos entrenados para identificar el mejor modelo para la previsión. Aplicaron ventana deslizante de 7 días en cada conjunto de datos como entrada y con horizonte de 3 horas. La evaluación fue a través del MSE. Concluyendo que LSTM y el TCN producen un alto rendimiento y con errores más pequeños en comparación con los enfoques clásicos de aprendizaje automático y los enfoques de pronóstico estadístico.

Aunque el cultivo se encuentre en invernadero, al sector agrícola le interesa el pronóstico de fenómenos climáticos para usar de forma eficiente los sistemas de calefacción, ventilación y acondicionamiento agroclimático en el interior del invernadero, con el fin de obtener un mejor rendimiento agrícola. Castañeda-Miranda y Castaño [19] analizaron distintos modelos matemáticos capaces de pronosticar la temperatura interior de un invernadero, tomando como variables de entrada la temperatura del aire exterior, humedad relativa del aire exterior e interior, velocidad del viento y radiación solar global. Desarrollaron dos modelos, uno autorregresivo con entrada externa (ARX) y otro con una red neuronal artificial (ANN) perceptrón multicapa entrenada con el algoritmo *backpropagation* Levenberg-Marquardt. El análisis de precisión de los modelos se hizo mediante el coeficiente de determinación (R^2), el error estándar porcentual de la predicción (% SEP) y el error porcentual absoluto medio (MAPE). En este trabajo concluyeron que los modelos basados en ANN producían mejores resultados en la predicción de la temperatura interior del invernadero, con un nivel de confianza de 95%.

Los autores de [20] desarrollaron modelos de predicción de temperatura mínima para horizontes de 6, 12, 24, 30, 36, y 48 horas, utilizando algoritmos Random Forest, red neuronal profunda (DNN) totalmente conectada y red neuronal convolucional de memoria a corto plazo (CNN). Los datos para el entrenamiento corresponden a un periodo de diez años, obtenidos de la estación meteorológica del Servicio de Conservación de Recursos Naturales en la región de Alcalde, Nuevo México. Para el test usaron datos recopilados de dispositivos IoT en las cercanías de Freshies, Nuevo México; durante dos años, con una frecuencia de diez minutos. Las variables de entrada fueron el valor máximo, mínimo y medio de las variables temperatura, radiación, humedad relativa, temperatura del suelo, precipitación, velocidad y dirección del viento, a través de ventanas móviles de 24 horas. En el entrenamiento realizaron validación cruzada de k-fold con 3 fold sobre el conjunto de datos de entrenamiento. El modelo DNN más exitoso contiene 13 capas densas de 39 nodos cada una, con una capa de salida adicional de 1 nodo. Con función de activación *elu*, el optimizador *nadam* y *log-cosh* la función de pérdida. El modelo CNN contiene dos capas de convoluciones unidimensionales, una capa LSTM (8 nodos) y 2 capas profundas adicionales completamente conectadas (10 nodos) con *drouput* de 0,2. Con *relu* como función de activación, *Adam* optimizador y MSE función de pérdida. En el estudio observaron que los valores de RMSE y R^2 se deterioran luego de las 100 épocas de entrenamiento. Además, la velocidad del viento, la dirección del viento, la presión relativa y la presión de vapor tenían relativamente poca importancia. En los modelos analizados obtuvieron pronósticos precisos que mejoran con la longitud del registro de datos. El tiempo de anticipación de 6 horas proporcionó la mejor predicción de temperatura, tanto para Random Forest como para las redes neuronales.

Los autores de [10] desarrollaron un índice llamado IG (del portugués Índice de Geada) que permite la predicción de la ocurrencia de heladas en Brasil y países adyacentes (Argentina, Paraguay, Uruguay y Bolivia) con una anticipación de hasta 120 horas. Utilizaron los datos obtenidos de distintas estaciones meteorológicas, desde 2.012 a 2.018, tomando sólo los meses más fríos (de mayo a septiembre). Las variables que se consideraron en el proceso de desarrollo del modelo fueron la temperatura, la humedad relativa, la velocidad del viento, la presión atmosférica y la nubosidad. El índice

toma un enfoque multivariado y se basó principalmente en el análisis de la media y desviación estándar de las variables. Para el análisis de los resultados obtenidos usaron el error cuadrático medio y el R^2. Con este trabajo los autores concluyeron que la temperatura es la variable que mostró tener la mayor influencia en los resultados obtenidos por el IG. Por otro lado, el índice requería un ajuste particular para los estados de Minas Gerais, Río de Janeiro y São Paulo.

Los autores de [3] implementaron un modelo de predicción de heladas que utiliza información agroclimatológica de la localidad de Panguilemo, provincia de Talca, Chile. Las variables que consideraron fueron la temperatura, humedad relativa, radiación solar, punto de rocío, velocidad y dirección del viento; obtenidas de una estación meteorológica automática. Observaron que los eventos de helada ocurren principalmente entre los meses de mayo y agosto, siendo julio el mes con mayor número de heladas. Evaluaron diversos métodos de clasificación como Naive Bayes, Random-Forest, Arboles de Decisión y SVM. El conjunto de datos empleado correspondía a los registros de heladas y no heladas ocurridas en período de 2.010 a 2.016, el cual se encontraba desbalanceado (la cantidad de casos de heladas era notoriamente menor a los casos de no heladas), ocasionando problemas de rendimiento en los modelos. Este inconveniente fue resuelto aplicando la técnica SMOTE (Técnica de sobremuestreo de minorías sintéticas) para generar datos artificiales. Evaluaron la performance de los modelos utilizando validación cruzada, específicamente el método de LOOCV (Leave One-Out Cross Validation). La conclusión de este trabajo manifiesta que los mejores resultados se obtuvieron con el método Random Forest, con un rendimiento global del 90% y un 15% de error en heladas no detectadas. Con la capacidad de predecir con una antelación de 12 horas, la ocurrencia de un evento de helada en una zona específica cercana a una estación meteorológica.

Otro caso en Chile, se da por Fuentes et all. [9], con el objeto de predecir la temperatura mínima a partir de la información histórica recopilada en estaciones meteorológicas desarrollaron una red neuronal artificial (ANN) de tres capas usando el algoritmo backpropagation. La información meteorológica que utilizaron fue provista por la Red Nacional de Agrometeorología (AGROMET) de Chile generada por diez estaciones meteorológicas automáticas ubicadas en los valles interiores de las Regiones

del Maule y Biobío en Chile. El conjunto de datos usado corresponde al período de 2.010 a 2.017. Las variables meteorológicas consideradas fueron la temperatura del aire, la humedad relativa, la magnitud y la dirección del viento, la precipitación, la radiación solar y la temperatura del punto de rocío. En el proceso de validación se utilizaron distintos índices estadísticos para evaluar las predicciones de heladas, incluida la sensibilidad, especificidad, precisión, exactitud, tasa de error y F1- score. Los autores concluyeron que la ANN con 25 neuronas en la capa oculta fue el modelo que proporcionó los mejores resultados. Respecto de las variables meteorológicas usadas destacan la importancia de los datos del viento para la región analizada porque están asociados con la cordillera de Los Andes.

En Argentina, Diedrichs et all. en [8] abordan la predicción de heladas mediante un modelo desarrollado con redes Bayesianas y Random Forest. Los datos para el desarrollo del modelo fueron obtenidos de cinco estaciones agrometeorológicas de la provincia de Mendoza, Argentina. Las variables consideradas fueron la temperatura y la humedad relativa desde 2.001 hasta 2.016. Los datos reales disponibles fueron insuficientes para construir un sistema de pronóstico de heladas preciso, debido a la pequeña cantidad de heladas que se producen durante el año, esto motivó que, al igual que en el trabajo de Möller-Acuña et all. [3], utilizaron la técnica SMOTE en la generación de datos artificiales. Para analizar los resultados de los modelos en la predicción de la temperatura eligieron RMSE y R^2 como métricas de regresión. Los autores concluyeron que los modelos implementados con Redes Neuronales eran competitivos en términos de sensibilidad y precisión. Además, que la inclusión de información de sensores vecinos ayuda a mejorar la precisión del modelo de pronóstico.

En resumen, para el análisis y pronóstico de la temperatura, algunos estudios han usado métodos numéricos [3] [8], pero es mayor la cantidad de casos que en los últimos años aplicaron algoritmos de aprendizaje automático como redes neuronales de distinto tipo [5] [7] [16] [17] [18], redes Bayesianas y Random Forest [2][5][6] por ser superiores en rendimiento.

En la mayoría de los casos, los datos son obtenidos de estaciones meteorológicas, siendo de relevancia la temperatura, la humedad relativa, la velocidad y dirección del viento, la presión atmosférica y la nubosidad, y en algunos casos el punto de rocío como

variables de entrada para los modelos. Ante el inconveniente de que la cantidad de datos sea insuficiente se ha utilizado técnicas de sobremuestreo como SMOTE [3] [8].

El horizonte con el que se ha logrado predecir heladas varía desde 12 horas [3] [16] hasta 120 horas [10] de antelación. En el análisis de los resultados obtenidos se ha usado validación cruzada (método de LOOCV) [3], el error cuadrático medio [9] [10] [8] y el coeficiente de determinación R^2 [8] [10] [19].

Así, el problema del pronóstico de la temperatura ha sido y sigue siendo objeto de análisis en diversas regiones del mundo.

Capítulo 3

Marco Teórico

3. Marco Teórico

El desarrollo del marco teórico presenta en la primera sección conceptos referidos a Ciencia de datos, como el *pipeline*, redes neuronales y evaluación de modelos. La segunda sección refiere a series temporales y la tercera trata el fenómeno meteorológico de la helada.

3.1. Ciencia de Datos

Comúnmente se define Ciencia de Datos como una metodología [21] mediante la cual se puede extraer conocimiento a partir de los datos, con el fin de utilizar este conocimiento para predecir eventos futuros, comprender el pasado/presente, crear nuevos productos, entre otros usos. En síntesis, la ciencia de datos tiene un objetivo ambicioso, pretende generar opiniones basadas en los datos para que sean usadas en la toma de decisiones [22].

3.1.1. Proceso de ciencia de datos

El proceso de la ciencia de datos (Figura 1) es un proceso sistemático y disciplinado que involucra un conjunto de fases. En primer lugar, se debe entender el problema y plantear el objetivo de análisis de los datos. A continuación, comienza la etapa de exploración de los datos, que consiste en comprender los datos y las cosas del mundo real que estos datos describen, para poder extraer características significativas. Estas características se usan en herramientas de modelado y análisis, finalmente se presentan los resultados obtenidos [23]. En este punto, hay un ciclo de retroalimentación, el nuevo conocimiento disponible permite volver a la primera fase, a generar nuevas preguntas, nuevos problemas que se deben enmarcar y un nuevo proceso comienza.

3.1.1.1. Entender el problema

El primer paso es comprender el negocio de forma que permita enmarcar el problema y elaborar un objetivo de análisis bien definido a partir de él. Los problemas suelen ser muy confusos al principio. Se necesita mucho tiempo y esfuerzo para seguir un curso de acción hasta el resultado final. Entonces, antes de invertir ese esfuerzo, es fundamental garantizar que se está trabajando en el problema correcto y aclarar exactamente qué constituye una posible solución al problema [23].

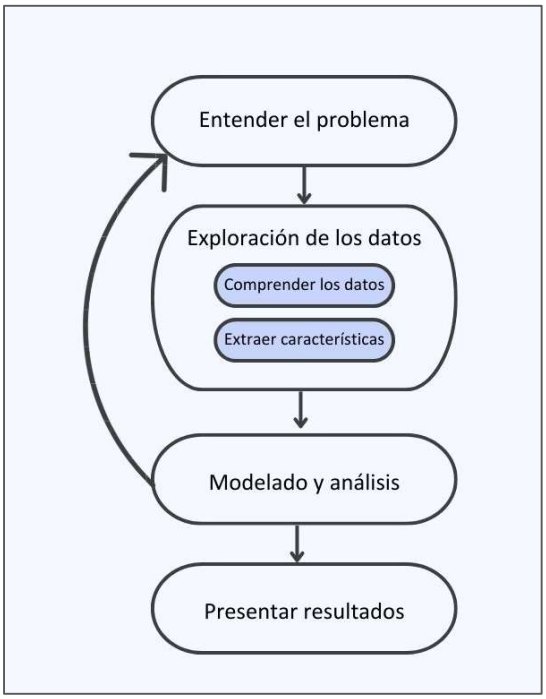


Figura 1 - Proceso de Ciencia de datos Fuente: C. Field – The Data Science Handbook [23]

3.1.1.2. Exploración de los datos

Al momento de obtener los datos pueden presentarse distintos inconvenientes, uno de ellos es la dificultad para disponer de ellos. Esto debido a que el propietario de los datos puede tener impedimentos legales para entregarlos o simplemente negarse a entregarlos para su procesamiento, por razones de negocio, patentes, etc. Otro inconveniente es la cantidad de datos disponibles, en caso de ser una cantidad escasa, no se obtendrán buenos resultados. En el caso contrario, una gran cantidad de datos demandará un tiempo considerable de exploración, con alta probabilidad de obtener resultados de utilidad.

Una vez obtenidos los datos, es importante determinar si estos datos pueden resolver el problema que se intenta abordar [23]. En caso afirmativo, se realizan tareas para preparar los datos en un formato apropiado. Los datos se recopilan de diferentes fuentes, en diversas formas presentando casos de datos faltantes, incompletos o incorrectos; en escalas diferentes lo que dificulta operar sobre ellos.

El científico de datos, comienza por conocer el tipo de los datos, los clasifica, los agrupa, encuentra valores máximos, mínimos y medios, cuartiles. Los visualiza de diferentes maneras [23], estudia la distribución de los datos observando en gráficos las

distribuciones a las que tienden. Los limpia, los filtra y extrae las características significativas para el problema, generando un conjunto de datos apropiados para el procesamiento [24].

La extracción de características es una parte creativa de la ciencia de datos que está estrechamente relacionada con la experiencia en el dominio. Una característica buena corresponderá a algún fenómeno del mundo real. Los científicos de datos deben trabajar en estrecha colaboración con expertos en el dominio, y comprender qué significan estos fenómenos y cómo descomponerlos en números. Extraer buenas características es una tarea relevante, marcará el éxito o fracaso del análisis [23]. Además de beneficiar el procesamiento de los datos con la reducción de la dimensión de los datos, dado que una alta dimensión de los datos puede ser un obstáculo a la hora del procesamiento de los mismos.

Finalmente, a los datos se les da la estructura adecuada para el procesamiento. La cual dependerá del objetivo planteado, si se trata de obtener una descripción, un agrupamiento, una predicción entre otros.

Para identificar características relevantes se utilizan distintas estrategias, una de ellas es la matriz de correlación, se trata de una estrategia de visualización [24] [25].

En términos más generales, la correlación es una métrica que mide qué tan estrechamente vinculadas están dos variables X e Y [23]. Los valores de la correlación varían entre -1 y 1, siendo ambos extremos una buena aproximación. El valor -1 indica una correlación inversa, mientras una variable crece la otra decrece en similar proporción. En cambio, para el valor 1 ambas variables crecen o decrecen en la misma proporción.

3.1.1.3. Modelado y análisis

La mayoría de los proyectos de ciencia de datos involucran algún tipo de modelo de aprendizaje automático, como puede ser un clasificador, un modelo de regresión o un algoritmo de agrupación [23].

Con el aprendizaje automático (machine learning), se ingresan datos de casos relevantes y se obtienen reglas que luego, se pueden aplicar a nuevos datos para producir respuestas originales [26]. Este tipo de aprendizaje se ha convertido en un

término general que cubre diferentes áreas. Existen dos tipos principales de aprendizaje automático, no supervisado y supervisado.

3.1.1.3.1. Aprendizaje no supervisado

El aprendizaje no supervisado incluye algoritmos que aprenden de un conjunto de casos de entrenamiento no etiquetados. Estos algoritmos se utilizan para encontrar patrones, resumir y explicar características o estructuras clave de los datos [21][23]. La mayoría de las técnicas se pueden resumir en los siguientes grupos de problemas: 1) Clustering para dividir el conjunto de casos en grupos; 2) Reducción de dimensionalidad de los datos; 3) Detección de valores atípicos para encontrar eventos inusuales; 4) Detección de novedades, se ocupa de los casos en los que se producen cambios en los datos [21].

3.1.1.3.2. Aprendizaje supervisado

El aprendizaje supervisado encuentra asociaciones entre las características de un conjunto de datos y una variable objetivo [22]. Los algoritmos aprenden de un conjunto de casos de casos etiquetados con los que se realiza el entrenamiento, para luego generalizar al conjunto de todas las entradas posibles [21].

La pregunta planteada al iniciar un proceso de ciencia de datos determina el tipo de respuesta que busca el científico de datos. Según el tipo de respuesta se determina el conjunto de técnicas a emplear. Si la pregunta admite solo un conjunto discreto de respuestas (un número finito de opciones), se enfrenta a un problema de clasificación. En cambio, si la pregunta es una predicción de un valor real, se trata de un problema de regresión [21].

La clasificación es una herramienta de aprendizaje automático supervisado para la predicción de resultados discretos (en la variable dependiente u objetivo), aun cuando los datos o variables independientes puedan ser continuas o discretas. De acuerdo con la cardinalidad del conjunto objetivo, se distingue entre clasificadores binarios o clasificadores multiclase [23].

En los problemas de regresión, el valor a predecir se puede expresar como una combinación de una o más variables independientes (también llamadas predictores). El

objetivo de la regresión es construir un modelo para predecir la respuesta a un conjunto de variables de entrada [21].

Dentro de estas técnicas de aprendizaje supervisado se encuentra la regresión logística, las máquinas de vectores soporte, los árboles de decisión, random forest, las redes neuronales, entre otras [21]. En particular, las redes neuronales básicas son herramientas estándar en el aprendizaje automático por ser fáciles de usar y bastante efectivas [23].

Aprendizaje profundo

El aprendizaje profundo (deep learning) es un área específica del aprendizaje automático, que pone énfasis en aprender a través de sucesivas representaciones en capas, cada vez más significativas. La cantidad de capas que contribuyen a un modelo de datos se denomina profundidad del modelo. Estas representaciones en capas corresponden a modelos de redes neuronales [26], y las variantes sofisticadas de estas redes se utilizan para el aprendizaje profundo [23].

3.1.1.4. Presentar resultados

Esta última fase consiste en elaborar un informe que describa el trabajo que realizado y cuáles fueron sus resultados [23].

3.1.2. Redes neuronales

Las redes neuronales artificiales (ANN) son técnicas populares de aprendizaje automático que simulan el mecanismo de aprendizaje en organismos biológicos. El sistema nervioso humano contiene neuronas conectadas entre sí mediante el uso de axones y dendritas. Las regiones de conexión entre axones y dendritas se denominan sinapsis. La fuerza de las conexiones sinápticas cambia en respuesta a estímulos externos, así se lleva a cabo el aprendizaje en los organismos vivos.

Este mecanismo biológico se simula en redes neuronales artificiales, que contienen unidades de cómputo denominadas neuronas (Figura 2). Las unidades computacionales están conectadas entre sí a través de pesos, que cumplen la misma función que la fuerza de las conexiones sinápticas en los organismos biológicos [27].

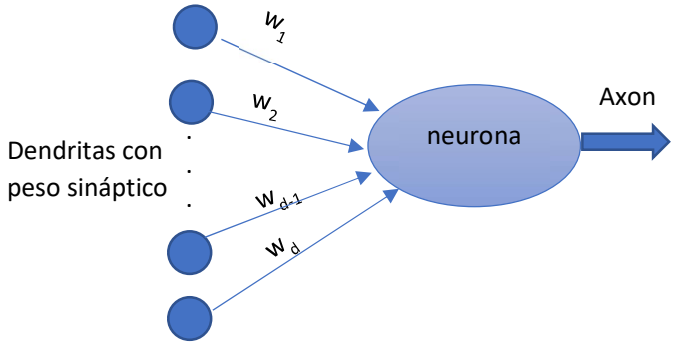


Figura 2 - Red neuronal artificial Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook[27]

Las redes neuronales son una herramienta no estándar de análisis estadístico, por medio de la cual es posible estudiar y modelar cualquier conjunto de datos [19]. Una propiedad importante de estas redes es su capacidad de aprender de diferentes procesos que ocurren. Donde el aprendizaje comúnmente se define como un proceso que optimiza el rendimiento de la red con respecto a una tarea determinada a través del entrenamiento.

Los modelos con redes neuronales artificiales (ANN) se han vuelto cada vez más importantes en el procesamiento y análisis de series de tiempo y cálculos de predicciones futuras. Pueden aprender y adaptarse a cualquier modelo, adquieren conocimiento al detectar los patrones y las relaciones en los datos y los almacenan en conexiones interneuronales artificiales conocidas como pesos sinápticos [9]. Se han aplicado con éxito a los problemas de predicción y clasificación de patrones [19].

De forma particular se han utilizado ampliamente para investigar el mecanismo del cambio climático y predecir la tendencia del cambio climático, por poseer la ventaja de que las redes neuronales artificiales hacen uso completo de cierta información desconocida oculta en los datos [28].

3.1.2.1. Funcionamiento básico de una red neuronal

Toda red neuronal artificial posee al menos una capa de entrada y una capa de salida de neuronas. Las entradas se proyectan sobre la capa de salida mediante el uso de una función lineal [28]. Cada entrada a una neurona se escala con un peso, lo que afecta la función calculada en esa unidad, propagando los valores calculados hacia las

neuronas de salida y utilizando los pesos como parámetros intermedios. El aprendizaje ocurre cambiando los pesos que conectan las neuronas [27].

El estímulo externo que necesita una red neuronal artificial para el aprendizaje lo proporcionan los datos de entrenamiento, que contienen casos de pares de entrada-salida de la función a aprender. Estos pares de datos de entrenamiento se introducen en la red neuronal mediante el uso de representaciones de entrada para hacer predicciones sobre la salida. Los datos de entrenamiento brindan retroalimentación sobre la corrección de los pesos en la red neuronal según la coincidencia de la salida predicha para una entrada en particular con la etiqueta de salida registrada en los datos de entrenamiento [27].

Un método típico es dividir el conjunto de datos en tres conjuntos disjuntos, cada uno con diferentes proporciones. Estos conjuntos son el conjunto de entrenamiento que permite al modelo aprender sobre la dinámica y el comportamiento del problema que está tratando, el conjunto de validación con el cual se obtienen resultados que se comparan con la información existente y el conjunto de pruebas con el que se testea el modelo con nuevos datos [9].

Los pesos entre neuronas se ajustan de una manera matemáticamente justificada en respuesta a los errores de predicción. El objetivo de cambiar los pesos es modificar la función calculada para que las predicciones sean más correctas en iteraciones futuras. Esta capacidad de calcular las salidas con precisión a partir de entradas no vistas, entrenando sobre un conjunto finito de datos, se conoce como generalización del modelo. La principal utilidad de todos los modelos de aprendizaje automático se obtiene de su capacidad para generalizar su aprendizaje a partir de datos de entrenamiento vistos a datos no vistos [27].

3.1.2.2. Perceptrón

El perceptrón fue la primera red neuronal artificial descrita algorítmicamente. Es un tipo de red neuronal que puede decidir si una entrada pertenece a una clase específica. Los perceptrones se pueden clasificar en perceptrón de Rosenblatt y perceptrón multicapa (MPL) [28].

El perceptrón de Rosenblatt es una red neuronal simple, de una sola capa con *m* nodos de entrada y una única neurona de salida. Consta de una sola capa con peso y sesgo sinápticos ajustables [28]. La red perceptrón multicapa consta de una capa de entrada de *m* nodos de origen, *i* capas ocultas de neuronas y una capa de salida de *n* neuronas. Cada neurona de la red tiene una función de activación no lineal diferenciable.

El perceptrón es el tipo más simple de red neuronal, su arquitectura básica contiene una única capa de entrada y un nodo de salida (Figura 3). Cada instancia de entrenamiento tiene la forma $(\bar{X}, y)$, donde cada $\bar{X} = [x_1,...x_d]$ contiene *d* variables de característica, e *y* es el valor observado de la variable de clase binaria [27]. Además, desde la capa de entrada se transmiten las *d* características con los pesos $\bar{W} = [w_1,...w_d]$.

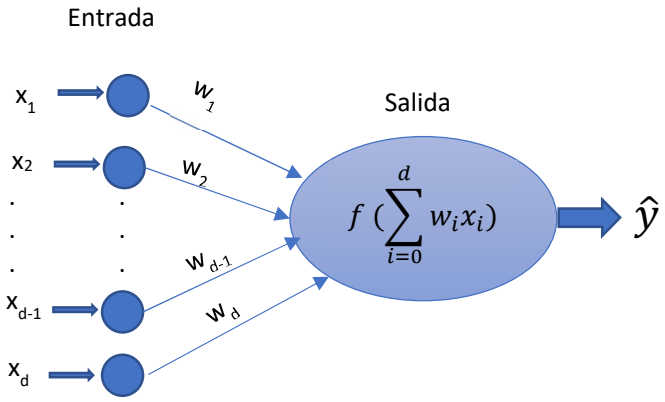


Figura 3 – Perceptrón Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook[27]

En el nodo de salida se calcula la función lineal $\bar{W} \cdot \bar{X}$. El valor de esta función es transformado a +1 o −1 (clasificación binaria) a través de la función *sign*, para predecir a $\hat{y}$ (Ecuación 1), la variable dependiente de $\bar{X}$.

$$\hat{y} = sign\{\bar{W} \cdot \bar{X}\} = sign\{\sum_{i=0}^{d} w_i x_i\} \quad (1)$$

La función *sign* cumple el papel de una función de activación. Se pueden usar diferentes funciones de activación para simular diferentes tipos de modelos utilizados en el aprendizaje automático.

El error de la predicción es por lo tanto la diferencia entre el valor observado y el valor predicho (Ecuación 2). En los casos en que el valor de error $E(\bar{X})$ no sea cero, los

pesos en la red neuronal deben actualizarse en la dirección (negativa) del gradiente de error.

$$E(\bar{X}) = y - \hat{y} \tag{2}$$

En muchos casos, hay una parte invariable de la predicción, que se denomina sesgo. Esto tiende a ocurrir en situaciones en las que la distribución de la clase binaria está muy desequilibrada. En tal caso, es necesario incorporar una variable de sesgo adicional *b* (Ecuación 3) que capture esta parte invariable de la predicción [27].

$$\hat{y} = sign\{\bar{W} \cdot \bar{X} + b\} = sign\{\sum_{i=0}^{d} w_i x_i + b\} \tag{3}$$

El sesgo se puede incorporar como un peso mediante una neurona (Figura 4).

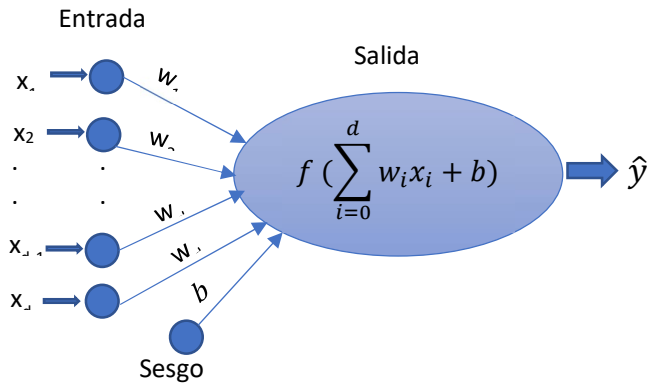


Figura 4 - Perceptrón con sesgo Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook[27]

Cuando Rosenblatt propuso el algoritmo del perceptrón, estas optimizaciones se realizaron de forma heurística, y no se presentó en términos formales de optimización en el aprendizaje automático. El objetivo siempre fue minimizar el error en la predicción, por lo tanto, el algoritmo del perceptrón se diseñó heurísticamente para minimizar el número de errores de clasificación, y se disponía de pruebas de convergencia que garantizaban la corrección del algoritmo de aprendizaje en entornos simplificados. Este tipo de función objetivo de minimización se denomina función de pérdida [27].

El algoritmo de entrenamiento de las redes neuronales se alimenta de cada instancia de los datos de entrada $\bar{X}$ (uno por uno o en lotes pequeños) para calcular la predicción $\hat{y}$. Luego, los pesos se actualizan en función del valor de error (Ecuación 2). Específicamente, cuando el vector de datos $\bar{X}$ se introduce en la red, el vector de peso $\bar{W}$ se debe actualizar (Ecuación 4).

$$\bar{W} \Leftarrow \bar{W} + \propto (y - \hat{y})\bar{X} \quad (4)$$

El parámetro α regula la tasa de aprendizaje de la red neuronal. El algoritmo del perceptrón recorre repetidamente todos los casos de entrenamiento en orden aleatorio y ajusta iterativamente los pesos para minimizar el error cuadrático de la predicción hasta que alcanza la convergencia. Cada uno de estos ciclos se denomina época.

El algoritmo del perceptrón siempre converge para proporcionar un error cero en los datos de entrenamiento cuando los datos son linealmente separables. Sin embargo, no se garantiza que el algoritmo del perceptrón converja en instancias donde los datos no son separables linealmente, en estos casos se requiere del uso de arquitecturas neuronales más complejas [27].

3.1.2.3. Red neuronal multicapa

Una red neuronal multicapa presenta una arquitectura que comprende tres tipos de capas neuronales. La primera o capa más baja, es una capa de entrada donde se recibe información externa. La última o capa más alta, es una capa de salida donde se obtiene la solución del problema. La capa de entrada y la capa de salida están separadas entre sí por una o más capas intermedias llamadas capas ocultas [19]. Las neuronas de sesgo pueden estar presentes tanto en las capas ocultas como en las capas de salida (Figura 5 b).

La arquitectura específica de las redes neuronales multicapa se conoce como redes de avance (feed-forward), porque las capas sucesivas se alimentan entre sí en la dirección de avance desde la entrada hasta la salida [27]. Los nodos en capas adyacentes generalmente están completamente conectados por arcos acíclicos desde una capa inferior a una capa superior [19], propagando hacia adelante las señales de la función y hacia atrás las señales de error que se identifican [28].

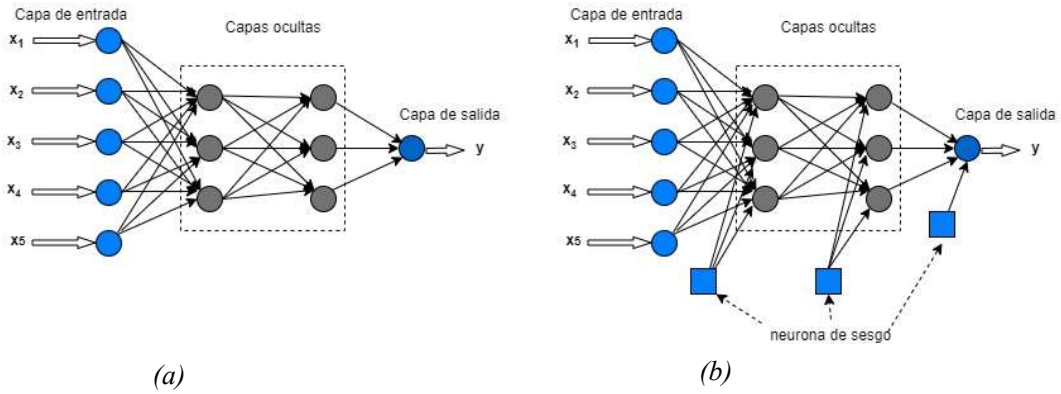


Figura 5 - Arquitectura básica de una red feed-forward con dos capas ocultas y una capa de salida. (a) Arquitectura sin sesgo y (b) Arquitectura con sesgo. Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook[27]

El comportamiento de una red neuronal está determinado por las funciones de transferencia de sus neuronas, por la función de pérdida que se optimiza en la capa de salida, por la tasa de aprendizaje y por la arquitectura misma [19]. La definición de la arquitectura de la red neuronal incluye la determinación del número de capas y el número de neuronas en cada capa (dimensionalidad) [27].

Si una red neuronal contiene $p_1 \ldots p_k$ unidades en cada una de sus k capas, entonces las representaciones vectoriales (columna) de estas salidas, indicadas por $\overline{h_1} \ldots \overline{h_k}$ tienen dimensiones $p_1 \ldots p_k$. Cada caso de entrenamiento contiene d variables de característica, los pesos de las conexiones entre la capa de entrada y la primera capa oculta están contenidos en una matriz W_1 con tamaño $d \times p_1$. De forma general, los pesos entre la r-ésima capa oculta y la $(r+1)$-ésima capa oculta están dados por la matriz $p_r \times p_{r+1}$ denotada por W_r. Si la capa de salida contiene o nodos, entonces la matriz final W_{k+1} tiene un tamaño de $p_k \times o$ (Figura 6).

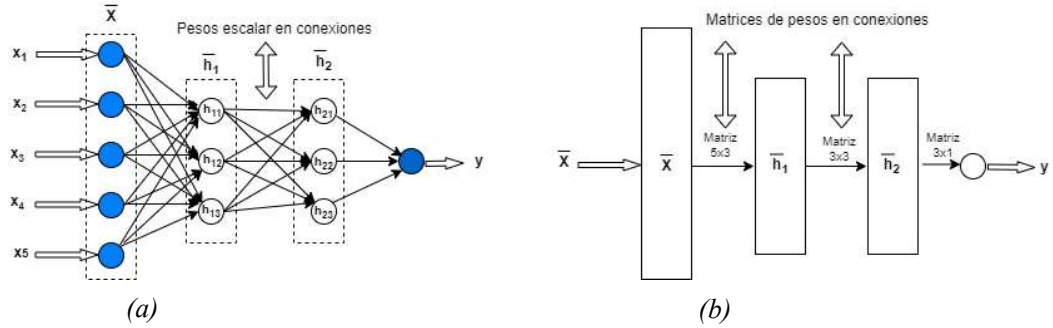


Figura 6- Arquitectura de red neurona con distinta notación. (a) Con notación escalar. (b) Con notación vectorial. Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook [27]

Aunque existen numerosos tipos de ANN, el utilizado comúnmente es el Perceptrón Multicapa (MLP) con *backpropagation* [9][19]. Se ha aplicado ampliamente en diferentes áreas, incluyendo reconocimiento de patrones, procesamiento de imágenes, control inteligente, predicción de series de tiempo, entre otras.

3.1.2.4. Backpropagation

En una red neuronal de una sola capa, el proceso de entrenamiento es relativamente sencillo porque el error (o función de pérdida) se calcula como una función directa de los pesos, lo que permite obtener fácilmente el gradiente. En el caso de redes multicapa, la pérdida es una función de composición de los pesos en capas anteriores. El gradiente de una función de composición se calcula usando el algoritmo de *backpropagation* [27].

Este proceso, en forma simplificada, consiste en un ciclo de dos fases de propagación hacia adelante (forward) y adaptación hacia atrás (backward). Se requiere la fase hacia adelante para calcular los valores de salida y las derivadas locales en varios nodos, y la fase hacia atrás para acumular los productos de estos valores locales en todas las rutas desde el nodo hasta la salida.

En la fase *forward*, las entradas para una instancia de entrenamiento alimentan a la red neuronal. Esto produce una cascada de cómputos hacia adelante a través de las capas, utilizando el conjunto actual de pesos. La salida prevista final se compara con la de la instancia de entrenamiento y se calcula la derivada de la función de pérdida con respecto a la salida. Luego, es necesario calcular la derivada de esta pérdida con respecto a los pesos en todas las capas en la fase inversa [27].

El objetivo de la fase hacia atrás es aprender el gradiente de la función de pérdida con respecto a los diferentes pesos, utilizando la regla de la cadena del cálculo diferencial. Esta información del error se transmite hacia atrás, comenzando desde la capa de salida hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida, asignando un porcentaje aproximado del error a la participación de cada neurona en la salida original. Este proceso se repite hacia atrás, capa por capa, hasta que todas las neuronas de la red hayan recibido un error que describa su contribución en relación con el error total. Según esta información recibida, todos los pesos se

ajustan, de modo que la próxima vez, la diferencia entre la salida calculada y la observada disminuya [9].

3.1.2.5. Redes neuronales recurrentes (RNN)

Existen ciertos tipos de datos que contienen dependencias secuenciales entre los atributos, como el caso de las series temporales, el texto o datos biológicos [27]. Una red neuronal recurrente (RNN), es un caso especial de red neuronal donde el objetivo es hacer uso de tipos de datos secuenciales, y aprender de las etapas anteriores para pronosticar tendencias futuras [29].

Estas redes poseen un tipo primitivo de memoria, en forma de capas recurrentes que toman dos tipos de entrada: 1) la salida de la capa anterior y 2) la salida de la misma capa recurrente desde el último punto que procesó. Es decir, la salida de una capa recurrente al clasificar un punto, se pasará de nuevo a esta misma capa al clasificar el siguiente punto. Así, la salida podría codificar una predicción sobre cuál será el siguiente punto, dado el punto actual [23] .

En una RNN, de forma general, el vector de entrada en el tiempo *t* es $\bar{x}_t$, el estado oculto en el tiempo *t* es $\bar{h}_t$, y el vector de salida en el tiempo *t* (por ejemplo, predicción para *t + 1)* es $\bar{y}_t$. Entonces, el estado oculto en el tiempo *t* viene dado por una función del vector de entrada en el tiempo *t* y el vector oculto en el tiempo *(t − 1)* (Ecuación 5) [27].

$$\bar{h_t} = f(\bar{h}_{t-1}, \bar{x}_t) \tag{5}$$

El estado oculto evoluciona con el tiempo, pero los pesos y la función subyacente *f(·, ·)* permanecen fijos en todas las marcas de tiempo (elementos secuenciales). El valor de salida será dado por la función (Ecuación 6).

$$\bar{y_t} = g(\bar{h}_t) \tag{6}$$

Considerando la matriz W_{xh} de entrada oculta, la matriz capa oculta W_{hh} y una matriz de salida oculta W_{hy}, la (Ecuación 5) de forma extendida se obtiene (Ecuación 7).

$$\bar{h_t} = f(W_{xh}\bar{x}_t + W_{hh}\bar{h}_{t-1}) \tag{7}$$

$$\bar{y_t} = W_{hy}\bar{h}_t$$

Una representación gráfica se presenta en la Figura 7.

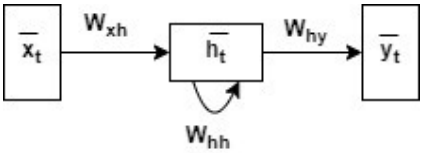


Figura 7- Red Neuronal Recurrente Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook [27]

La arquitectura multicapa se utiliza para construir modelos de mayor complejidad. En la Figura 8 se muestra un ejemplo de una red profunda que contiene tres capas. Para la marca de tiempo *t*, en la capa *k,* el vector de estados ocultos se identifica por $\bar{h}_t^{(k)}$.

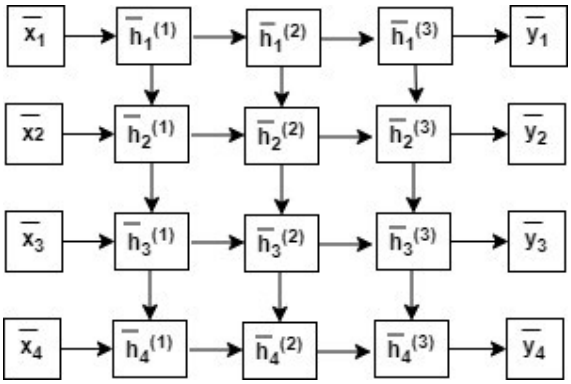


Figura 8 - RNN multicapa Fuente: C. C. Aggarwal, Neural networks and deep learning : a textbook [27]

Las redes recurrentes básicas, sólo recuerdan la última salida de la capa y la retroalimentará como entrada durante la siguiente iteración, por lo tanto, no son adecuadas para recordar secuencias de datos más largas [29]. Este problema se resuelve utilizando redes de memoria a corto y largo plazo (LSTM - Long Short-Term Memory) que pueden recordar eventos seleccionados del pasado, como el caso de las series de tiempo [23].

3.1.2.5.1. Redes de Memoria a Corto Plazo

El modelo LSTM se ha ocupado del problema de pérdida del gradiente de las redes neuronales recurrentes. Es uno de los modelos de redes neuronales artificiales más potentes para datos secuenciales como texto, series temporales y otras secuencias discretas como secuencias biológicas [27].

El modelo LSTM consta de conexiones de celdas que contienen una capa oculta y una capa de salida. La capa oculta recibe una entrada y valores de contexto de la capa

oculta de la celda anterior, y luego deriva un vector de contexto oculto para la siguiente capa a partir de las entradas, y una capa de salida calcula el vector de salida [17].

Las redes LSTM plantean una mejora a la arquitectura de la red neuronal recurrente de la Figura 8, cambiando las condiciones recurrentes de propagación de los estados ocultos $\bar{h}_t^{(k)}$. Para lograr este objetivo se incluye un vector oculto adicional denominado estado de celda, denotado como $\bar{c}_t^{(k)}$. Este vector funciona como una memoria a largo plazo, retiene parte de la información de estados anteriores mediante una combinación de operaciones parciales de olvido e incremento en los estados de la celda anteriores [27].

En las actualizaciones de los estados de celda y los estados ocultos se utilizan cuatro variables vectoriales intermedias: 1) la variable de entrada $\bar{\imath}$, 2) la variable de olvido $\bar{f}$, 3) la variable de salida $\bar{o}$ y 4) la variable de contenidos propuestos $\bar{c}$.

Los vectores $\bar{\imath}, \bar{f}\ y\ \bar{o}$ contienen un valor continuo en *(0, 1).* Se denominan puertas de entrada, olvido y salida respectivamente, porque conceptualmente se utilizan como puertas booleanas para decidir (i) si agregar a un estado de celda, (ii) si olvidar un estado de celda y (iii) si permitir la fuga a un estado oculto desde un estado de celda. La actualización del estado de celda (Ecuación 8) usa de algunas de estas variables intermedias.

$$\bar{c}_t^{(k)} = \underbrace{\bar{f} \odot \bar{c}_{t-1}^{(k)}}_{\text{Reset}} + \underbrace{\bar{\imath} \odot \bar{c}}_{\text{Incremento}} \tag{8}$$

La primera parte de (Ecuación 8) (reset) usa los bits de olvido en $\bar{f}$ para decidir qué parte del estado de celda de la marca de tiempo anterior se restablecerá a 0, de este modo olvida información del pasado. La otra parte (incremento), usa los bits de entrada en $\bar{\imath}$ para decidir si agregar los componentes de $\bar{c}$ a los estados de celda, así incorpora nueva información en la memoria a largo plazo.

Los estados ocultos $\bar{h}_t^{(k)}$ se actualizan (Ecuación 9) utilizando fugas del estado de celda, copiando una forma funcional (función tangente) de cada uno de los estados de celda en cada uno de los estados ocultos, dependiendo de la puerta de salida $\bar{o}$.

$$\bar{h}_t^{(k)} = \bar{o} \odot \tanh(\bar{c}_t^{(k)}) \quad (9)$$

Los estados de celda a largo plazo funcionan como autopistas de gradiente, que se filtran en los estados ocultos de la red, esto hace que los modelos LSTM proporcionen mejores flujos de gradiente que las RNN estándar, evitando los problemas asociados con la pérdida y explosión del gradiente.

3.1.2.6. Hiperpárametros de una red neuronal

Elegir la arquitectura de red adecuada es más un arte que una ciencia; y aunque existen algunas mejores prácticas y principios en los que se puede confiar, es la práctica lo que ayuda a encontrar un modelo adecuado empleando redes neuronales [26].

Al definir una red neuronal se establecen las características estructurales y funcionales de la misma, como son la cantidad de capas, cantidad de neuronas, función de activación de la capa, función de pérdida, optimizador, cantidad de épocas de entrenamiento, entre otras. Estas características son los parámetros de configuración o hiperparámetros de la red neuronal, y determinan el rendimiento de la misma.

El entrenamiento de una red neuronal gira en torno a las capas, los datos de entrada, los objetivos planteados, la función de pérdida y el optimizador. La red está compuesta por capas encadenadas, transformando los datos de entrada en predicciones. La función de pérdida compara estas predicciones con los objetivos, produciendo un valor de pérdida (medida de qué tan bien las predicciones coinciden con lo esperado). El optimizador utiliza este valor de pérdida para actualizar los pesos de la red (Figura 9) [26]. El objetivo a lograr con el modelo guiará la elección de la función de pérdida y de la métrica de evaluación.

3.1.2.6.1. Tipos de capas

Diferentes capas son apropiadas para diferentes tipos de procesamiento de datos. Así, los datos vectoriales almacenados en tensores de forma 2D (muestras, características), se procesan mediante capas densamente conectadas, también llamadas capas totalmente conectadas o densas. Los datos de secuencia, almacenados en tensores de forma 3D (muestras, intervalos de tiempo, características), generalmente se procesan mediante capas recurrentes, como una capa LSTM [26].

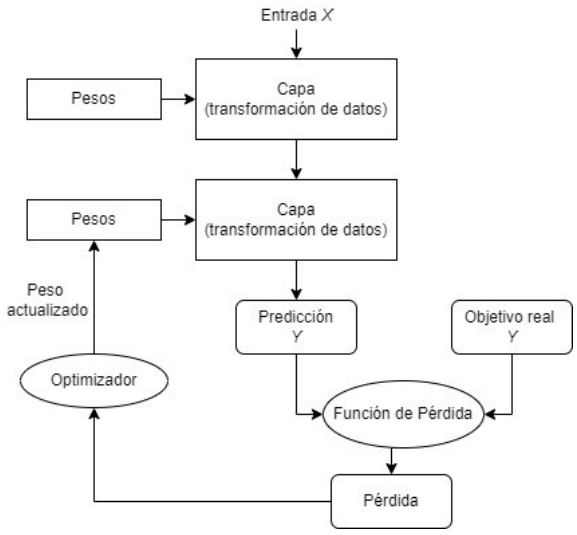


Figura 9 - Relación entre la red, capas, función de pérdida y optimizador. Fuente: F. Chollet, Deep Learning with Python [26].

3.1.2.6.2. Función de activación

Las redes neuronales reciben entradas (valores cuantitativos) y emiten otros valores cuantitativos [22]. La neurona, recibe la entrada que es combinada con los pesos, luego se somete a una función de activación y emite la señal de salida (Ecuación 10).

$$\hat{y} = f(w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m) \tag{10}$$

El conjunto de entradas es $X = \{x_0, x_1, x_2, \ldots, x_m\}$, el conjunto de pesos está dado por $W = \{w_0, w_1, w_2, \ldots, w_m\}$ y f es la función de activación.

Inicialmente, las matrices de peso toman pequeños valores aleatorios que no producen representaciones útiles, pero son un punto de partida para ajustar gradualmente estos pesos, basándose en la señal de retroalimentación [26]. Es decir que los pesos contienen la información aprendida por la red cuando procesa los datos de entrenamiento.

La función de activación define la forma en que la suma ponderada de la entrada se transforma en una salida, por lo tanto, tiene un gran impacto en la capacidad y el rendimiento de la red neuronal. La función de activación más básica es la identidad o activación lineal (Ecuación 11), que se usa normalmente en el nodo de salida, cuando el objetivo es un valor real [27].

$$\Phi_{(v)} = v \tag{11}$$

Las funciones de activación clásicas, que se utilizaron al principio del desarrollo de las redes neuronales fueron las funciones de signo (*sign*), sigmoide (*sigmoid*) y tangente hiperbólica (*tanh*). En los últimos años, se ha vuelto más popular la función de activación ReLU (*Rectified Lineal Unit*) (Ecuación 12), representada en la Figura 10. Ésta ha reemplazado en gran medida las funciones sigmoide y *tanh* debido a la facilidad para entrenar redes neuronales multicapa [27] .

$$\Phi_{(v)} = \max\{v, 0\} \tag{12}$$

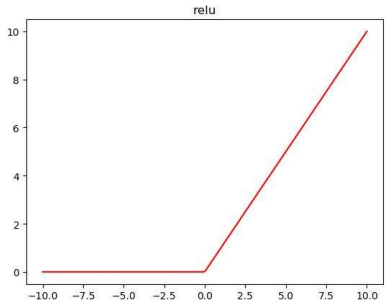


Figura 10 - Función de activación ReLU

3.1.2.6.3. Función de pérdida (loss)

Esta función mide el rendimiento de la red con los datos de entrenamiento. El algoritmo siempre tratará de minimizar el valor de esta función. Existen pautas simples a seguir para elegir la función de pérdida correcta cuando se trata de problemas comunes como la clasificación, la regresión y la predicción de secuencias.

Normalmente se utiliza la entropía cruzada binaria para un problema de clasificación de dos clases, la entropía cruzada categórica para un problema de clasificación de muchas clases, el error cuadrático medio (MSE) para un problema de regresión y la clasificación temporal conexionista para un problema de aprendizaje de secuencias [26].

3.1.2.6.4. Optimizador

El optimizador establece el mecanismo de actualización de los pesos de la red a partir de la función de pérdida. Implementa una variante específica de descenso de gradiente estocástico (SGD) [26]. Entre los principales optimizadores se encuentra Adagrad, Adadelta, RMSProp, Adam (RMSProp con momentum) y otros. El más usado

es Adam, porque incorpora la mayoría de las ventajas de otros algoritmos, y normalmente presenta mejor desempeño que el resto de los métodos [27].

3.1.2.6.5. Tasa de aprendizaje

El aprendizaje en la red ocurre al tomar lotes aleatorios de muestras de datos, calcula el gradiente de los parámetros de red con respecto a la pérdida en el lote. Luego, los parámetros de la red se actualizan en la magnitud (valor positivo en el rango de 0 a 1) definida por la tasa de aprendizaje o ratio de aprendizaje (*learning rate*) en una dirección opuesta al gradiente [26].

El hiperparámetro de la tasa de aprendizaje de la red, es posiblemente el hiperparámetro más importante [30], porque interviene directamente en el entrenamiento de la red afectando el rendimiento y el tiempo de respuesta. Durante el entrenamiento, la retropropagación del error estima la "cantidad de error" de la que son responsables los pesos de un nodo en la red. En lugar de actualizar el peso con la cantidad total, se escala según la tasa de aprendizaje. Esto significa que con una tasa de aprendizaje de 0,1 (valor predeterminado muy común) los pesos en la red se actualizan 10% del error de peso estimado cada vez.

El incremento inapropiado de esta tasa puede producir oscilaciones (zigzagueo) en la pérdida al momento de actualizar los parámetros, provocando aumento en la pérdida e incluso desviaciones del mínimo [30]. De forma general, en Figura 11 se visualiza el aspecto de la curva de aprendizaje con una tasa de aprendizaje que toma valores extremos. Si el valor es demasiado alto el algoritmo diverge. Un valor muy alto produce que la curva decaiga abruptamente en las primeras épocas y luego se mantiene casi estable, en este caso el algoritmo puede ser subóptimo por haber encontrado un mínimo parcial de la función de pérdida. Con un valor muy bajo, el modelo tiene alta probabilidad de encontrar el mínimo de la función, pero será lento. Una buena curva decae a lo largo de las épocas.

Figura 11- Curvas de aprendizaje para distintas tasas de aprendizaje. Fuente: Hands-On Machine Learning with Scikit-Learn and TensorFlow [30]

3.1.3. Entrenamiento

Para el entrenamiento de una red se debe especificar el número de épocas (*epochs*), este determina cuántas veces se pasa el conjunto de datos completo a través del modelo. El rendimiento del modelo resulta afectado por este parámetro. El incremento en el número de épocas puede mejorar, como así también lo puede deteriorar por un sobreajuste del modelo al conjunto de datos de entrenamiento [20].

3.1.3.1. Sobreajuste (*overfit*)

El sobreajuste refiere al hecho de que los modelos de aprendizaje automático tienden a funcionar peor con datos de prueba no vistos que con los datos de entrenamiento. La solución encontrada por el modelo no generaliza bien, existiendo siempre una brecha entre el rendimiento de los datos de entrenamiento y los de prueba, que es particularmente grande cuando los modelos son complejos y el conjunto de datos de entrenamiento es pequeño [26][27].

Para evitarlo, el conjunto de datos con el que se entrena el modelo debe ser distinto al conjunto de datos usado en la evaluación [26]. Además, aumentar el número de instancias de entrenamiento mejora el poder de generalización del modelo, mientras que aumentar la complejidad del modelo reduce su poder de generalización. Por otro lado, cuando hay muchos datos de entrenamiento disponibles, es poco probable que un modelo demasiado simple capture relaciones complejas entre las características y el objetivo [27].

El sobreajuste es un obstáculo para lograr que los modelos generalicen, es decir, que funcionen bien con datos nunca antes vistos. Es importante poder medir de manera confiable la capacidad de generalización de un modelo.

El problema fundamental en el aprendizaje automático es la puja entre optimización y generalización. La optimización ajusta un modelo para obtener el mejor rendimiento posible con los datos de entrenamiento, mientras que la generalización se refiere a qué tan bien se desempeña el modelo entrenado con datos que nunca antes ha visto.

Al comienzo del entrenamiento, la optimización y la generalización están correlacionadas, a menor pérdida en los datos de entrenamiento, menor es la pérdida de datos de validación. Mientras esto sucede, el modelo no es adecuado, es decir, la red aún no ha modelado todos los patrones relevantes de los datos de entrenamiento.

Después de un cierto número de iteraciones en el entrenamiento, la generalización deja de mejorar, las métricas de validación se detienen, para luego comenzar a degradarse, en ese momento, el modelo está comenzando a sobreajustarse. Es decir, está comenzando a aprender patrones que son específicos de los datos de entrenamiento, pero que son engañosos o irrelevantes cuando se trata de datos nuevos.

Las formas más comunes de evitar el sobreajuste en las redes neuronales son: obtener más datos de entrenamiento, reducir la capacidad de la red, agregar regularización de peso y agregar abandono (dropout) [26].

3.1.3.2. Regularización

La regularización consiste en restricciones a la complejidad de una red, obligando a que los pesos sólo tomen valores pequeños para simplificar el modelo (la distribución de los pesos es más regular) y reducir el riesgo de sobreajuste. Esto se denomina regularización de peso y se realiza agregando a la función de pérdida de la red un costo asociado por tener pesos grandes. El valor de los hiperparámetros de regularización se aplica durante el aprendizaje permaneciendo constante durante el entrenamiento [26][30].

Este costo se presenta en dos valores: 1) Regularización L1, el costo agregado es proporcional al valor absoluto de los coeficientes de ponderación, 2) Regularización L2, el costo agregado es proporcional al cuadrado del valor de los coeficientes de ponderación [26]. Un valor muy grande de regularización, obtendrá un modelo casi plano (una pendiente cercana a cero), seguramente el algoritmo de aprendizaje no

sobreajustará los datos de entrenamiento, pero será poco probable que encuentre una buena solución [30].

3.1.3.3. Deserción o abandono (dropout)

La deserción (*dropout*) es una de las técnicas de regularización más efectivas y más utilizadas para las redes neuronales. La deserción, aplicada a una capa, consiste en eliminar aleatoriamente un conjunto de características de la salida de la capa durante el entrenamiento. La tasa de abandono indica la fracción de características que se ponen a cero; generalmente se establece entre 0.2 y 0. 5 [26].

3.1.4. Evaluación de un modelo

En el aprendizaje automático, el objetivo es lograr modelos que generalicen, que funcionen bien con datos nunca antes vistos, por lo que es crucial poder medir de manera confiable la capacidad de generalización del modelo[26].

La evaluación de un modelo se reduce a dividir los datos disponibles en tres conjuntos: entrenamiento, validación y prueba. Se entrena con los datos de entrenamiento y se evalúa el modelo con los datos de validación. Una vez que el modelo está listo (desempeño aceptable), se prueba una última vez con los datos de prueba.

Existen distintas métricas para analizar el desempeño de un modelo. De acuerdo al interrogante planteado al inicio del proceso y al tipo de modelo, serán las métricas que se empleen para analizarlo.

3.1.4.1. Métricas para modelos de clasificación

La forma común de mostrar las métricas de rendimiento para un clasificador es con una matriz de confusión [23]. Considera los conceptos del resultado del clasificador y la verdad básica real [21]. En un problema binario, la matriz posee dos filas y dos columnas y muestra la cantidad de predicciones de cada categoría frente a la categoría en la que deberían haber sido predichos.

Entonces se presentan cuatro casos posibles: 1) Verdaderos positivos (TP), el clasificador predice una muestra como positiva y realmente es positiva; 2) Falsos positivos (FP), el clasificador predice una muestra como positiva, pero en realidad es negativa; 3) Verdaderos negativos (TN), el clasificador predice una muestra como negativa y realmente es negativa; 4) Falsos negativos (FN): El clasificador predice una

muestra como negativa, pero en realidad es positiva. Se resume esta información en la Tabla 1.

		Predicción	
		Positiva	**Negativa**
Real	**Positiva**	**TP**	**FN**
	Negativa	**FP**	**TN**

Tabla 1 - Matriz de confusión

La matriz de confusión permite definir diferentes métricas como son la exactitud (accuracy), precisión, sensibilidad (recall) y el F1-score.

La exactitud es la fracción de casos predichos correctamente respecto del total de casos (Ecuación 13).

$$Exactitud = \frac{TP + TN}{TP + FN + FP + TN} \tag{13}$$

La precisión es la proporción de casos clasificados como positivos correctamente, respecto del total de casos positivos predichos (Ecuación 14).

$$Precisión = \frac{TP}{TP + FP} \tag{14}$$

La sensibilidad (recall) mide la cobertura de un clasificador, la fracción de casos predichos como positivos correctamente, respecto del total de casos positivos reales (Ecuación 15).

$$Sensibilidad = \frac{TP}{TP + FN} \tag{15}$$

La métrica F1-score, es la media armónica de la precisión del clasificador con su sensibilidad(recall) (Ecuación 16).

$$F1 = \frac{2 * Precisión * Sensibilidad}{Precisión + Sensibilidad} \tag{16}$$

El valor de F1 será de 1 para un clasificador perfecto y de 0 en el peor de los casos.

3.1.4.2. Métricas para modelos de regresión

Existen tres métricas principales cuando se utilizan modelos de aprendizaje automático de regresión, ellas son 1) El error absoluto medio (MAE), 2) El error

cuadrático medio (MSE), y 3) Raíz del error cuadrático medio (RMSE). Cada métrica intenta describir y cuantificar la eficacia de un modelo de regresión comparando una lista de predicciones con una lista de respuestas correctas. Todas son funciones de pérdida, por lo tanto, se busca minimizar sus valores [22].

Se *n* la cantidad de observaciones, y_i el valor real e $\hat{y}_i$ el valor predicho, se calcula cada una de las métricas.

El error absoluto medio (MAE) calcula la media del valor absoluto de los errores. Es fácil de entender y denota, en promedio, qué tan equivocado está el modelo (Ecuación 17) [22] .

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

El error cuadrático medio (MSE) calcula la media del cuadrado de los errores. Castiga los errores más grandes, lo que tiende a ser mucho más útil en el mundo real (Ecuación 18) [22].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

Raíz del error cuadrático medio (RMSE) calcula la raíz cuadrada de la media del cuadrado de los errores (Ecuación 19).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

El coeficiente de determinación R^2 se describe como la fracción de la varianza que el modelo tiene en cuenta. Un valor de 1 significa una coincidencia perfecta y un valor de 0 significa que no capturó ninguna de las variaciones (Ecuación 20) [23].

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} \tag{20}$$

3.2. Series Temporales

La Ciencia de Datos resuelve problemáticas sobre la lluvia, la temperatura atmosférica, el crecimiento de la población, el producto interno bruto (PBI) y similares,

utilizando datos que se miden a intervalos regulares de tiempo real. Estos datos representan las variaciones temporales de una entidad en puntos de tiempo fijos, dentro de un intervalo de tiempo finito y que juntos describen una serie de tiempo.

La inclusión del intervalo de tiempo real finito en la definición de una serie temporal tiene sentido desde la perspectiva de su predicción en un punto de tiempo no incluido en el conjunto de datos registrados. Es decir, si la motivación de la serie temporal es sólo preservar datos históricos, ignorando predicciones futuras, la restricción en el intervalo fijo y finito de serie temporal puede ser eliminada.[31].

La predicción en una serie temporal es de gran utilidad y tiene amplia aplicación en el mundo real, y la predicción errónea en series temporales puede generar grandes pérdidas para las instituciones. Las redes neuronales juegan un papel vital en el aprendizaje del comportamiento dinámico de una serie temporal [31].

Para manejar el comportamiento no determinista de la serie temporal, los investigadores generalmente no se arriesgan a predecir el valor absoluto del siguiente punto de la serie, sino que ofrecen un pequeño rango del valor predicho. Esto se logra dividiendo el rango dinámico de la serie temporal en un número fijo de particiones, generalmente de igual longitud, lo que reduce la incertidumbre en la predicción ya que la predicción se refiere a identificar una partición que contiene el siguiente valor de punto de tiempo en lugar de un valor absoluto [31].

3.2.1. Pronóstico sobre series temporales

La mayoría de los algoritmos de aprendizaje automático funcionan mejor cuando los conjuntos de datos están equilibrados, si esto no sucede el algoritmo tiende a favorecer las muestras de clase mayoritaria [32]. El desequilibrio en los datos ocurre cuando ciertos rangos de valores están sobrerrepresentados en comparación con otros.

Para hacer frente a este problema de aprendizaje desequilibrado, existen distintas técnicas de submuestreo y sobremuestreo, una de las más usadas es SMOTE. El submuestreo elimina instancias de la clase mayoritaria de forma aleatoria, lo que puede llevar a pérdida de información importante. Mientras que en el sobremuestreo, se agregan nuevas muestras a la clase minoritaria para equilibrar el conjunto de datos [32].

Las técnicas de remuestreo (submuestreo y sobremuestreo) no siempre son adecuadas para aplicaciones que tratan con series temporales. Esto se debe a que estas estrategias implican la pérdida de la correlación del tiempo y la dependencia entre las mediciones [33].

En el pronóstico a través de ANN se han utilizado otras técnicas para mitigar los efectos del problema de los datos desequilibrados en series temporales. En la fase de preprocesamiento de datos se aplica ventana deslizante, normalización de datos y K-Fold. Luego se entrena la red neuronal de forma que se garantice la dependencia temporal de los datos [33].

Existen varias técnicas para el pronóstico de datos de series temporales de manera efectiva como Autoregresivo univariado (AR), Promedio móvil univariado (MA), Suavizado exponencial simple (SES) y Promedio móvil integrado autorregresivo (ARIMA). De todos ellos, ARIMA ha demostrado un rendimiento superior en precisión y exactitud al predecir los próximos valores de las series temporales. Sin embargo, los estudios empíricos realizados e informados por Siaamimi-Nni et all. [29] muestran que los algoritmos basados en el aprendizaje profundo, como el LSTM, presentan mejores resultados que los modelos ARIMA. Más específicamente, el algoritmo basado en LSTM mejoró la predicción en un 85% en promedio en comparación con ARIMA.

3.3. Heladas

Técnicamente, la palabra “helada” se refiere a la formación de cristales de hielo sobre las superficies, tanto por congelación del rocío como por un cambio de fase de vapor de agua a hielo [13]. Se produce cuando la superficie terrestre y el aire que se siente sobre ella alcanza una temperatura por debajo de los 0°C [2]. Algunos autores definen la helada como la ocurrencia de una temperatura menor o igual 0°C medida en una garita “tipo Stevenson” ubicada a una altura entre 1,25 y 2 metros [10][13].

La atmósfera recibe energía proveniente del sol en forma de radiación. Una fracción de la energía es absorbida por la tropósfera (capa de la atmósfera más cercana a la Tierra donde se presentan los fenómenos meteorológicos). Otra parte se dirige al exterior, al ser difundida desde la atmósfera hacia el espacio y el resto llega a la superficie de la Tierra. En las noches con cielo cubierto por nubes, gran parte de la

energía que se difunde desde la corteza de la Tierra (radiación de calor proveniente del suelo) es reflejada por estas masas de humedad hacia el planeta; otra parte de ella es absorbida y la restante es enviada al espacio. Cuando de una región de la superficie terrestre se desprende una mayor cantidad de calor que el recibido, ocurre un enfriamiento que favorece la formación de la helada. Esto generalmente se presenta en la madrugada o cuando está saliendo el sol [34].

Los balances de radiación en una zona de la superficie terrestre no son los mismos a lo largo del tiempo, y dependen de la ubicación sobre la Tierra, porque la inclinación de los rayos solares que llegan a la zona influye en la cantidad de energía que ésta recibe.

Las situaciones favorables para que se produzcan las heladas se pueden distribuir en dos clases distintas, una viene dada por las condiciones locales y la otra por las condiciones meteorológicas actuales. Entre las condiciones locales se encuentra la exposición del terreno, la proximidad de bosques, la latitud y altitud [10]. Respecto de las condiciones meteorológicas, los principales elementos que influyen son el viento, la nubosidad, la humedad atmosférica y la radiación solar.

3.3.1. Clasificación de las heladas

Según el origen climatológico, las heladas se clasifican en heladas por advección y heladas por radiación. Las heladas por advección están asociadas con incursiones a gran escala de aire frío con una atmósfera con viento y bien mezclada, y una temperatura que a menudo está por debajo de cero, incluso durante el día. Las heladas por radiación se presentan por la pérdida de calor del suelo durante las noches despejadas y en calma, y con inversiones de temperatura [13]. Las heladas por advección suelen ser esporádicas, mientras que las heladas de radiación ocurren más a menudo.

Como se mencionó anteriormente, durante el día el suelo se calienta, pero al anochecer pierde calor por radiación, con mayor cantidad en las noches largas de invierno. Los lugares más propensos a la formación de heladas por radiación son tanto los valles como las cuencas y hondonadas próximas a las montañas. Ello se debe a la acumulación del aire frío que desciende durante la noche. Se originan cuando el aire cercano a la superficie del suelo tiene una humedad relativa baja y disminuye aún más por la llegada de un viento con aire seco. Esto último causa la evaporación del agua que se encuentra sobre las plantas, lo que provoca su enfriamiento [34].

Hay dos subcategorías de heladas de radiación: la blanca y la negra. Una helada blanca ocurre cuando el vapor de agua se deposita sobre la superficie y forma una capa blanca de hielo que se denomina normalmente escarcha. Cuando la humedad es alta es más probable que se produzca una helada blanca. La helada negra ocurre cuando la temperatura cae por debajo de 0 °C y no se forma hielo sobre la superficie. Si la humedad es suficientemente baja, entonces la temperatura de la superficie puede que no alcance la temperatura del punto de formación de hielo y no se formará escarcha.

Una característica de la temperatura del aire en las heladas de radiación nocturnas es que su mayor caída se produce en unas pocas horas después de la puesta del sol, cuando la radiación neta sobre la superficie cambia rápidamente de positiva a negativa. Este cambio rápido en la radiación neta ocurre porque la radiación solar disminuye desde el valor más alto en mediodía hasta cero en la puesta de sol [13].

Es posible la ocurrencia de heladas mixtas, que se producen por una combinación de condiciones advectivas y radiactivas. No es extraño tener condiciones advectivas que traen una masa de aire frío en una región provocando una helada advectiva. Esto puede venir seguido por varios días despejados, con condiciones de calma que conducen a heladas de radiación [13].

3.3.2. Daños de la helada

Los cultivos son vulnerables a la helada, el proceso de deterioro de las plantas depende de la especie a la que pertenece y del estado vegetativo en que se encuentre. El cultivo es más resistente a la helada cuando se encuentra en el periodo de germinación, mientras que en la floración es mayor el daño que sufre. Los efectos que causa la helada en el interior de la planta es la ruptura de las membranas de la célula por el crecimiento de cristales de hielo. Y en el exterior se produce la muerte de hojas y tallos tiernos, destrucción de un gran porcentaje de flores y frutos pequeños, e incluso la muerte total de la planta [34].

En las heladas por advección se produce un continuo movimiento de aire frío sobre los cultivos. Los daños que sufre la planta dependen de su naturaleza y de la etapa de desarrollo en que se encuentre. Las heladas por radiación afectan principalmente a las plantas con flores y a las hortalizas.

A mayor duración e intensidad de la helada los daños son mayores. La intensidad se refiere a la velocidad de descenso de la temperatura, es decir, cuantos grados desciende la temperatura por hora. La mínima alcanzada será la magnitud. A una misma temperatura puede o no haber daños, en función de que la helada haya tenido una duración larga o muy efímera. Para una temperatura dada, el tiempo de exposición al frío está muy relacionado con la aparición gradual de daños. Así, suele resultar más perjudicial una temperatura de -2°C a -3° C durante varias horas que una temperatura mucho más baja en períodos menores de media hora [4].

Los daños para una misma temperatura negativa dependen no solamente de la duración de la exposición, sino también de las temperaturas de los días anteriores, así como de la sequedad de la atmósfera, siendo las condiciones más desfavorables un descenso brusco de la temperatura, después de un período suave y húmedo [4].

Analizando las heladas desde un punto de vista geográfico, los daños por heladas pueden producirse en cualquier localidad, fuera de las zonas tropicales. En gran medida, la probabilidad de temperaturas bajo cero está afectada por las condiciones locales. Muchos agricultores tienen una idea acertada sobre la localización de las zonas frías en su localidad. Es menos probable el daño cuando la masa de tierra es un área donde sopla el viento o está rodeada de grandes cuerpos de agua, por el efecto moderador del ambiente marítimo sobre la humedad y la temperatura [13].

3.3.3. Protección contra las heladas

La preocupación de los agricultores para proteger sus cultivos de las heladas se debe a las fuertes pérdidas económicas y naturales que pueden presentarse durante el ciclo agrícola. Existen varios métodos para reducir los efectos de las heladas en cultivos, los cuales se agrupan en indirectos (o pasivos) y directos (o activos).

Los métodos indirectos son los que actúan en términos de prevención, normalmente para un periodo largo de tiempo. Se relacionan con técnicas biológicas y ecológicas, e incluyen prácticas llevadas a cabo antes de las noches de helada para reducir el potencial de daño [13]. Disminuyen la afectación durante el periodo de helada, por la elección apropiada de las especies, variedades, épocas de cultivo y ubicación de las distintas plantas [34].

Los métodos directos o activos se basan en acciones temporales tomadas antes y durante el periodo de peligro de la helada. Se basan en métodos físicos e intensivos desde el punto de vista energético [13] que buscan reducir la pérdida de calor del suelo o producir el calentamiento directo del aire. En algunos casos se protege con cajones, cestos, entablillados de madera u otros elementos vegetales cuando las características del cultivo lo permiten. Otro recurso es producir nieblas o humos (quemando productos naturales o sustancias químicas) en la capa de aire adyacente a la superficie del suelo y reponen las pérdidas de calor agregando una cierta cantidad de él. También se suele recurrir al calentamiento directo del aire y la planta mediante calentadores, antes de que la temperatura sea crítica para las plantas. O bien, recurriendo al método más antiguo que es el uso del agua ya sea con inundación o por aspersión [34].

Capítulo 4

Metodología

4. Metodología

En la investigación se ha recopilado conjuntos de datos numéricos, los mismos se han analizado y procesado a través de modelos matemáticos y estadísticos para predecir la temperatura a futuro y determinar la ocurrencia o no de la helada; por lo tanto, se ha llevado a cabo una de investigación experimental de carácter cuantitativo.

4.1. Metodología aplicada en el desarrollo del modelo

Específicamente se ha aplicado el proceso de Ciencia de Datos, que establece una secuencia de fases para la resolución de un problema. Esta secuencia no es estrictamente lineal, en cualquier momento del proceso es factible que sea necesario regresar a la fase anterior para hacer modificaciones. La Figura 12 resume el flujo del trabajo realizado, iniciando con el entendimiento del contexto del problema a través de información brindada por expertos. Se han obtenido los datos desde dos estaciones meteorológicas, los cuales fueron sometidos a la instancia de preprocesamiento, donde se analizaron, limpiaron, unificaron formatos e identificaron los datos más relacionados con la variable a predecir para identificar la ocurrencia del fenómeno de la helada. El desarrollo detallado de estas tareas se presenta en el capítulo 5.

Al completar este análisis exploratorio de los datos, se han preparado los datos en una estructura de ventana deslizante, adecuada para el procesamiento con algoritmos redes neuronales recurrentes de tipo LSTM. Luego, en la fase de modelado se ha desarrollado una serie de modelos; en primer lugar, para identificar los valores adecuados de los hiperparámetros de las redes LSTM que procesarían los datos, como son la tasa de aprendizaje y regularización, los cuales afectan el sobreajuste de la red. Una vez encontradas estos hiperparámetros de red, se aplicaron a modelos en los cuales se han variado la cantidad de atributos de entrada. En el capítulo 6 se explicita la fase de modelado

Finalmente se evaluaron los modelos, a través de los resultados obtenidos por las métricas y la visualización de las predicciones realizadas, identificando en el conjunto de modelos desarrollados el más simple (menor cantidad de datos de entrada y estructura de red con menos componentes) que predice la helada con mejores resultados.

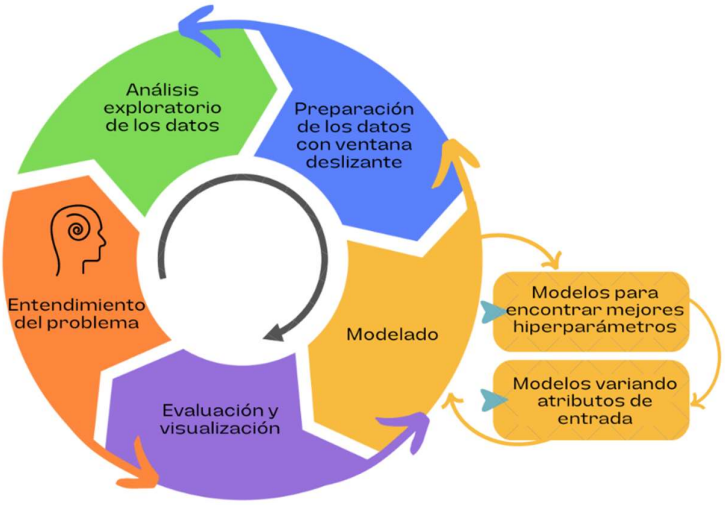


Figura 12- Proceso de Ciencia de Datos aplicado a en la investigación. Fuente: Autora

4.2. Datos

El análisis del fenómeno meteorológico de la helada en la zona sur del Valle del Tulum, se ha realizado a partir de los datos registrados por dos estaciones agrometeorológicas. Los datos han sido provistos por Instituto de Automática de la Facultad de Ingeniería de la Universidad Nacional de San Juan, en el caso de la estación situada en el Establecimiento San Francisco S.A. (explotación privada), ubicado en la localidad de Cañada Honda, departamento Sarmiento, provincia de San Juan, Argentina. Y por el Servicio de Agrometeorología de la Estación Experimental Agropecuaria San Juan dependiente del Instituto Nacional de Tecnología Agropecuaria (INTA) en el departamento Pocito donde se encuentra instalada la otra estación meteorológica. Existiendo entre ellas una distancia de 37 km aproximadamente (Figura 13).

Figura 13 - Ubicación geográfica de las estaciones meteorológicas

Se trata de una estación meteorológica Davis modelo Vantage Pro 2. El software que usa es Weatherlink 5.8.2. Se encuentra en abrigo a 1,50 m de altura desde la superficie. Posee sensores que miden distintas variables y a partir de ellas calcula otras variables de interés. Registrando valores para las variables: temperatura exterior, humedad, velocidad y dirección del viento, precipitación, presión atmosférica, radiación solar, punto de rocío, evapotranspiración, índice de calor, índice de temperatura y humedad, entre otras. La medición y registro de los datos de las variables se realiza cada 10 minutos.

De la estación situada en INTA se consideraron datos desde el año 2.016 hasta julio del año 2.021, y del Establecimiento San Francisco desde el mes de abril de 2.013 al año 2.018.

4.2.1. Unificar y limpiar de datos

Los datos obtenidos presentan dos tipos de problemas. En el primer tipo se incluyen problemas de formato, como caracteres inconsistentes, espacios en blanco extraños y entre otros. Estos suelen ser resueltos con el preprocesamiento adecuado. El segundo tipo implica el contenido actual de los datos, como son valores atípicos y los valores nulos o vacíos. Estos requieren un análisis que permita descubrir el significado de estos problemas en una situación particular y cómo deben abordarse [23].

Para el tratamiento de los valores faltantes existen distintas opciones. Una de ellas es eliminar los registros que poseen valores nulos. Otra es reemplazar los valores inexistentes por el valor medio de la columna, sin indicar que se hizo dicho reemplazo; o la opción anterior, pero agregando una columna al conjunto de datos, en dicha columna se identifica si el registro posee o no valores faltantes [24]. En este caso se optó por la primera opción.

Por cada año se han seleccionado los datos correspondientes al periodo comprendido entre la fecha de la primera helada hasta la fecha de la última helada, evidenciando que el periodo de heladas se constituye desde el mes de mayo hasta el mes de octubre. Las heladas son frecuentes entre las 00 horas y 8 horas y, considerando que los datos se han estructurado en ventanas de entrada de 3 horas y horizonte de 3 horas, se han seleccionado sólo los registros que se encuentran en el rango horario desde las 18 horas hasta las 9:50 horas. De este modo se ha garantizado la disponibilidad

de una cantidad suficiente de las lecturas previas, a la hora en la cual podría ocurrir el fenómeno meteorológico bajo estudio.

Para la selección de las variables meteorológicas se ha elaborado la matriz de correlación lineal con las variables: temperatura mínima, humedad relativa, punto de rocío, velocidad del viento, presión atmosférica y radiación solar.

4.2.2. Balanceo de datos

Con frecuencia el conjunto de datos reales con el que se entrena el algoritmo está desequilibrado, situación muy común cuando se modelan eventos inusuales o temporales como son las heladas en el área de la meteorología. El sesgo en el conjunto de datos de entrenamiento suele reflejarse en el rendimiento del modelo, obteniendo un modelo más preciso para los casos mayoritarios, que para los casos que son minoría.

Un enfoque para abordar los problemas de desequilibrio en los datos consiste en aplicar durante la etapa de preprocesamiento métodos de remuestreo. Uno de ellos es el sobremuestreo, que incrementa de forma sintética la cantidad casos de la clase minoritaria; otro es el submuestreo que elimina casos de la clase mayoritaria de forma aleatoria; y otro método consiste en aplicar ambos tipos de muestreo. Algunos métodos de remuestreo son SMOTE, ADASYN, SMOTEtomek y SMOTEENN [35]. En este trabajo se ha aplicado este último.

4.2.3. Estructura de los datos de entrada al modelo

Proveer a una red neuronal con datos que toman valores grandes o datos heterogéneos, puede provocar grandes actualizaciones del gradiente que evitarán que la red converja. Para facilitar el aprendizaje de una red neuronal, los datos deben ser homogéneos y de valores pequeños, en lo posible, variar en el rango de 0 a 1 o de -1 a 1. Para que los datos adquieran estas características se lleva a cabo una tarea de normalización para cada característica que conforman el conjunto de datos [26].

En un conjunto de datos de series temporales usado para alimentar el modelo, los valores de las sucesivas marcas de tiempo están estrechamente relacionados con los valores en la ventana anterior. Si el modelo trata de predecir la i-ésima temperatura T_i, a partir de una ventana de tamaño t. Las temperaturas T_{i-t}, T_{i-t+1}, …T_{i+t-1}, T_{i+t} se utilizan para predecir la temperatura objetivo. Las ventanas analizadas han sido de 2, 3 y 4 horas

previas. El horizonte es de 3 horas, de modo que permite generar alarmas que avisen al productor para que active medidas de mitigación contra el fenómeno.

Para los sistemas de aprendizaje automático los datos se deben estructurar en tensores. Un tensor es un contenedor (arreglo multidimensional) de datos numéricos. Independientemente de los datos que se vayan a procesar, primero deben ser convertidos en tensores, a esta tarea se la denomina vectorización de datos [26].

4.2.4. Datos de entrenamiento, validación y prueba

Tanto el conjunto de datos de entrenamiento como el de prueba deben ser representativos de los datos disponibles. En modelos que tratan de predecir el futuro a partir del pasado (por ejemplo, el clima de mañana, los movimientos de las acciones, etc.), no se deben mezclar aleatoriamente los datos antes de dividirlos, ya que al hacerlo se creará una fuga temporal [26]. La estrategia empleada para conformar estos conjuntos de datos fue usar los datos de la estación San Francisco para el entrenamiento y validación, y los datos de la estación INTA para test.

4.3. Modelo

Las RNN de tipo LSTM son potentes para procesar datos secuenciales como las series temporales, se debe tener presente que el tamaño (cantidad de datos) de la serie impacta directamente en el tiempo de procesamiento requerido por la red. Por ello, se ha usado este tipo de red neuronal para desarrollar modelos de regresión que reciben como entrada distinta cantidad de lecturas secuenciales (serie temporal).

Para la predicción de la temperatura se han desarrollado variados modelos, con redes neuronales densas y con redes neuronales recurrentes LSTM. Para la entrada se han constituido distintos conjuntos de datos, con variaciones en las variables meteorológicas consideradas, tomando las lecturas cada 10 minutos. En todos los casos se ha mantenido el periodo previo (ventana) de 3 horas y el horizonte de 3 horas. La variable meteorológica temperatura es entrada en todos los modelos, en algunos de ellos se incorpora la humedad relativa, por existir entre ellas mayor correlación entre el conjunto de variables analizadas a través de la matriz de correlación.

Se han desarrollado modelos de RNN LSTM con diferentes arquitecturas, variando la capa de entrada en función del conjunto de datos considerado. Una capa oculta y la

capa de salida de una neurona se mantuvo en todos los modelos. La función de activación usada en la capa oculta es RELU. En la función de pérdida (*loss*) se utiliza el MSE. En cuanto al algoritmo de optimización es Adam.

En cuanto a la evaluación de los modelos se han analizado el MSE, la RMSE y el R^2 en general para el modelo de regresión. Y en particular para las heladas, el resultado de la predicción (valor continuo) se ha clasificado con una etiqueta indicando la predicción de helada o no, según corresponda. A partir de esto se ha analizado la matriz de confusión y las métricas de sensibilidad y F1-score para los casos de heladas.

Capítulo 5

Análisis Exploratorio de los Datos

5. Análisis exploratorio de los Datos

La problemática de predicción de las heladas ha sido enmarcada como un problema de Ciencia de Datos, buscando la solución a través del tipo aprendizaje automático supervisado con algoritmos de regresión. Se han usado modelos basados en redes neuronales de tipo LSTM que, a partir de datos obtenidos de estaciones meteorológicas, pronostican la temperatura hacia un horizonte de tres horas.

En todo el proceso se ha usado el lenguaje de programación Python haciendo uso de las funcionalidades que proveen distintas librerías de código abierto, como NumPy para las distintas operaciones numéricas para el preprocesamiento y estructuración de los datos [36]. La librería scikit-learn [37] proporciona algoritmos de clasificación, agrupamiento y regresión que se complementa con TensorFlow [38] para el desarrollo de los modelos de aprendizaje automático. La librería Matplotlib que permite crear distintos tipos de visualizaciones [39].

5.1. Comprensión de los datos

Los valores de cada lectura realizada por cada estación agrometeorológica se almacenan en un archivo de formato .txt. El registro sucede cada 10 minutos y se dispone de un archivo por año y por estación. De la estación situada en INTA se han considerado datos desde el año 2.016 hasta el año 2.021, y del Establecimiento San Francisco desde el mes de abril del año 2.013 al año 2.018.

Los valores corresponden a distintas variables, como la temperatura exterior, máxima, mínima; la humedad; punto de rocío; características del viento, velocidad, dirección, velocidad máxima, velocidad de la ráfaga; sensación térmica; precipitación; presión atmosférica; radiación solar y ultravioleta; evapotranspiración y un conjunto de índices que son calculados (índice calor, índice THW, índice THSW, entre otros).

Los datos son de tipo estructurado que pueden ser tabulados. Los encabezados de las tablas son distintos, los de la estación INTA están en inglés y los de la estación San Francisco en español y los nombres de columnas que están abreviados, incluyen caracteres especiales como el punto '.' y el espacio en blanco. Cada tabla posee 36 columnas (variables) y la cantidad de filas es variable, debido a que distintos tipos de

interrupciones en el funcionamiento de las estaciones han ocasionado la ausencia de registros para estos periodos de falla.

Respecto al tipo de dato, ambas estaciones registran la variable fecha en el formato 'dd/mm/aa'; las variables que refieren a la dirección del viento son cualitativas (punto cardinal) y el resto de variables meteorológicas son de tipo real. En cuanto a la variable hora de lectura, San Francisco registra en formato de 24 horas, mientras que la estación INTA lo hace en formato de 12 horas AM y PM, representado por 'a.m.' y 'p.m.' respectivamente. Un valor que no haya sido leído o calculado por una falla técnica en la estación se registra con tres guiones medios '---'.

Para el análisis del fenómeno de la helada, la temperatura es la variable meteorológica más importante. La estación registra temperatura exterior, la cual refiere al exterior de la garita donde se encuentra ubicada; además la temperatura máxima y mínima registrada durante el intervalo de 10 minutos. De estas tres temperaturas se ha analizado la temperatura mínima, presentando en la Tabla 2 una breve descripción de esta variable para cada conjunto de datos y de forma gráfica en la Figura 14.

	San Francisco	INTA
Cantidad de casos	299.671	301.899
Media	16,27	18,36
Desviación estándar	9,55	8,99
Valor mínimo	-9,60	-5,10
Valor máximo	42,70	43,20
25%	9,30	11,70
50%	16,40	18,60
75%	23,10	25,10

Tabla 2 - Descripción de la variable temperatura mínima para cada estación

Se observan ciertas similitudes en los conjuntos de datos. La cantidad de datos es aproximadamente 300.000. El valor del primer cuartil es muy superior a 0, indicando que la cantidad de valores correspondientes a heladas (menores o iguales a cero) son menos del 25% del total de los datos. La temperatura máxima es próxima a los 43°. En cambio, el valor mínimo es notablemente menor para San Francisco. Se percibe que ninguno de los conjuntos de datos posee valores atípicos (*outliers*).

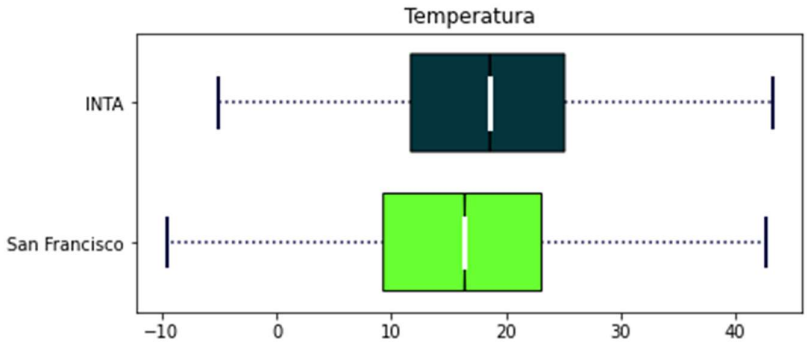


Figura 14 - Descripción de la variable temperatura para las dos estaciones meteorológicas.

La variable temperatura mínima se ha representado en un histograma (Figura 15), resultando un gráfico con gran simetría en los datos. Donde se aprecia claramente que en ambas estaciones la mayor cantidad de datos corresponden a temperaturas mayores a cero. También se observa que la cantidad de temperaturas menores o iguales a cero de la estación San Francisco (Figura 15 (a)) se extiende hacia el valor -10, siendo mayor la cantidad de heladas que las registradas por la estación INTA (Figura 15 (b)).

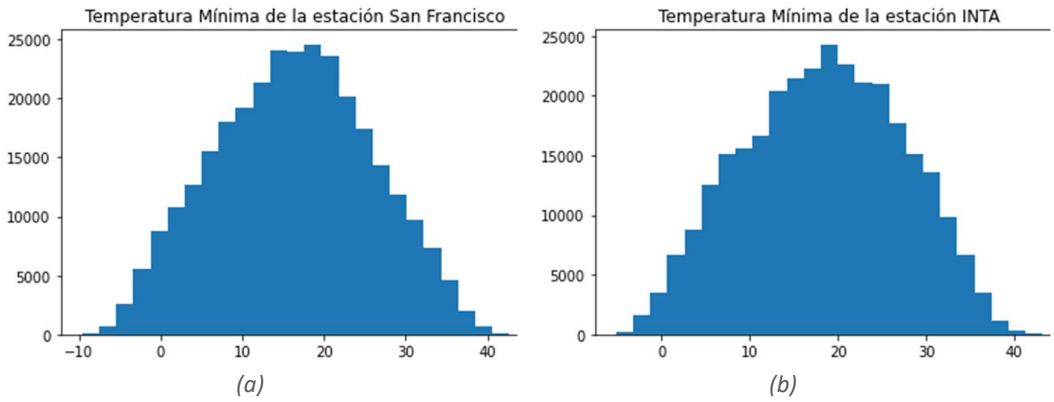


Figura 15- Distribución de la temperatura. (a) Estación San Francisco (b) Estación INTA

5.2. Preprocesamiento de los datos

Entre las tareas de preprocesamiento fue necesario:

a) Determinar un formato para los encabezados de las tablas de datos, seleccionando el formato de la estación INTA porque los nombres de las columnas no poseen caracteres especiales, característica que sí presentan los datos de San Francisco y pueden ocasionar inconvenientes al momento de ser procesados por el lenguaje de programación.

b) Unificar el formato de la hora. Se optó por el formato de 24 horas.

c) Datos no registrados, se eliminaron los registros donde el valor de la temperatura había sido registrado como '---'.

En la estación San Francisco, para el periodo de heladas del año 2.018 se ha identificado la falta de un registro correspondiente a la fecha 2018-10-18 13:20. En cambio, para la estación de INTA se ha detectado la falta de registro de datos para distintos períodos durante la época de heladas, estos se presentan en la Tabla 3.

Desde	Hasta	Periodo faltante
2020-05-06 05:10	2020-05-06 14:20	0 días 9:10
2020-05-25 06:40	2020-05-27 11:20	2 días 04:40
2020-06-06 11:40	2020-06-10 10:00	3 días 22:20
2020-06-10 11:00	2020-06-10 12:10	0 días 1:10
2020-06-16 08:40	2020-06-16 19:40	0 días 11:00
2020-06-24 08:10	2020-06-24 08:30	0 días 00:20
2020-06-24 08:50	2020-06-24 11:40	0 días 02:50
2020-06-24 12:00	2020-06-24 12:20	0 días 00:20
2020-07-08 11:40	2020-07-08 12:30	0 días 00:50
2020-07-08 12:40	2020-07-08 13:00	0 días 00:20
2021-06-01 00:00	2021-06-12 14:30	11 días 14:30

Tabla 3 - Datos faltantes en la estación INTA

Un primer análisis, para contextualizar el fenómeno de la helada en la zona donde se encuentran las estaciones meteorológicas, ha consistido en conocer el período de ocurrencia del fenómeno, la cantidad de casos por mes y magnitud de la helada.

La Tabla 4 muestra el mes en que se produjo el fenómeno de la helada por primera y por última vez en cada año por cada estación meteorológica. Se observa que, para los nueve años analizados, en la mayoría de los años el período de heladas comienza en el mes de mayo o junio; mientras que la finalización del periodo se produce entre los meses de agosto, septiembre u octubre. Se ha considerado como heladas tardías aquellas producidas a partir del mes de agosto, época en que algunas especies, como el almendro, inician su floración.

Años	San Francisco		INTA	
	Mes primera helada	Mes última helada	Mes primera helada	Mes última helada
2013	Mayo	Septiembre	--	--
2014	Mayo	Septiembre	--	--
2015	Mayo	Septiembre	--	--
2016	Abril	Octubre	Junio	Septiembre
2017	Mayo	Octubre	Junio	Agosto
2018	Mayo	Octubre	Junio	Agosto
2019	--	--	Junio	Septiembre
2020	--	--	Mayo	Agosto
2021	--	--	Mayo	Agosto

Tabla 4 - Mes de ocurrencia de la primera y última helada por año para cada estación meteorológica

La cantidad de días en los que se producen heladas varía entre los años. De los años 2.016, 2.017 y 2.018 se disponen los datos de ambas estaciones, se ha observado que la estación ubicada en el Establecimiento San Francisco ha registrado mayor cantidad de días con heladas (Tabla 5) que la estación instalada en EEA INTA (Tabla 6). Esto era previsible debido a que San Francisco (Departamento Sarmiento) se encuentra ubicada aproximadamente 37 km hacia el sur de la EEA INTA (Departamento Pocito). Del conjunto de años analizados se destaca el año 2.018 con mayor cantidad de días con heladas registradas en las dos estaciones meteorológicas.

	2013	2014	2015	2016	2017	2018
Abril	0	0	0	2	0	0
Mayo	7	6	6	0	5	3
Junio	17	21	20	17	17	27
Julio	23	23	26	24	20	25
Agosto	22	13	4	6	13	23
Septiembre	11	2	6	5	4	3
Octubre	0	0	0	1	2	1
Total de días con helada	**80**	**65**	**62**	**55**	**61**	**82**
Total de días con heladas tardías	**33**	**15**	**10**	**12**	**19**	**27**

Tabla 5 - Cantidad de días con heladas registrados por la estación San Francisco

	2016	2017	2018	2019	2020	2021
Abril	1	0	0	0	0	0
Mayo	0	0	0	0	1	1
Junio	7	8	12	6	10	9
Julio	9	8	15	12	15	16
Agosto	0	0	11	11	10	2
Septiembre	1	0	0	2	0	0
Octubre	0	0	0	0	0	0
Total de días con helada	**18**	**16**	**38**	**31**	**36**	**28**
Total de días con heladas tardías	**1**	**2**	**11**	**13**	**10**	**2**

Tabla 6 - Cantidad de días con helada registrados por la estación INTA

La Figura 16 representa la temperatura para el periodo de heladas (desde la primera hasta la última) para el año 2.018 de la estación San Francisco. Donde se visualiza en primera instancia que la cantidad de valores correspondientes a no heladas es notablemente superior a la cantidad de heladas. Además, se observan heladas tardías en los meses de agosto, septiembre y octubre.

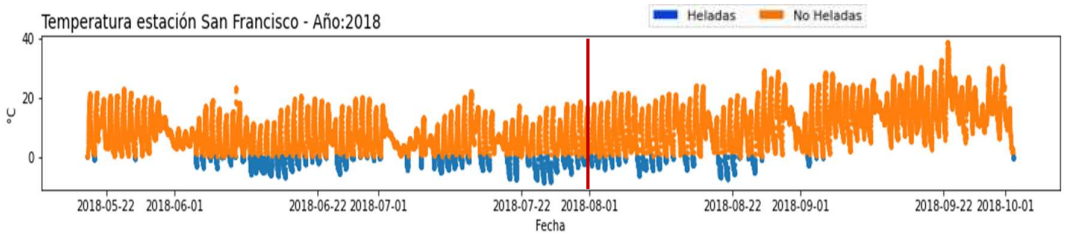


Figura 16 - Temperatura para el periodo de heladas del año 2018 en la estación San Francisco

Del conjunto de años considerados en cada estación, la cantidad total de días en los que se produjo el fenómeno de la helada es de 405 días para la estación San Francisco y 167 días en la estación INTA.

Del análisis de los registros efectuados por las estaciones meteorológicas, se ha determinado que de San Francisco se dispone un total de 299.671 registros, de los cuales 13.872 corresponden a temperatura mínima menor o igual a cero, es decir que aproximadamente el 4,63% del total de los registros corresponden a casos de helada.

Para la estación INTA, se cuenta con un total de 301.899 registros que incluyen 3.854 casos de helada, lo que significa que, respecto del total de registros, el 1,28% es de heladas. Estos valores están representados gráficamente en la Figura 17. Claramente

existe un desbalance entre la cantidad de registros o casos de heladas y la cantidad de registros de no heladas.

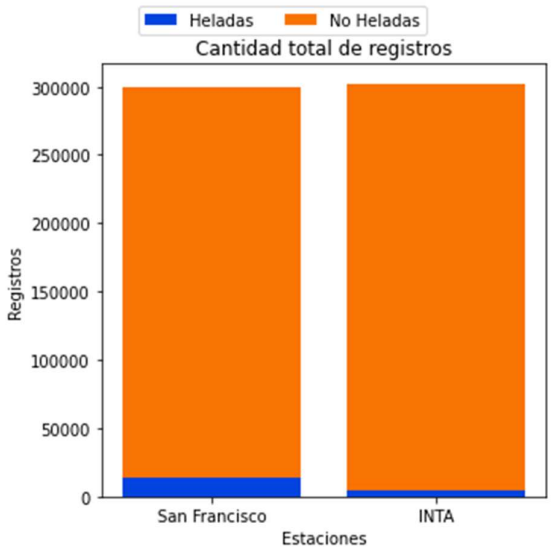


Figura 17- Cantidad de casos de heladas y no heladas para cada estación

Un detalle por año se presenta en la Tabla 7, con cantidad total de registros, cantidad de registros de helada, cantidad de registros de heladas tardías para la estación y cantidad de registros de heladas tardías para la estación San Francisco. Del mismo modo en la Tabla 8 para la estación INTA.

	2013	2014	2015	2016	2017	2018	Total
Cantidad total de registros	38.050	51.939	52.559	52.076	41.016	64.031	**299.671**
Cantidad de registros de helada	2.524	2.227	2.293	1.517	2.118	3.193	**13.872**
Cantidad de registros de helada tardía	936	393	187	234	472	933	**3.155**

Tabla 7 - Cantidad total de registros, cantidad de registros de heladas y de heladas tardías por año para la estación San Francisco

	2016	2017	2018	2019	2020	2021	Total
Cantidad total de registros	52.704	52.560	44.436	51.835	50.814	49.550	**301.899**
Cantidad de registros de helada	316	379	885	569	859	846	**3.854**
Cantidad de registros de helada tardía	28	38	186	209	229	27	**717**

Tabla 8 - Cantidad total de registros, cantidad *de registros de heladas y de registros de heladas tardías por año para la estación INTA*

La menor temperatura mínima registrada por cada estación en cada año procesado se presenta en la Tabla 9.

	2013	2014	2015	2016	2017	2018	2019	2020	2021
San Francisco	-9,6°	-6,0°	-5,7°	-5,0°	-9,0°	-8,6°	-	-	-
INTA	-	-	-	-2,4°	-4,9°	-4,5°	-3,2°	-4,3°	-5,1°

Tabla 9 – Menor temperatura mínima de cada estación para cada año.

Del análisis de los datos se evidencia que la zona donde se ubica la estación San Francisco es más fría, presenta mayor cantidad de casos de heladas, con intensidades que prácticamente duplican a las heladas de la estación INTA. Por tanto, los datos de la estación San Francisco son más convenientes para el entrenamiento del algoritmo de aprendizaje que los de la estación INTA.

5.3. Reducción del conjunto datos

La reducción de la dimensionalidad en los datos es un requisito previo para casi cualquier análisis que se desee realizar, una de las causas que la motivan es por preocupaciones computacionales. Un algoritmo de aprendizaje automático que opera con *d* características, en realidad procesa vectores de dimensión *d*. Si *d* es muy grande, el rendimiento de estos algoritmos comienza a fallar, y esta decadencia se entiende como un problema de la dimensionalidad. Normalmente es conveniente reducir los datos en un espacio de menor dimensión [23].

5.3.1. Cantidad de casos

El desbalance tan importante que presenta el conjunto de datos, puede afectar el rendimiento de cualquier modelo, por lo tanto, es necesario tratar esta problemática, principalmente porque el interés radica en la predicción de heladas (casos minoritarios).

Para balancear los datos, en caso de aplicar un método de sobremuestreo ocasionaría la generación de una gran cantidad de casos sintéticos. Si se aplicara submuestreo, se eliminaría una gran cantidad de casos de no heladas, existiendo la posibilidad de quitar del conjunto de datos casos probablemente importantes. Esto ha ocasionado que se plantee en primera instancia, una reducción de la cantidad de registros de no heladas considerando el período del año y el momento del día en que se producen las heladas.

De cada año se han seleccionado los registros o casos correspondientes al periodo comprendido entre el día anterior a la fecha de la primera helada hasta el día posterior de la fecha de la última helada. Además, las heladas son frecuentes entre las 0 horas y las 8 horas, por esto se han seleccionado solo los registros que se encuentran en el rango horario desde las 18 horas hasta las 9:50 horas. De este modo se reduce la cantidad de casos de no heladas, pero se mantiene la cantidad de casos de heladas.

Para la estación San Francisco, la cantidad de registros se redujo de 299.671 (Tabla 7) a 82.975 (Tabla 10), obteniendo 13.814 registros de heladas que corresponden al período diario comprendido entre las 18 horas y las 9:50 horas. Es decir, 58 registros menos que la cantidad indicada en la Tabla 7, esta pérdida se debe a que corresponden a registros de temperaturas negativas ocurridas a partir de las 10 horas. Estos registros son descartados porque a partir de esta hora el calor del sol contrarresta el impacto de la helada.

En el caso de los datos de la estación INTA, la reducción fue de 301.899 (Tabla 8) a 49.973 (Tabla 11) registros, obteniendo 3.830 registros de heladas, en este caso se perdieron 24 registros. La Figura 18 muestra la cantidad de heladas y de no heladas con el conjunto de datos reducido por el filtro aplicado.

	2013	2014	2015	2016	2017	2018	Total
Cantidad total de registros	13.315	11.440	13.368	17.302	14.298	13.252	**82.975**
Cantidad de registros de helada	2.517	2.222	2.291	1.515	2.108	3.161	**13.814**

Tabla 10 - Cantidad total de registros y de registros de heladas por año para estación San Francisco luego de filtrar

	2016	2017	2018	2019	2020	2021	Total
Cantidad total de registros	12.975	5.779	7.590	9.311	8.421	5.897	**49.973**
Cantidad de registros de helada	316	375	880	567	851	841	**3.830**

Tabla 11 - Cantidad de registros y de registros de heladas por año para estación INTA luego de filtrar

En este conjunto de datos filtrado, para la estación San Francisco, la cantidad de registros de heladas representa el 16,65% respecto de la cantidad total de registros (en el conjunto de datos original esta proporción era de 4,63%). En el caso de los datos de la estación INTA, la cantidad de heladas respecto de la cantidad de registros es del 7,66%

(anteriormente era de 1,28%). De este modo se alcanza una mejora en el balance del conjunto de datos.

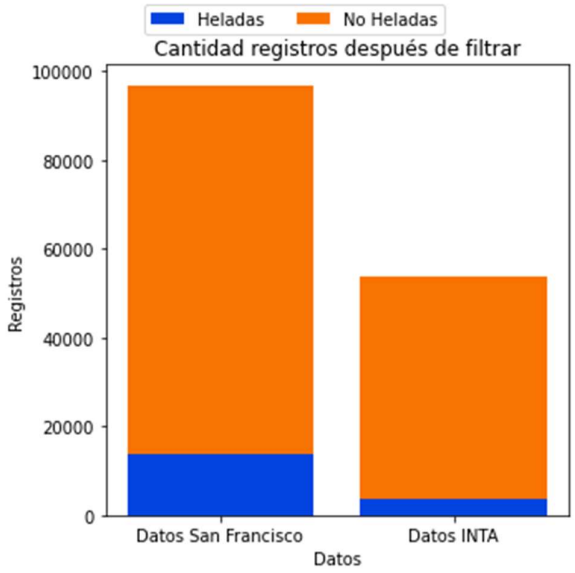


Figura 18 - Cantidad de registros de helada y no helada para cada estación luego de filtrar

Para estos conjuntos reducidos de datos la estadística descriptiva de la variable temperatura mínima (Tabla 12) muestra que el valor del primer cuartil es superior a cero, indicando que los datos de las heladas son menores del 25%, característica que también se presentaba considerando el conjunto de todos los datos (conjuntos originales).

La mejora está dada por la disminución del valor del primer cuartil. Para la estación San Francisco, de un valor 9,30 (datos originales) se ha alcanzado el de 1,50 (datos reducidos). Y en la estación INTA, el valor de 11,70 obtenido con todos los datos originales, ha bajado a 3,10 con el conjunto reducido. Esto indica que con los datos reducidos ha aumentado la proporción de datos referidos a heladas respecto al total de datos (heladas y no heladas).

Por otro lado, en la comparación del valor del segundo cuartil (50%) para los datos filtrados de cada estación (Tabla 12) se observa una diferencia de 0,7. Y para el tercer cuartil, los valores coinciden. La diferencia en la media es poco significativa, mientras que la desviación estándar, para San Francisco está alrededor de un punto por encima de la otra estación. Esto muestra similitudes en los conjuntos de datos, lo cual es importante ya que uno de ellos se usa para entrenar el modelo y el otro para probarlo.

	San Francisco	INTA
Cantidad datos	**82.975**	**49.973**
Media	**6,31**	**7,01**
Desviación estándar	**6,45**	**5,32**
Valor mínimo	**-9,60**	**-5,10**
Valor máximo	**36,40**	**32,30**
25%	**1,50**	**3,10**
50%	**5,70**	**6,40**
75%	**10,30**	**10.30**

Tabla 12 – Descripción de la variable temperatura para cada estación.

La distribución de la variable temperatura mínima de cada estación meteorológica presenta asimetría, como se observa en la Figura 19. Mostrando la existencia de un sesgo hacia los valores positivo, producto del desbalance en los datos.

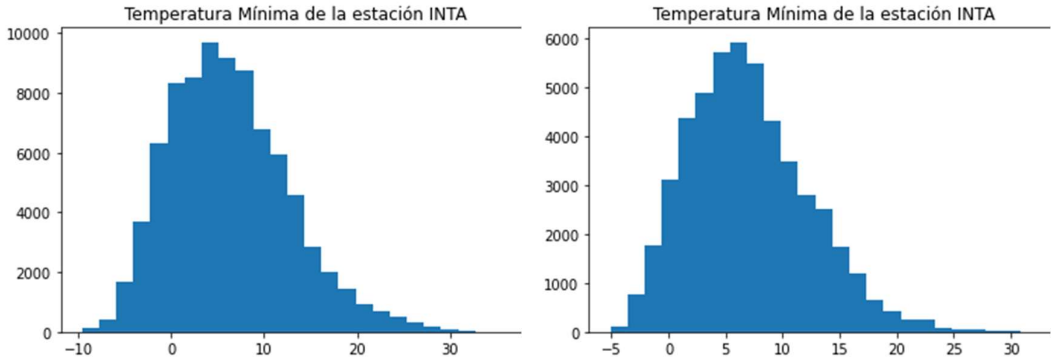


Figura 19 - Distribución de la variable temperatura mínima (a) Estación San Francisco (b) Estación INTA

5.4. Selección de variables

Las estaciones meteorológicas registran los valores de distintas variables, además de calcular algunos índices. En el análisis del fenómeno de la helada, que particularmente refiere a la temperatura, no todas las variables censadas tienen relación con ella, lo que requiere una selección de variables.

En una primera instancia, a partir del concepto de helada y de la consulta a un experto, se determina que las variables relacionadas a la helada podrían ser la temperatura, la humedad relativa, el punto de rocío, el viento, la presión atmosférica y la radiación solar. Luego, en una segunda instancia se ha aplicado una técnica matemática para determinar la correlación entre estas variables.

En segunda instancia se ha realizado la matriz de correlación lineal para las variables meteorológicas seleccionadas anteriormente, se presenta en la Figura 20. A partir de ella se han elegido las variables para los modelos. En el análisis visual de la matriz se ha observado que las variables más relacionadas con la temperatura es la humedad con un valor de -0,67, ante un aumento de la temperatura decrece la humedad y viceversa. En tanto entre el punto de rocío y la temperatura la correlación es directa y alcanza un valor de 0,63.

Las variables que se han seleccionado para el modelo predictivo son la temperatura, humedad relativa y punto de rocío. La radiación solar ha sido descartada debido a que arroja un valor 0,51; encontrándose muy cerca del umbral de aceptación. Lo que podría ocasionar que esta variable no haga un gran aporte a la capacidad predictiva del modelo, y perjudique su desempeño al incrementar la cantidad de datos a procesar si se la incluye. Además, en estudios previos sobre algoritmos de árboles aleatorios se ha constatado que la inclusión de esta variable entre los datos de entrada no aporta mejores resultados [25].

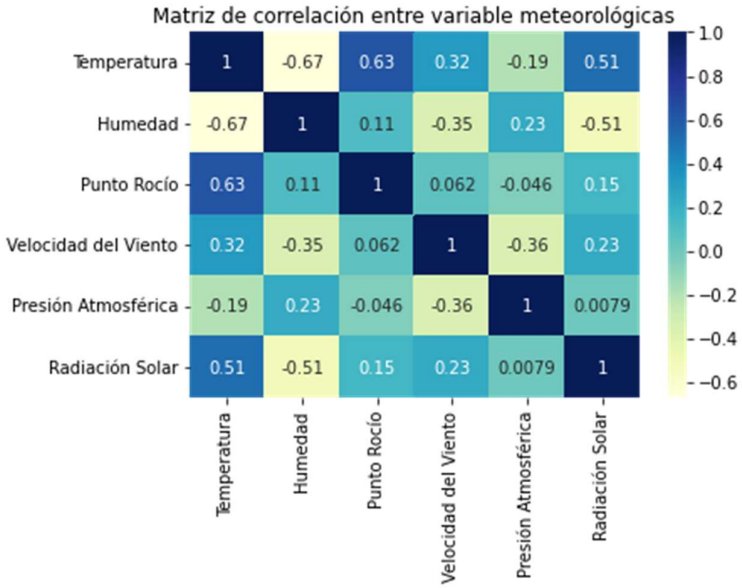


Figura 20 -.Matriz de Correlación Lineal entre las variables Temperatura mínima, Humedad, Punto de rocío, Velocidad del viento, Presión atmosférica y Radiación solar.

A partir de los datos del último año disponible de la estación San Francisco (2.018) se ha seleccionado el periodo de heladas y se ha representado las tres variables (temperatura, humedad y punto de rocío) de forma gráfica (Figura 21) para analizar visualmente los resultados de la matriz de correlación.

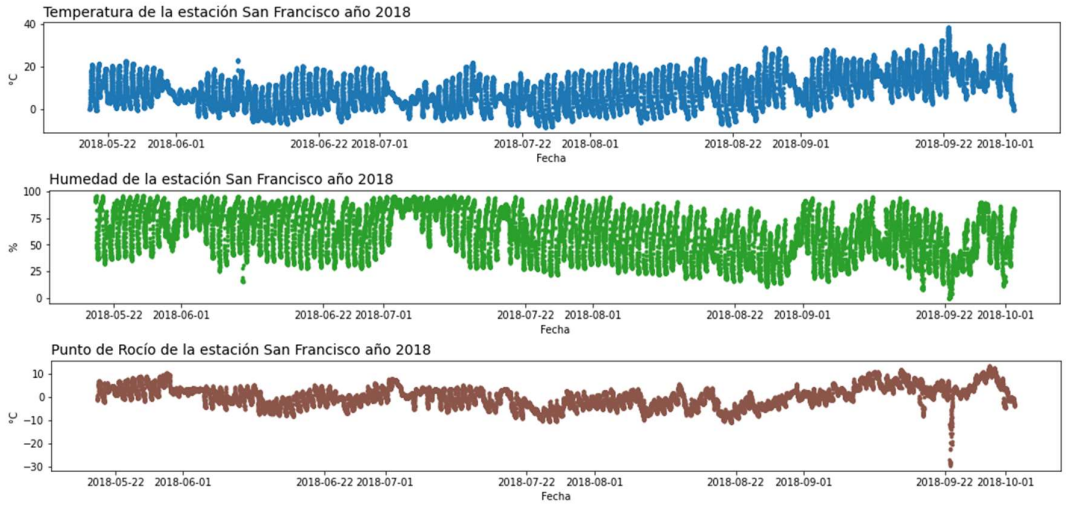


Figura 21 - Representación de las variables temperatura, humedad y punto de rocío del año 2.018 en la estación San Francisco

Desde una observación global de los gráficos, se puede apreciar claramente la correlación indirecta entre temperatura y humedad. Además de la correlación directa entre temperatura con el punto de rocío en los meses de junio, julio y agosto.

5.5. Preparación de los datos

Los datos disponibles son conjuntos de series temporales, lo que permite estructurar la entrada con una ventana móvil. Se posee evidencia de que algoritmos de aprendizaje automático como *Random Forest* arrojan buenos resultados con las lecturas de 3 horas previas a la ocurrencia del fenómeno [25].

Las estaciones toman lectura de los datos cada 10 minutos. Así, durante una hora se registran 6 lecturas para una variable. Entonces, la ventana de 3 horas contiene 18 valores para cada variable meteorológica considerada en el modelo.

Respecto al horizonte de la salida, se ha trabajado con 3 horas. Es decir, se pronostica la temperatura hacia la tercera hora posterior a la última hora de lectura de la temperatura incluida en la entrada. La determinación de este valor se basa en un aspecto operativo para el productor. La predicción de ocurrencia de helada con este período de anticipación, da al agricultor el tiempo suficiente para asistir a la zona de cultivo y desplegar las acciones de mitigación de daño de la helada.

A partir del conjunto de datos filtrado de acuerdo al horario de ocurrencia del fenómeno y el período de heladas de cada año, se ha generado el conjunto de datos

para el tamaño de ventana establecido. La cantidad de casos obtenidos se muestra en la Tabla 13.

	San Francisco	INTA
Cantidad total de casos	52.607	31.613
Cantidad de casos de heladas	12.681	3.760

Tabla 13 - Cantidad total de casos y cantidad de casos de heladas por cada estación estructurados en una ventana de 3 horas.

En la Figura 22 se representa la cantidad de casos descriptos anteriormente, donde se visualiza con claridad que los casos de no heladas son superiores a los casos de heladas.

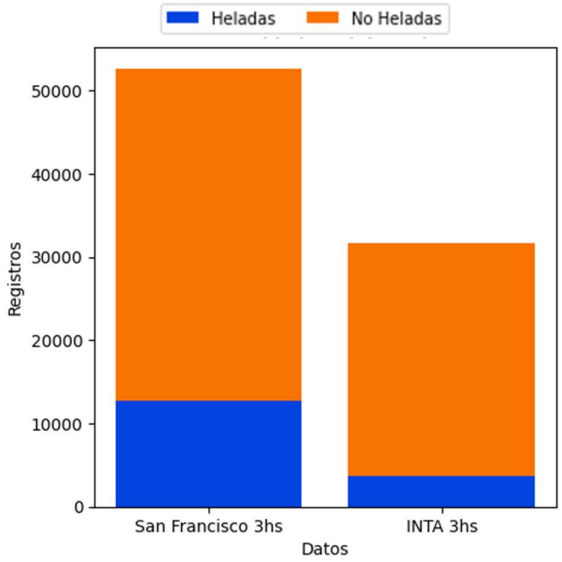


Figura 22-- Cantidad de casos de heladas y no heladas para cada estación para la ventana de 3 horas.

Dado que la magnitud del fenómeno y la cantidad de casos de heladas es superior para la estación San Francisco, se ha decidido entrenar los algoritmos de aprendizaje automático con los casos de esta estación. Mientras que los datos de la estación INTA se han usado para la validación de los modelos.

Capítulo 6

Modelado y Resultados

6. Modelado

El problema de la predicción de la temperatura, se ha tratado con modelos de regresión. Como se mencionó anteriormente, los datos tienen la característica de haber sido registrados a intervalos regulares de 10 minutos (series temporales), lo que permite proponer una solución a través de redes neuronales recurrentes de tipo LSTM [27].

Determinar la arquitectura de una red neuronal es una tarea de modelado difícil. Es necesario establecer parámetros (también llamados hiperparámetros) como el número de neuronas en las capas de entrada y salida, el número de capas ocultas y el número de neuronas en estas capas, criterios de parada para el entrenamiento, algoritmo de aprendizaje de optimización, funciones de activación en las capas ocultas y de salida; entre otros.

Ante la falta de estrategias para determinar los hiperparámetros de una red neuronal, se ha planteado una secuencia de experimentaciones de análisis con el fin de encontrar la arquitectura de red simple que arroje resultados aceptables. La Figura 23 presenta el flujo de experimentaciones llevadas a cabo.

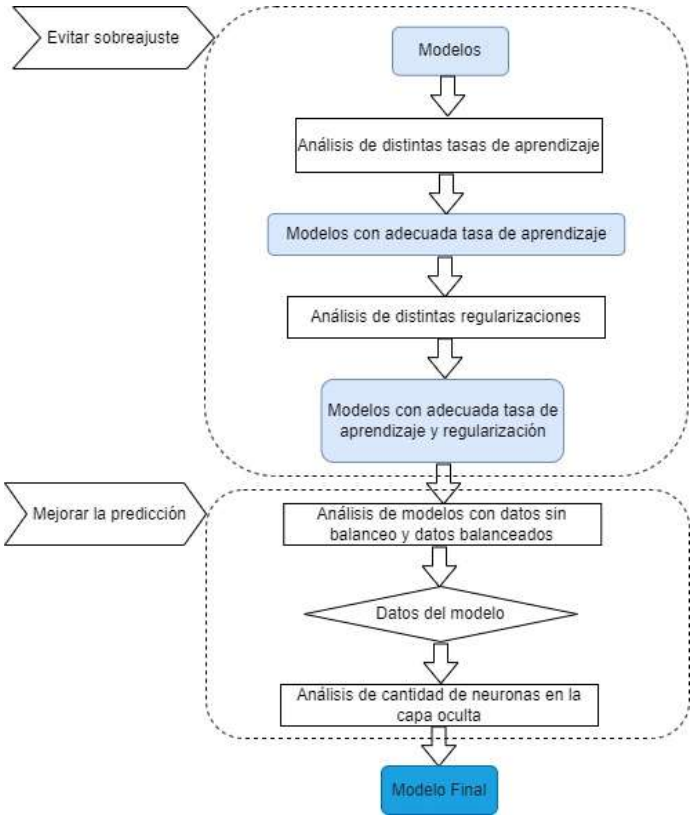


Figura 23 -Flujo de análisis en la búsqueda del modelo

Los hiperparámetros que se han analizado son la tasa de aprendizaje y la regularización para evitar sobreajuste. Luego, para una mejora en las predicciones, se compara el rendimiento de cada modelo con los datos originales (desbalanceados) contra los resultados obtenidos con los datos balanceados. Finalmente se ha analizado la cantidad de neuronas de la capa oculta.

6.1. Entrada y Salida

El número de neuronas en la capa de entrada y de salida está determinado por la naturaleza del problema. Como se indicó con anterioridad, para la entrada se ha trabajado con la técnica de ventana deslizante, la cual requiere determinar el tamaño de la misma. Para esto, además del estudio previo con el algoritmo de clasificación *Random Forest* [25], se han desarrollado modelos LSTM univariados (Figura 24) que procesan ventanas de 2, 3 y 4 horas (disponibles en el repositorio GitHub) [40] y se ha analizado el rendimiento en cada caso (Tabla 14). La capa de salida consta de una neurona que arroja un valor numérico representativo del pronóstico de la temperatura hacia un horizonte de 3 horas.

La arquitectura de red neuronal recurrente (Figura 24) usada consta de una capa oculta con función de activación RELU, función de pérdida (loss) el MSE y algoritmo de optimización Adam, con tasa de aprendizaje 1e-6. La capa de salida posee una neurona y tiene función de activación lineal.

```
def modelo(n_time_step, n_features, neuron, lr):
  model = Sequential()
  model.add(LSTM(neuron, activation= "relu",input_shape=(n_time_step, n_features)))
  model.add(Dense(1))
  model.compile(loss='mse',optimizer= tf.keras.optimizers.Adam(learning_rate=lr), metrics=['mse'])
  model.summary()
  return model
```

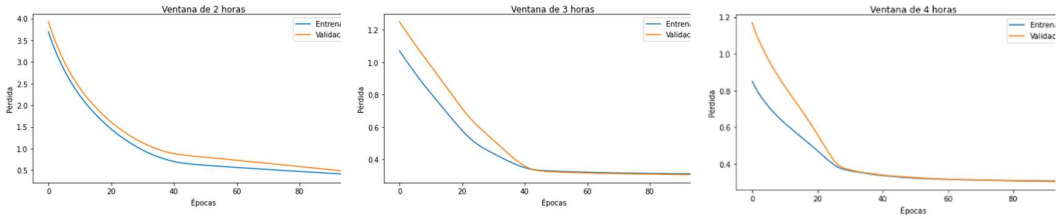
Figura 24 - Red neuronal LSTM univariada

El entrenamiento se ha realizado para 100 épocas con los datos de la variable temperatura de la estación San Francisco. La prueba se ha ejecutado con los datos de la estación INTA. Los resultados obtenidos se presentan en la Tabla 14.

	2 horas	3 horas	4 horas
Pérdida en entrenamiento	10,78	8,53	8,41
Pérdida en validación	12,30	8,39	8,34
RMSE entrenamiento	3,28	2,92	2,90
RMSE validación	3,50	2,90	2,89
RMSE test	2,82	2,66	2,74
R^2 entrenamiento	0,64	0,71	0,72
R^2 validación	0,64	0,75	0,76
R^2 test	0,60	0,64	0,62
Recall (heladas)	0,02	0,33	0,32
F1-score (heladas)	0,04	0,46	0,44

Tabla 14 - Análisis de resultados obtenidos con distintos tamaños de ventana deslizante

En la observación de la curva de aprendizaje, para los modelos con ventana de2 y 3 horas la curva es mejor que la de 4 horas donde se aprecia una caída abrupta de la curva de validación.



(a) *(b)* *(c)*

Figura 25 - Curvas de aprendizaje obtenidas con modelos univariados para distintas ventanas deslizantes. (a) ventana de 2 horas, (b) ventana de 3 horas, (c) ventana de 4 horas.

La representación gráfica de la predicción del modelo de 2 horas (Figura 26 (a)) muestra claramente que los valores predichos se alejan de los reales, especialmente en los valores de temperatura negativa (heladas). Esta situación es notoriamente distinta, para los modelos de 3 (Figura 26 (b)) y 4 horas (Figura 26 (c)), donde los valores predichos están más próximos a los reales.

(a)

(b)

(c)

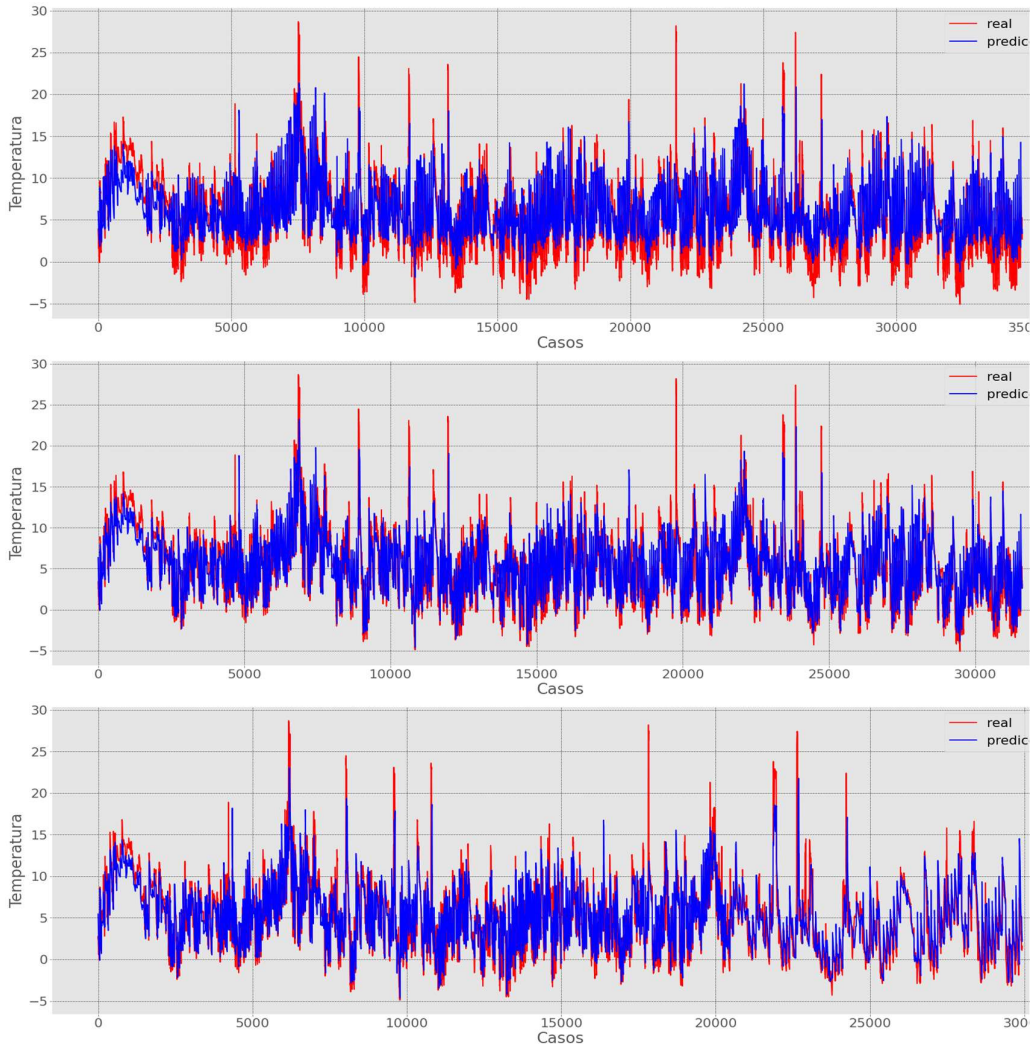


Figura 26 - Representación de la predicción de cada modelo. (a) con ventana de 2 horas (b) ventana de 3 horas (c) ventana de 4 horas

Observando las métricas (loss, RMSE, R^2, recall y F1-Score), el modelo con ventana de 2 horas tiene un rendimiento notablemente menor que los otros dos casos, principalmente en la función de pérdida, como así también en recall y F1-Score (Tabla 14). Esto se constata al observar la gráfica de las predicciones de este modelo. Los valores para el modelo con la ventana de 3 horas y el de 4 horas no arrojan diferencias notorias, tampoco la representación gráfica de la predicción.

Del análisis de las distintas métricas (Tabla 14), las curvas de aprendizaje (Figura 25Figura 1) y la representación gráfica de la predicción (Figura 26); sumado al estudio previo realizado con el algoritmo de clasificación *Random Forest* [25], se selecciona el modelo de 3 horas por procesar menor cantidad de datos y presentar mejor curva de aprendizaje.

Así la entrada queda constituida por una ventana de 18 pasos o marcas de tiempo, lo que representa 3 horas para cada variable que se ha incluido (Figura 27). Y el modelo entrega un valor numérico representativo del pronóstico de la temperatura hacia un horizonte de 3 horas.

Figura 27 – Ventana de entrada de 18 valores de la variable meteorológica y salida con horizonte de 3 horas.

En la Tabla 15 se presenta la cantidad de casos de cada conjunto de datos discriminando cantidad de casos de heladas y no heladas.

Conjunto	Total de casos	Casos de heladas	Casos de no heladas
Entrenamiento	44.203	9.838	34.365
Validación	16.808	5.686	11.122
Test	31.613	3.760	27.853

Tabla 15 – Cantidad de casos para cada conjunto de datos discriminado casos de heladas y no heladas.

Los datos de San Francisco desde el año 2.013 al año 2.017 se destinan al entrenamiento, el año disponible restante (2.018) para validación. Resultando 44.203 casos para entrenamiento, que representan el 72,45% del total de casos. Para validación son 16.808 casos, equivalente a un 27,55% del conjunto total. Estos valores se representan en la Figura 28.

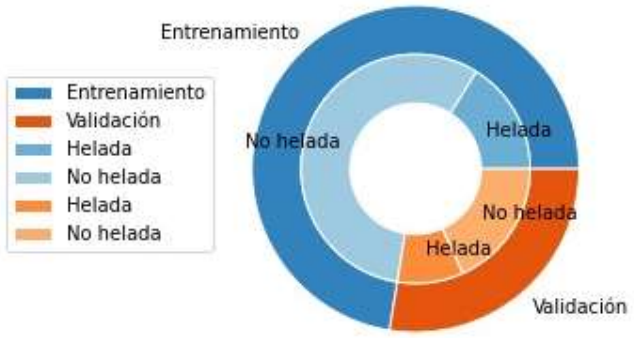


Figura 28 - Separación de los datos de la Estación San Francisco en conjunto de entrenamiento y de validación.

El conjunto de test se conforma de 31.613 casos de la estación INTA, de los cuales 3.760 corresponden a heladas (11,89%).

6.2. Funciones

La función de activación usada para las capas ocultas es ReLU, por la facilidad que presenta esta función para el entrenamiento de las redes neuronales [27]. De acuerdo a lo planteado en [26], para la función de pérdida (*loss*) se ha utilizado el error cuadrático medio (MSE) por tratarse de un problema de regresión. La capa de salida consta de una neurona con función de activación lineal.

6.3. Entrenamiento

El entrenamiento se realiza para distinta cantidad de épocas (epochs). Los resultados se analizan a través del valor del MSE, RMSE y el coeficiente R^2. Luego, cada valor predicho para los casos del conjunto de prueba, se etiqueta con 0 o 1, según corresponda a una temperatura de helada o no helada respectivamente. Con esto se construye la matriz de correlación y se observa el valor recall y el F1-Score para los casos de heladas.

6.4. Capas ocultas

Se ha iniciado en análisis con redes que poseen una capa oculta. Para esto se ha tenido en cuenta lo planteado por Géron, A. [30], quien indica que, para muchos problemas se puede iniciar con una o dos capas ocultas y se obtendrán buenos resultados. Sólo en caso de problemas muy complejos (clasificación de imágenes grandes o el reconocimiento de voz), se puede aumentar gradualmente la cantidad de capas ocultas, hasta que comience a sobreajustar el conjunto de entrenamiento.

6.5. Optimizador

El algoritmo de optimización usado es Adam por presentar buen desempeño lo que hace que sea muy usado [27]. En este algoritmo participa la tasa de aprendizaje y la regularización, para determinar los valores de estos hiperparámetros, se ha llevado a cabo un análisis empírico con 3 modelos de referencia (Figura 29): a) un modelo univariado que recibe como entrada 18 valores de la variable temperatura, b) un modelo bivariado cuya entrada son 18 valores de las variables temperatura y humedad relativa (36 valores en total), y c) un modelo trivariado cuyos datos de entrada son 54 valores que corresponden a las variables meteorológicas temperatura, humedad relativa y punto de rocío (18 valores de cada variable).

Para la selección de las variables de cada modelo se ha tenido en cuenta la correlación con la temperatura (Figura 20), eligiendo las variables más correlacionadas.

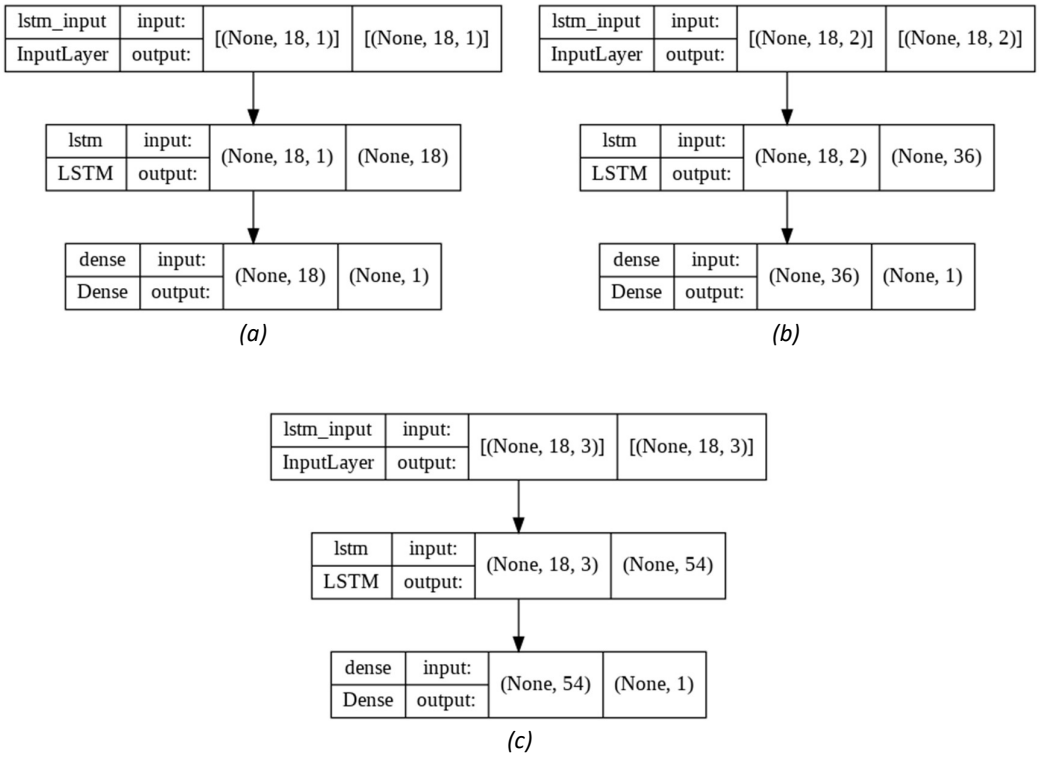


Figura 29 - Arquitectura de los modelos de referencia. (a) Modelo Univariado (b) Modelo bivariado (c) Modelo Trivariado

6.6. Tasa de aprendizaje

La tasa de aprendizaje interviene directamente en el rendimiento del modelo. El valor apropiado para este hiperparámetro depende de los datos y del modelo. Se ha utilizado diferentes tasas de aprendizaje en modelos que reciben como entrada distintas variables meteorológicas. En todos los casos, los modelos de referencia tienen una capa oculta con cantidad de neuronas igual a la cantidad de entradas. Es decir, 18 neuronas para el modelo univariado, 36 para el bivariado y 54 para el trivariado.

Los tres modelos se han entrenado con los siguientes valores de tasa de aprendizaje: 1e-3, 1e-4, 1e-5, 1e-6 y 1e-7 (disponibles en el repositorio GitHub) [40]. Se inicia en 1e-3 debido a que este es el valor por defecto para el optimizador Adam en la librería TensorFlow de Python. Los resultados obtenidos en la función de pérdida en entrenamiento y en validación se detallan en la Tabla 16.

		1e-3	1e-4	1e-5	1e-6	1e-7
Univariado	Pérdida en entrenamiento	0.267	0.275	0.289	0.309	0.822
	Pérdida en validación	0.276	0.287	0.278	0.302	1.032
Bivariado	Pérdida en entrenamiento	0.212	0.259	0.276	0.315	--
	Pérdida en validación	0.318	0.268	0.272	0.326	--
Trivariado	Pérdida en entrenamiento	0.178	0.244	0.269	0.291	--
	Pérdida en validación	0.178	0.267	0.270	0.285	--

Tabla 16- Pérdida de los modelos con distintas tasas de aprendizaje

La Figura 30 representa las curvas de aprendizaje para el modelo univariado con distintas tasas de aprendizaje, la Figura 31 las del modelo bivariado y la Figura 33 las del modelo trivariado.

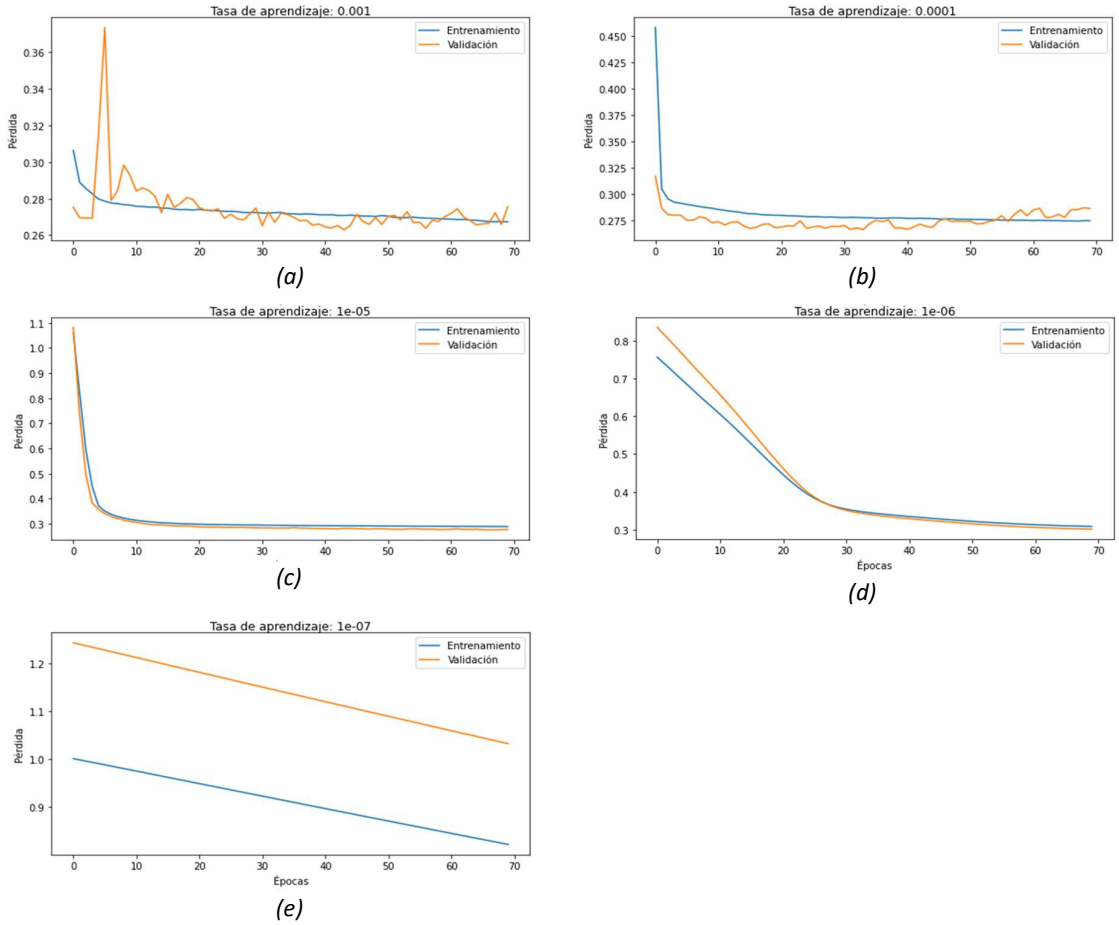


Figura 30 - Curva de aprendizaje del modelo univariado variando la tasa de aprendizaje. (a) Tasa de aprendizaje 1e-3 (b) Tasa de aprendizaje 1e-4 (c) Tasa de aprendizaje 1e-5 (d) Tasa de aprendizaje 1e-6 (e) Tasa de aprendizaje 1e-7

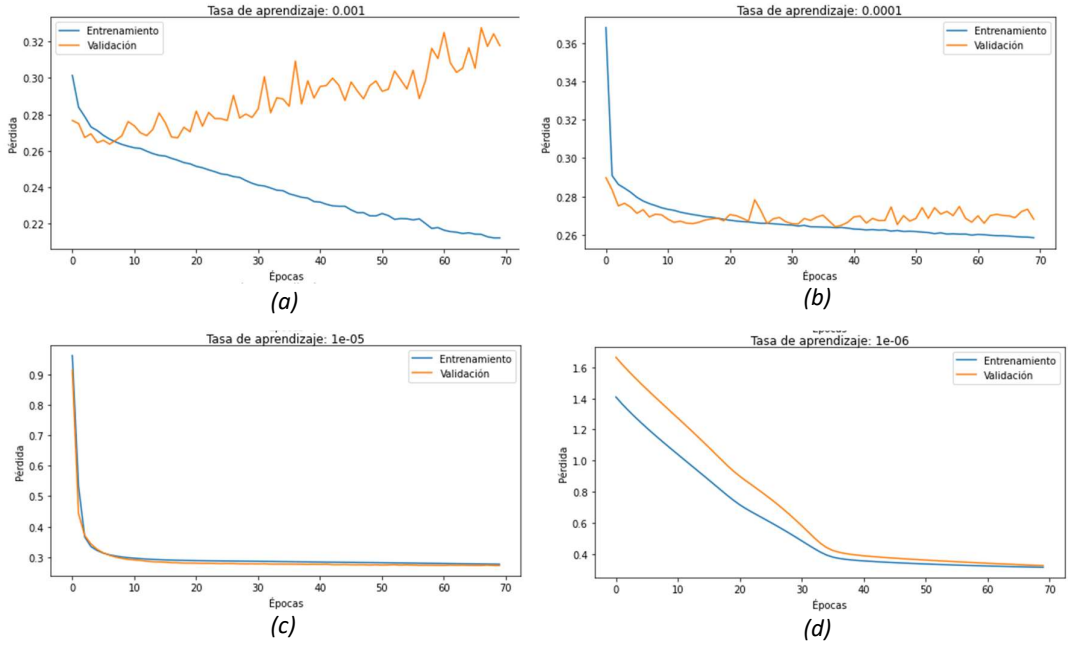


Figura 31- Curva de aprendizaje del modelo bivariado variando la tasa de aprendizaje. (a) Tasa de aprendizaje 1e-3 (b) Tasa de aprendizaje 1e-4 (c) Tasa de aprendizaje 1e-5 (d) Tasa de aprendizaje 1e-6

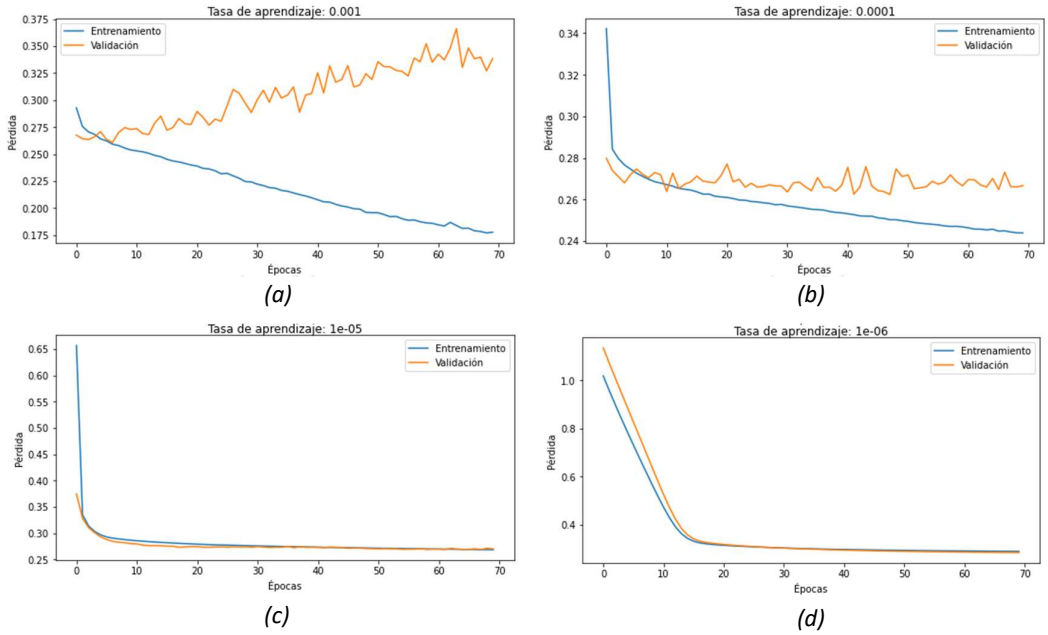


Figura 32 - Curvas de aprendizaje del modelo trivariado con distintas tasas de aprendizaje. (a) Tasa de aprendizaje 1e-3 (b) Tasa de aprendizaje 1e-4 (c) Tasa de aprendizaje 1e-5 (d) Tasa de aprendizaje 1e-6

Se ha observado en los tres modelos, que los valores de 1e-3 y 1e-4 en la tasa de aprendizaje generan una curva de aprendizaje con zigzag, lo que indica que estos valores son demasiado altos. El valor de 1e-7 sólo se analizó en el modelo univariado,

evidenciando que es un valor muy bajo (aprendizaje lento). El valor de 1e-5, si bien no presenta oscilaciones, cae abruptamente en las primeras épocas y luego se mantiene estable. Según lo expresado por Gerón, A. en [30], esta forma de la curva indica que el valor de la tasa es alto.

De los valores dados a la tasa de aprendizaje, el 1e-6 es el más adecuado para los tres modelos. En cuanto a los valores de la función de pérdida en entrenamiento y en validación, no hay diferencia significativa entre ellos.

Ante los resultados anteriores, se ha tratado de encontrar una mejor tasa de aprendizaje entrenando los modelos con una tasa menor a 1e-6, específicamente con el valor 8e-7. La Tabla 17 muestra los resultados obtenidos con ambas tasas los cuales no difieren significativamente entre sí, esto indica que en el conjunto analizado no hay un modelo que sea notablemente mejor.

Se puede observar que con el valor 8e-7 en la tasa de aprendizaje, sólo se obtiene una leve mejora en la pérdida del modelo bivariado. Siendo estos valores muy próximos a los obtenidos por el modelo trivariado con tasa 1e-6. El modelo univariado arroja mejores valores con la tasa 1e-6.

	Univariado		**Bivariado**		**Trivariado**	
	1e-6	**8e-7**	**1e-6**	**8e-7**	**1e-6**	**8e-7**
Pérdida en entrenamiento	0,309	0,333	0.315	0.302	0,290	0,295
Pérdida en validación	0,301	0,332	0,326	0,300	0,285	0,291
RMSE entrenamiento	2,917	3,030	2,946	2,885	2,830	2,849
RMSE test	2,633	2,696	2,603	2,558	2,528	2,549
R^2 entrenamiento	0,713	0,691	0,708	0,730	0,720	0,727
R^2 test	0,650	0,633	0,658	0,670	0,678	0,672
Recall para heladas	0,30	0,20	0,19	0,30	0,42	0,36
F1-score para heladas	0,44	0,32	0,31	0,43	0,54	0,49

Tabla 17 - Resultados de entrenar los modelos de referencia con la tasa 1e-6 y 8e-7

Además de los valores numéricos obtenidos, se han comparado las curvas de aprendizaje de cada modelo con cada tasa, éstas se observan en la Figura 33.

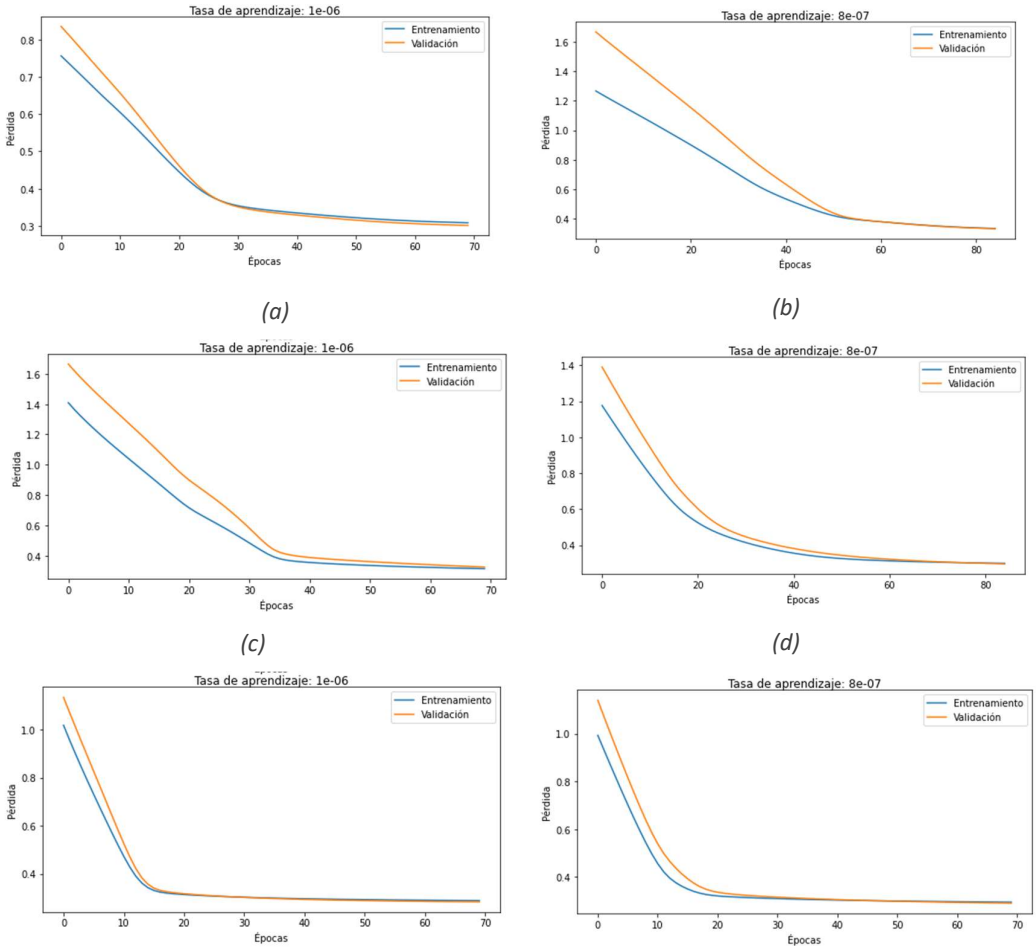


Figura 33 - Curvas de aprendizaje de los modelos (a) Univariado con tasa 1e-6 (b) Univariado con tasa 8e-7 (c) Bivariado con tasa 1e-6 (d) Bivariado con tasa 8e-7 (e) Trivariado con tasa 1e-6 (f) Trivariado con tasa 8e-7.

Claramente se aprecia una mejora en el aprendizaje del modelo bivariado con tasa 8e-7 (Figura 33 (d)) en comparación con la otra tasa (Figura 33 (c)). La curva del modelo univariado para el valor 8e-7 (Figura 33 (b)), se ha aplanado respecto a la curva del valor 1e-6 (Figura 33 (a)). Para el trivariado hay una leve mejora en la curva con tasa de aprendizaje de 8e-7 (Figura 33 (f)), esto se observa más claro en la Figura 34 que presenta las dos curvas de aprendizaje del modelo trivariado.

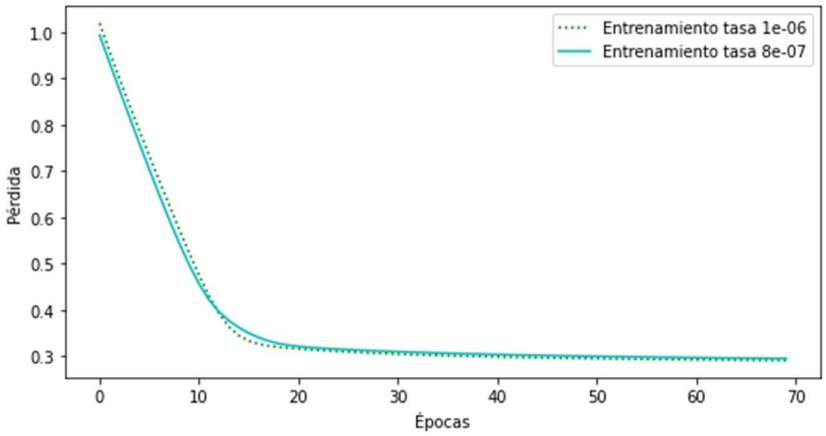


Figura 34- Comparación de las curvas de entrenamiento con tasa 1e-6 y 8e-7 del modelo trivariado,

El modelo trivariado con ambas tasas presenta los mejores valores de recall y F1-Score para las heladas predichas, esto es un aspecto importante ya que el interés está en la predicción del fenómeno. Por otro lado, los valores del R^2 para los datos de entrenamiento y los datos de test, son similares a los resultados obtenidos con el modelo bivariado con tasa 8e-7. Se debe considerar que la curva de aprendizaje del modelo bivariado es mejor porque no presenta sobreajuste, tampoco cae de rápidamente en las primeras épocas. Además, el entrenamiento es más ágil con el modelo bivariado por procesar menor cantidad de datos de entrada y tener menos neuronas en la capa oculta. Cabe aclarar que no se ha continuado analizando una tasa de aprendizaje menor para el modelo trivariado debido a que una tasa muy pequeña hace que el rendimiento de la red sea lento.

En síntesis, para continuar el análisis de otros hiperparámetros se ha adoptado la tasa 1e-6 para el modelo univariado, ya que una reducción en la tasa no ha presentado mejoras significativas. Para el bivariado la tasa será 8e-7 por las mejoras observadas en las métricas y curva de aprendizaje. Y finalmente, para el trivariado el valor de tasa de aprendizaje será 8e-7.

6.7. Regularización

La regularización es un hiperparámetro que interviene en el rendimiento de la red, siendo una estrategia para evitar el sobreajuste. Consiste en aplicar penalizaciones en los parámetros o en la actividad de la capa durante la optimización. Estas penalizaciones se suman a la función de pérdida que optimiza la red. En Python, a través de la librería

TensorFlow, a las capas LSTM se le puede aplicar penalización al núcleo, al *bias* y/o a la salida de la capa [41].

El modelo univariado con tasa 1e-6, el bivariado con tasa 8e-7 y el trivariado con tasa 8e-7 se han entrenado con cuatro variantes de regularización sobre el núcleo de la capa oculta (disponibles en el repositorio GitHub) [40]. Las variantes son: 1) L1=0 y L2=0; 2) L1= 0,01 L2=0; 3) L1=0 y L2=0,01 y 4) L1=0,01 y L2=0,01. Los resultados del modelo univariado se presentan en la Tabla 18.

	L1=0 L2=0	**L1=0,01 L2=0**	**L1=0 L2=0,01**	**L1=0,01 L2=0,01**
Pérdida en entrenamiento	0,306	0,387	0,326	0,405
Pérdida en validación	0,300	0,384	0,325	0,407
MSE entrenamiento	8,447	8,570	8,602	8,984
MSE validación	8,272	8,475	8,582	9,074
MSE test	7,054	7,169	7,100	7,147
RMSE entrenamiento	2,906	2,927	2,933	2,997
RMSE validación	2,876	2,911	2,929	3,012
RMSE test	2,656	2,678	2,665	2,673
R^2 en entrenamiento	0,716	0,711	0,710	0,698
R^2 validación	0,759	0,754	0,751	0,736
R^2 test	0,644	0,638	0,642	0,639
Recall (Helada)	0,37	0,29	0,27	0,16
F1-Score (Helada)	0,49	0,42	0,40	0,26
Recall (No helada)	0,98	0,99	0,99	0,99
F1-Score (No helada)	0,95	0,95	0,95	0,95

Tabla 18 – Resultados del entrenamiento del modelo univariado con tasa de aprendizaje de 1e-6 con distintos valores de regularización

La representación de curva de aprendizaje del modelo univariado variando los parámetros de regulación se presenta en la Figura 35.

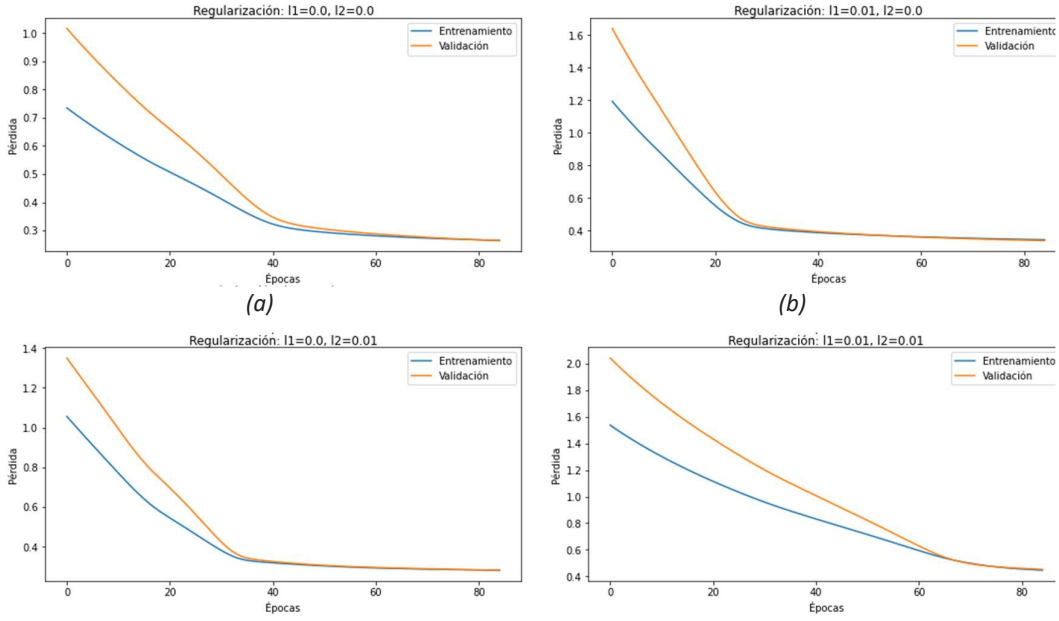


Figura 35- Curvas de aprendizaje del modelo unvariado con distintos valores en el parámetro de regularización. (a) l1 = 0 y l2 = 0, (b) l1 = 0,01 y l2 = 0, (c) l1 = 0 y l2 = 0,01, (d) l1 = 0,01 y l2 = 0,01

La Tabla 19 muestra los resultados obtenidos por el modelo bivariado con tasa de aprendizaje de 8e-7.

	L1=0 L2=0	**L1=0,01 L2=0**	**L1=0 L2=0,01**	**L1=0,01 L2=0,01**
Pérdida en entrenamiento	**0,296**	0,494	0,337	0,491
Pérdida en validación	**0,291**	0,493	0,331	0,488
MSE entrenamiento	8,149	8,185	8,416	8,222
MSE validación	8,015	8,171	8,276	8,166
MSE test	6,333	6,438	6,509	6,447
RMSE entrenamiento	2,855	2,861	2,901	2,867
RMSE validación	2,831	2,859	2,877	2,858
RMSE test	2,516	2,537	2,551	2,539
R^2 en entrenamiento	0,726	0,724	0,717	0,723
R^2 validación	0,767	0,762	0,759	0,763
R^2 test	0,681	0,675	0,672	0,675
Recall (Helada)	0,31	0,32	0,41	0,33
F1-Score (Helada)	0,44	0,45	0,53	0,46
Recall (No helada)	0,99	0,99	0,99	0,98
F1-Score (No helada)	0,95	0,95	0,95	0,95

Tabla 19 - Resultados del entrenamiento del modelo bivariado con distintos valores de regularización

La curva de aprendizaje de cada variante de regularización con la que se entrenó el modelo bivariado se presenta en la Figura 36 .

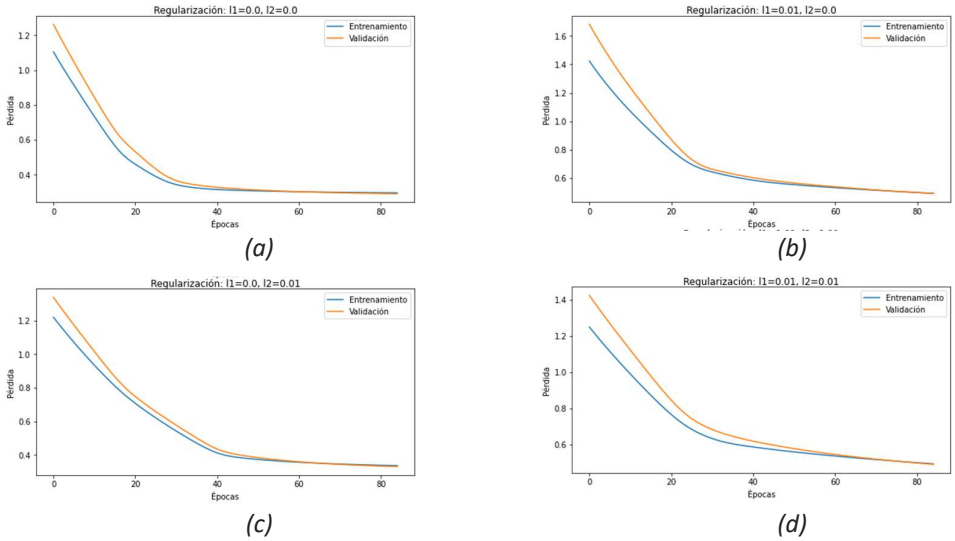


Figura 36 - Curvas de aprendizaje del modelo bivariado variando los parámetros de regulación. (a) l1 = 0 y l2 = 0, (b) l1 = 0,01 y l2 = 0, (c) l1 = 0 y l2 = 0,01, (d) l1 = 0,01 y l2 = 0,01

El modelo trivariado con tasa de aprendizaje de 1e-8 ha arrojado los resultados que se presentan en la Tabla 20.

	L1=0 L2=0	**L1=0,01 L2=0**	**L1=0 L2=0,01**	**L1=0,01 L2=0,01**
Pérdida en entrenamiento	0,292	0,530	0,340	0,556
Pérdida en validación	0,287	0,524	0,339	0,552
MSE entrenamiento	8,056	8,044	8,158	8,013
MSE validación	7,912	7,904	8,153	7,927
MSE test	6,373	6,366	6,464	6,324
RMSE entrenamiento	2,838	2,836	2,856	2,831
RMSE validación	2,813	2,811	2,855	2,815
RMSE test	2,524	2,523	2,542	2,515
R^2 en entrenamiento	0,729	0,732	0,725	0,730
R^2 validación	0,770	0,775	0,763	0,770
R^2 test	0,679	0,681	0,674	0,681
Recall (Helada)	0,40	0,40	0,40	0,39
F1-Score (Helada)	0,52	0,53	0,52	0,51
Recall (No helada)	0,98	0,98	0,98	0,98
F1-Score (No helada)	0,95	0,95	0,95	0,95

Tabla 20 - Resultados del entrenamiento del modelo trivariado con distintos valores de regularización

La Figura 37 muestra la representación de curva de aprendizaje del modelo trivariado variando los parámetros de regulación.

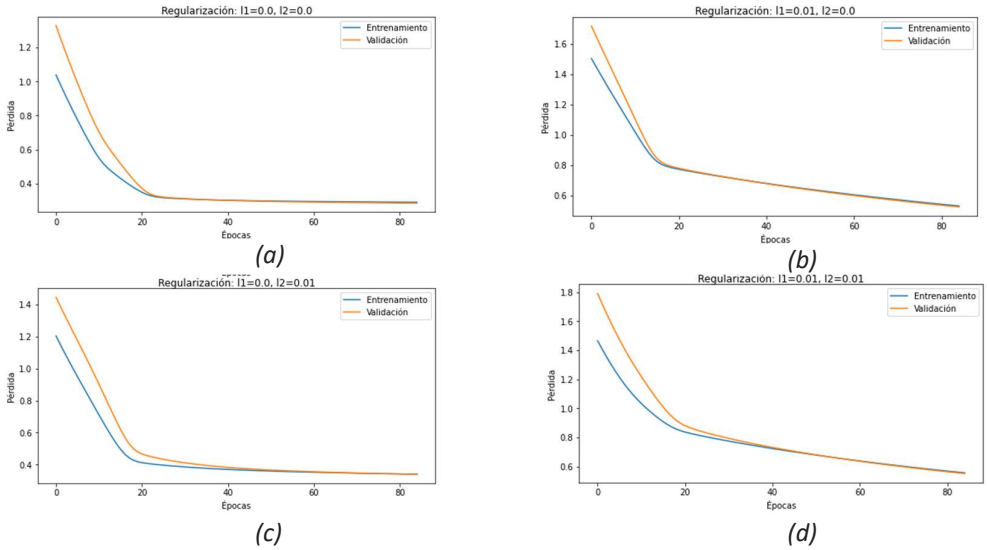


Figura 37 - Curvas de aprendizaje del modelo trivariado variando los parámetros de regulación. (a) l1 = 0 y l2 = 0, (b) l1 = 0,01 y l2 = 0, (c) l1 = 0 y l2 = 0,01, (d) l1 = 0,01 y l2 = 0,01

Al observar comparativamente los resultados obtenidos por cada modelo con distintos valores de regularización (Tabla 18, Tabla 19 y Tabla 20), se puede observar que ninguno de los tres modelos ha mejorado el desempeño en cuanto a valores de pérdida en entrenamiento y en validación, MSE, RMSE, R^2, recall y F1-Score. En cuanto a las curvas de aprendizaje no se observan mejoras en los casos en que el modelo fue entrenado con una regularización distinta de cero.

Por lo analizado con la tasa de aprendizaje y la regularización, con el modelo univariado se ha obtenido 0,64 para el coeficiente R^2 con los datos de test. Valor superado por 0,68 arrojado por el modelo bivariado y el trivariado con tasa de 8e-7 sin regularización. Por tanto, se ha continuado analizando estos dos modelos.

6.8. Balanceo de datos

Los datos procesados en los modelos analizados anteriormente presentan un claro desbalanceo (Tabla 13), superando la cantidad de casos de no heladas notablemente a los casos de heladas, siendo estos últimos en los cuales se centra el interés de esta predicción.

Las técnicas de remuestreo son muy usadas en algoritmos de clasificación cuando el conjunto de casos de dos o más clases se encuentra desbalanceados. En este caso, si bien se está trabajando con redes neuronales para un problema de regresión, se analiza mejorar la predicción balanceando los datos.

En esta experimentación, el conjunto de datos de entrenamiento ha sido balanceado con la técnica de remuestreo SMOTEEN. En un estudio previo [25] se ha concluido que esta técnica arroja buenos resultados al aplicarla sobre estos datos para ser usados en el algoritmo de clasificación Random Forest.

Luego del remuestreo de los datos, la cantidad de casos que conforman cada conjunto se detalla en la Tabla 21. El conjunto de entrenamiento original posee 44.203 registros, posterior al sobremuestreo queda conformado por 65.581 registros. Para el conjunto de validación se disponen de 16.808 registros que al ser sobremuestreados ascendieron a 21.812 registros.

	Entrenamiento	**Validación**
Casos de heladas	34.000	11.058
Casos de no heladas	31.643	10.810
Total de casos	65.643	21.868

Tabla 21 - Cantidad de casos en los conjuntos remuestreados

De forma gráfica (Figura 38) se aprecia claramente como el conjunto de datos se encuentra balanceado respecto a casos de heladas y no heladas.

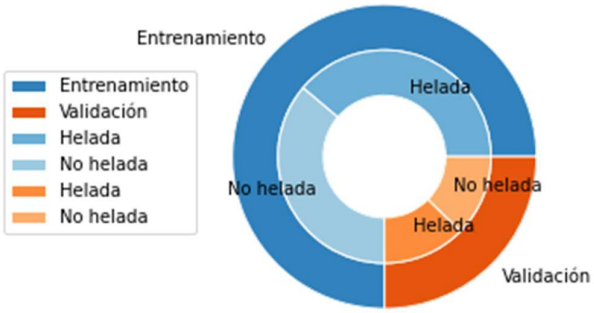


Figura 38 - Composición de los conjuntos de datos remuestreados

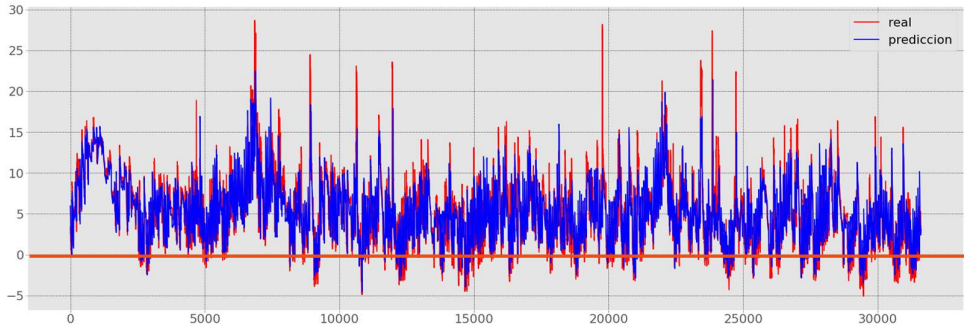
El modelo bivariado (con tasa de aprendizaje 8e-7 sin regulación) ha mostrado valores similares de RMSE y R^2 en el test con el entrenamiento con datos sin remuestreo y con los datos sobremuestreados (Tabla 22). Pero, analizado como clasificación, se observa que el pronóstico de casos de heladas es notablemente mejor con los datos sobremuestreados. El modelo trivariado (con tasa 8e-7 sin regulación) ha presentado la misma situación que el bivariado descrito anteriormente (modelos disponibles en el repositorio GitHub) [40].

	Bivariado		Trivariado	
	Sin remuestreo	**Con remuestreo**	**Sin remuestreo**	**Con remuestreo**
R^2 entrenamiento	0,73	0,80	0,73	0,81
R^2 test	0,68	0,66	0.68	0.67
RMSE entrenamiento	2,86	2,52	2,84	2,48
RMSE test	2,52	2,60	2,52	2,58
Recall (Helada)	0.31	**0.73**	0.40	**0.74**
F1-Score (Helada)	0.44	**0.69**	0.52	**0.69**
Recall (No helada)	0.99	0.95	0.98	0.94
F1-Score (No helada)	0.95	0.96	0.94	0.95

Tabla 22 - Comparación de métricas entre el modelo bivariado y trivariado sin remuestreo y con remuestreo

En la Figura 39 es posible ver gráficamente los resultados del modelo bivariado con cada conjunto de datos, apreciando que, para el conjunto sin remuestreo (Figura 39 (a)) varios casos reales de heladas han sido predichos con temperatura bajo cero, pero mayor a la real, o lo que es peor el pronóstico fue con temperatura positiva. En cambio, con los datos sobremuestreados (Figura 39 (b)) se visualiza mayor cantidad de casos de heladas pronosticados.

(a)



(b)

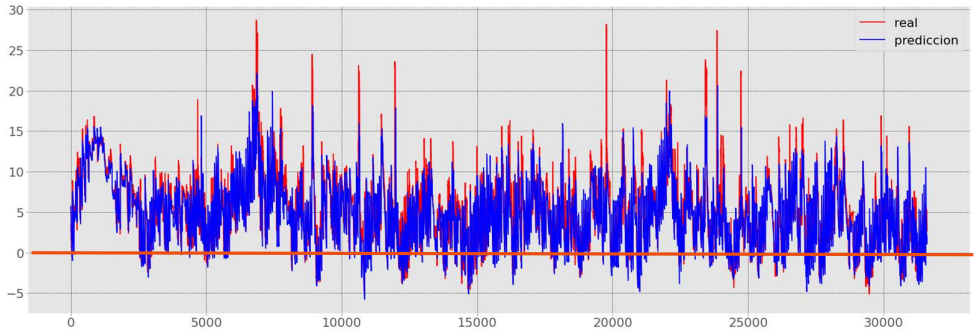


Figura 39- Predicción de temperatura del modelo bivariado. (a) Sin remuestreo (b) Con remuestreo

De forma gráfica la Figura 40 representa los resultados del modelo trivariado con cada conjunto de datos sin remuestreo (Figura 40 (a)) y remuestreados (Figura 40 (b)).

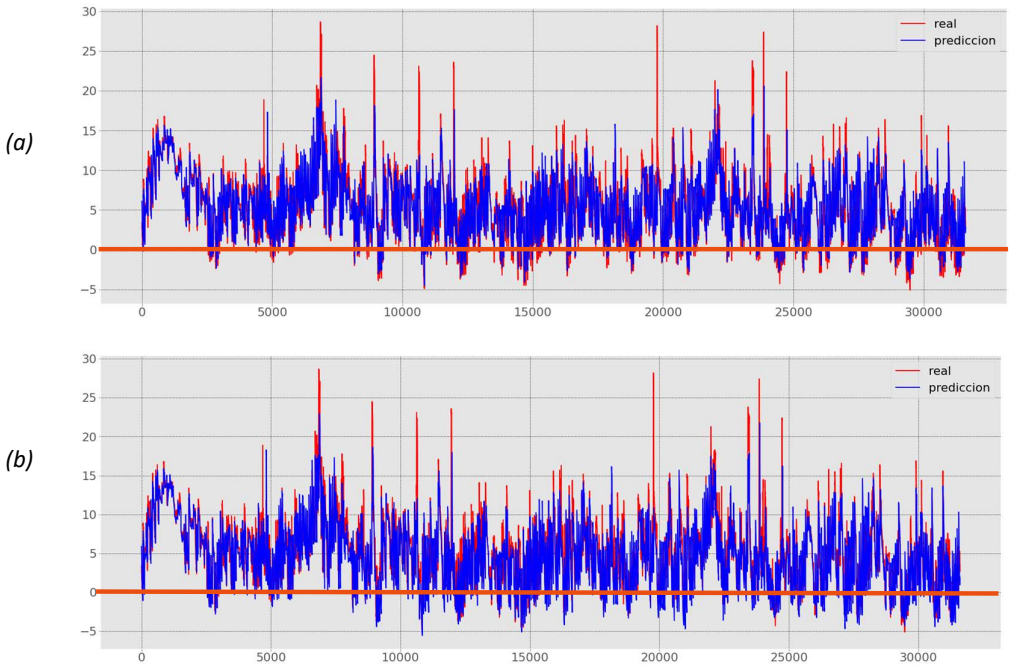


Figura 40 - Predicción de temperatura del modelo trivariado. (a) Sin remuestreo (b) Con remuestreo

Los valores arrojados por la matriz de confusión de los modelos que fueron entrenados con los datos remuestreados se presentan en la Tabla 23. Donde se puede observar la cantidad de casos de heladas acertados y no acertados, como así también para los casos de no heladas.

		Predicción Modelo bivariado		**Predicción Modelo trivariado**	
		Helada	**No Helada**	**Helada**	**No Helada**
Real	**Helada**	2.743	1.017	2.789	971
	No helada	1.484	26.369	1.565	26.288

Tabla 23 - Resultado de la matriz de confusión de los modelos entrenados con datos remuestreados

De los 2.743 casos de heladas predichos por el modelo bivariado, aproximadamente el 41% (específicamente 1.137) fueron con una temperatura menor a la real (sobre magnificación de la helada). Es decir que poco más de la mitad de las heladas fueron predichas con temperatura superior a la real. Para el modelo trivariado el porcentaje de heladas predichas con una temperatura menor a la real es de 51% aproximadamente respecto del total de heladas predichas.

El modelo trivariado predice mayor cantidad de temperaturas negativas menores a las reales. Esta situación, llevaría a que el productor planifique medidas de mitigación de la helada para una magnitud superior a la real, lo que puede terminar en un derroche de recursos. En los casos contrarios, donde se pronostica una temperatura menor a la real, se corre el riesgo de que sean escasos los recursos destinados a reducir los efectos de la helada, lo que termina siendo dañino para el cultivo, siendo esto último lo que se desea evitar.

6.8.1. Heladas tardías

Debido a que estos modelos alcanzan a predecir mayor cantidad de heladas, se ha realizado el análisis particular de las heladas tardías, aquellas que suceden a partir del mes de agosto y, dependiendo del cultivo, pueden causar severos daños.

El conjunto de datos de entrenamiento original 2.127 casos de heladas tardías, mientras que el conjunto de validación posee 1.732 casos. Cabe aclarar que estos conjuntos de datos fueron remuestreados, y en los conjuntos resultantes no es posible identificar casos de heladas tardías ya que el dato de fecha y hora no estuvo incluido en la operación de remuestreo. Los datos de testeo incluyen 717 casos de heladas tardías.

	Bivariado	**Trivariado**
Heladas predichas correctamente	513	506
Heladas predichas como "no helada"	204	211
Recall para casos de heladas	0,72	0,71
F1-Score para casos de heladas	0,83	0,83

Tabla 24 - Análisis de predicción de heladas tardías

Como se visualiza en la Tabla 24, para estos casos de heladas, los modelos predicen gran cantidad de ellos, presentando valores de recall y F1-Score similares a los detallados en la Tabla 22 para los casos de heladas con datos sobremuestreados.

En resumen, los dos modelos entrenados con datos sobremuestreados, mejoran el recall y F1-Score para los casos de heladas, pero no sucede lo mismo para el valor del coeficiente de determinación y el RMSE. Esto significa que son capaces de predecir más situaciones de heladas correctamente, pero en cuanto a la precisión en el valor de la temperatura no hay mejora. Por lo tanto, se continúa la investigación usando datos sobremuestreados.

6.9. Neuronas de la capa oculta

En los modelos analizados anteriormente la cantidad de neuronas de la capa oculta, es igual a la cantidad de entradas. Se plantea el siguiente interrogante, ¿podría mejorar el rendimiento si la capa oculta del modelo posee menos o más neuronas?

Para responder a esto, se ha realizado el análisis para el caso bivariado y trivariado con distinta cantidad de neuronas en la capa oculta (modelos disponibles en el repositorio GitHub) [40].

6.9.1. Modelos Bivariados

Manteniendo los hiperparámetros de la arquitectura de red entrenada con 36 neuronas en la capa oculta (función de activación ReLU para la capa oculta, optimizador Adam y tasa de aprendizaje 8e-7) y con datos sobremuestreados, se entrenaron modelos con 18, 54 y 72 neuronas en la capa oculta.

En la Tabla 25 se detallan los resultados obtenidos, donde se observa que el incremento de la cantidad de neuronas en la capa oculta no produce una mejora en la exactitud del modelo, el coeficiente de determinación R^2, ni en el RMSE. Una leve mejora se aprecia en el valor del recall para los casos de heladas entre el modelo de referencia analizado desde el inicio de las experimentaciones (36 neuronas) con del de 54 y 72 neuronas. También se advierte que no es significativa la diferencia en las métricas entre el modelo de 54 y el de 72 neuronas.

	Neuronas			
	18	**36**	**54**	**72**
R^2 entrenamiento	0,79	0,80	0,80	0,80
R^2 test	0,66	0,66	0,67	0,67
RMSE entrenamiento	2,57	2,52	2,51	2,49
RMSE test	2,58	2,60	2,57	2,58
Recall (Helada)	0,64	0,73	0,75	0,76
F1-Score (Helada)	0,66	0,69	0,69	0,69
Recall (No helada)	0,96	0,95	0,94	0,94
F1-Score (No helada)	0,95	0,96	0,95	0,95
Exactitud (accuracy) en test	0,92	0,92	0,92	0,92

Tabla 25 – Resultados obtenidos de entrenar el modelo bivariado con distinta cantidad de neuronas en la capa oculta

En la curva de aprendizaje de cada modelo representada en la Figura 41, se visualiza que conforme la cantidad de neuronas se incrementa la curva de la función pérdida de entrenamiento y la de validación se interceptan más rápido.

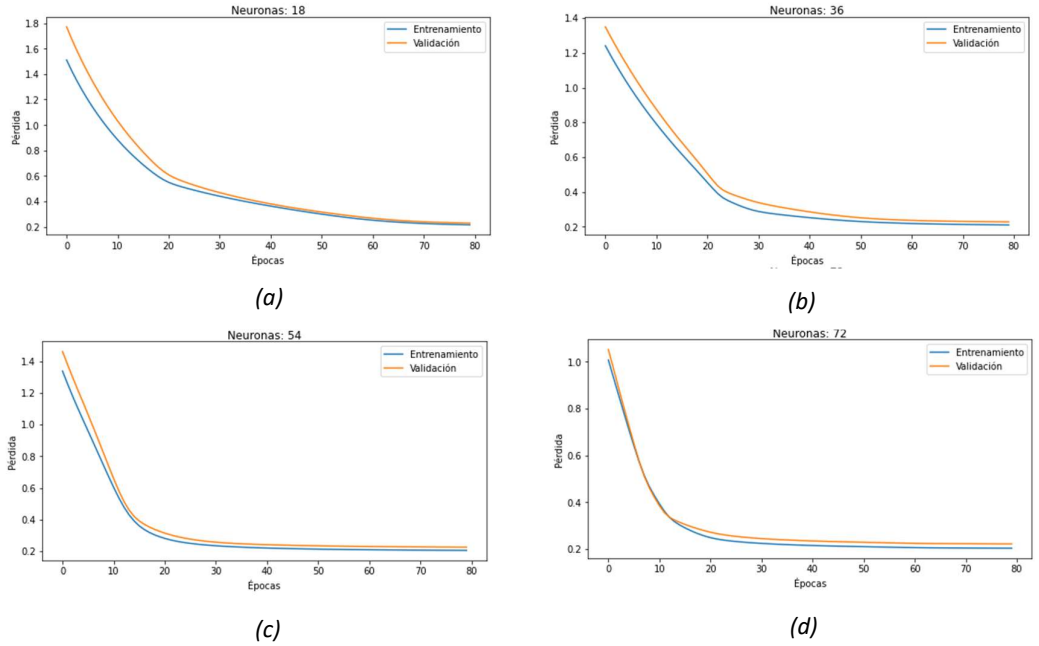


Figura 41 - Curva de aprendizaje para modelos bivariados con distinta cantidad de neuronas en la capa oculta. (a) 18 neuronas (b) 36 neuronas (c) 54 neuronas (d) 72 neuronas.

En síntesis, del conjunto de modelos bivariados analizados, el modelo de 36 neuronas es más adecuado por tener mayor recall para los casos de heladas que el modelo de 18 neuronas. Por otro lado, su curva de aprendizaje decae a través de las sucesivas épocas, contrario a los que sucede en los modelos de 54 y 72 neuronas.

6.9.2. Modelos Trivariados

Para el caso trivariado, se entrenaron modelos con 36, 72 y 90 neuronas manteniendo los parámetros de la arquitectura de red entrenada con 54 neuronas en la capa oculta (función de activación ReLU para la capa oculta, optimizador Adam y tasa de aprendizaje 8e-7) y con datos sobremuestreados.

Al igual que sucede para los modelos bivariados, se puede observar en la Tabla 26 que el incremento en la cantidad de neuronas en la capa oculta no produce una mejora en la exactitud, el coeficiente de determinación ni en el RMSE. La mejora del recall y el F1-Score no es significativa. Tampoco hay mejora en las métricas con disminución de la

cantidad de neuronas, sólo el beneficio de ser una capa oculta con menos componentes que por procesar.

	Neuronas			
	36	**54**	**72**	**90**
R^2 entrenamiento	0,81	0,81	0,81	0,81
R^2 test	0,67	0,67	0,67	0,67
RMSE entrenamiento	2,47	2,48	2,47	2,45
RMSE test	2,57	2,58	2,57	2,56
Exactitud (accuracy) en test	0,92	0,92	0,92	0,92
Recall (Helada)	0,75	0,74	0,77	0,77
F1-Score (Helada)	0,68	0,69	0,69	0,69
Recall (No helada)	0,94	0,94	0,94	0,94
F1-Score (No helada)	0,95	0,95	0,95	0,95

Tabla 26 - Resultados de entrenar modelos trivariados con 36, 54, 72 y 90 neuronas en la capa oculta.

Las respectivas curvas de aprendizaje no presentan sobreajuste y se observan en la Figura 42. Para los modelos de 72 y 90 neuronas, cae rápidamente antes de la época número 10, siendo más lenta la caída de curva para los modelos de 36 y 54 neuronas.

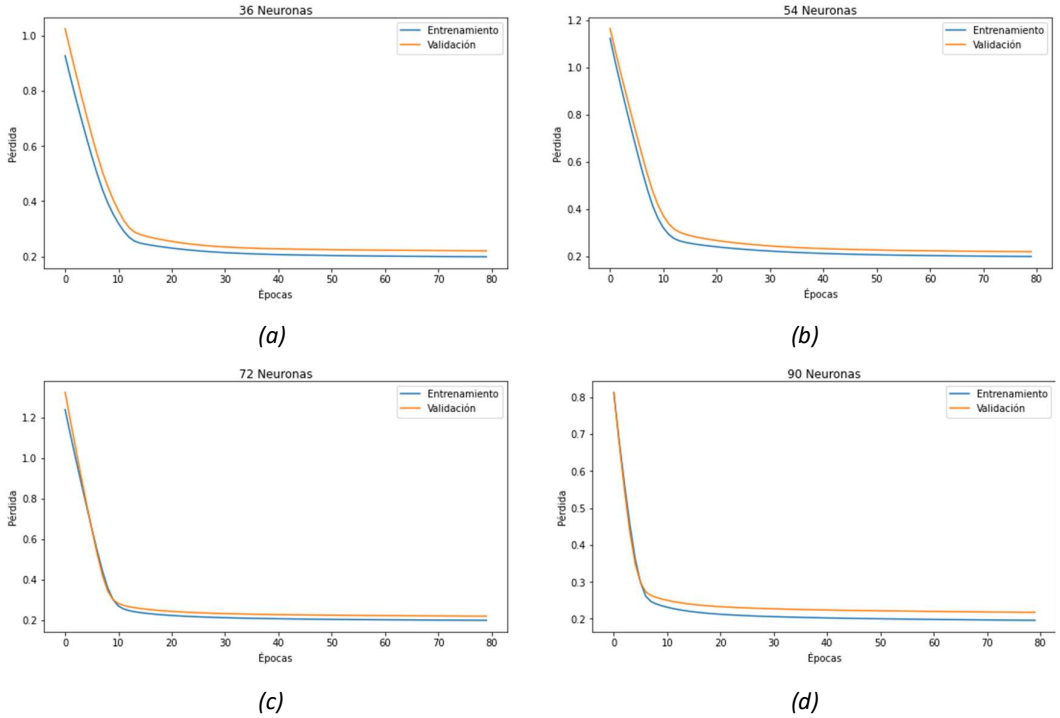


Figura 42 - Curvas de aprendizaje de modelos trivariados con distinta cantidad de neuronas en la capa oculta (a) 36 neuronas (b) 54 neuronas (c)72 neuronas (d) 90 neuronas

Del conjunto de modelos trivariados analizados, el modelo de 36 neuronas es el más adecuado porque no arroja diferencias relevantes en las métricas comparado con el resto de modelos de este conjunto; además por contener menos neuronas en la capa oculta es menor el procesamiento requerido. Por otra parte, su curva de aprendizaje es no presenta sobreajuste.

6.10. Modelo Final

La secuencia de experimentaciones y análisis realizados en los conjuntos de modelos establecidos, ha llevado a identificar como más apropiado dentro de los bivariados al modelo con 36 neuronas en la capa oculta, con tasa de aprendizaje de 8e-7 y sin regulación. En el caso de los trivariados, el modelo más adecuado es el de 36 neuronas en la capa oculta, con tasa de aprendizaje de 8e-7 y sin regulación. Ambos modelos entrenados con datos sobremuestreados. La comparativa de las métricas para estos modelos se presenta en la Tabla 27, donde se observa que para cada una de las métricas los valores son similares.

	Bivariado	**Trivariado**
R^2 entrenamiento	0,80	0,81
R^2 test	0,66	0,67
RMSE entrenamiento	2,52	2,47
RMSE test	2,60	2,57
Recall (Helada)	0,73	0,75
F1-Score (Helada)	0,69	0,68
Recall (No helada)	0,95	0,94
F1-Score (No helada)	0,96	0,95
Recall (Heladas tardías)	0,72	0,71
Exactitud (Accuracy) en test	0,92	0,92

Tabla 27 - Comparación de los modelos seleccionados

Finalmente, de los dos modelos comparados anteriormente, se ha establecido que es más apropiado el modelo bivariado. Esto se debe a que, como ya se ha mencionado, las métricas son similares, pero el modelo bivariado tiene menor cantidad de datos de entrada para procesar.

En cuanto a la evaluación el modelo presenta una exactitud del 92% en los casos de test, valor aceptable para la problemática tratada. Cabe aclarar que existe un

antecedente de modelo con redes neuronales convolucionales que logra el 98,86% de exactitud [7], esta diferencia en los resultados de los modelos se puede producir por diversos factores como es la arquitectura y tipo de red empleada, la cantidad de iteraciones durante el entrenamiento, el período de tiempo de registro de datos, la cantidad de dispositivos que censan las variables meteorológicas, entre otras.

En la predicción del fenómeno de la helada, los falsos negativos (se predice que no ocurrirá helada y en la realidad sucede) son los casos de gran atención, ya que estos llevan al productor a no desplegar medidas de mitigación y el cultivo termina dañado. La métrica que da información respecto a estos casos es la sensibilidad (recall), por ello en el análisis y evaluación de los modelos desarrollados se ha dado relevancia a esta métrica.

En este trabajo de investigación se ha desarrollado un modelo basado en una red neuronal LSTM que a partir de las variables meteorológicas temperatura y humedad relativa registradas a intervalos de 10 minutos durante un periodo de 3 horas, puede predecir la temperatura hacia un horizonte de 3 horas con una exactitud (accuracy) del 92% y la capacidad de pronosticar el 73% de casos de helada (sensbilidad).

Un análisis más específico refiere a que, de los casos de heladas predichos correctamente, aproximadamente el 41% fue pronosticado con una temperatura menor a la real, lo que llevaría a que el productor prepare recursos suficientes para mitigar el fenómeno meteorológico. Para los casos de heladas tardías el modelo presenta un recall de 0,72, valor similar al obtenido para esta métrica en la predicción de todos los casos de heladas, es decir que el modelo tiene la misma capacidad para predecir en época normal de helada como de helada tardía.

Con este trabajo se han logrado resultados satisfactorios, para comenzar a dar respuesta a la problemática del daño ocasionado por el fenómeno meteorológico de la helada en la provincia de San Juan.

Capítulo 7

Conclusiones

7. Conclusiones

En este trabajo se ha adoptado el proceso de Ciencia de Datos para el análisis de datos recopilados desde estaciones agrometeorológicas y se han desarrollado distintos modelos de pronóstico de la ocurrencia de heladas. De este conjunto de modelos se ha seleccionado el más adecuado según el análisis de los distintos resultados obtenidos. De este modo se está colaborando con los productores en la prevención del daño generado por este fenómeno agroclimático de la helada.

Las estaciones que han registrado los datos se encuentran separadas entre ellas por una distancia cercana a los 37 km. Se ha observado que en la zona de la estación San Francisco, en el departamento Sarmiento, presenta un período de heladas más amplio, con mayor cantidad de casos de heladas y de heladas tardías. Llegando a duplicar y triplicar a los casos de la estación INTA localizada en Pocito. Esto también se ve reflejado en la intensidad del fenómeno, siendo en San Francisco más intensas.

Se ha observado que los días con heladas tardías representan menos del 40% del total de días para un periodo de heladas.

El análisis de las variables a través de la matriz de correlación, ha permitido establecer que la temperatura se encuentra inversamente correlacionada con la humedad relativa y directamente con el punto de rocío.

En la fase de modelado se ha llevado a cabo un proceso empírico, basado en el entrenamiento de distintos modelos con variaciones en los hiperpárametros. Con posterior análisis basado en las métricas (R^2, RMSE, recall y F1-Score) y en las curvas de aprendizaje.

Para estructurar los datos se ha aplicado la técnica de ventana deslizante para el tamaño de 2, 3 y 4 horas de registro de variables. Para analizar y determinar el tamaño de ventana más adecuado se han desarrollado tres modelos LSTM univariados (uno para cada tamaño de ventana). De acuerdo a los resultados obtenidos se ha observado que la ventana de 3 horas es la más adecuada.

Se ha desarrollado una red neuronal de tipo LSTM capaz de predecir el fenómeno de la helada a partir de los valores registrados a intervalos de 10 minutos durante 3 horas, de la temperatura y la humedad relativa, con un horizonte de 3 horas. La cual

opera con datos balanceados con la técnica de SMOTEEN en cuanto a cantidad de casos de heladas y no heladas.

De forma general se puede concluir que entrenar las redes neuronales con datos balanceados otorga beneficios. Se ha comprobado que los modelos entrenados con datos sobremuestreados presentan mejor capacidad para predecir casos de heladas que el modelo con datos no sobremuestreados.

El incremento de la cantidad de neuronas en la capa oculta o en los datos de entrada no garantiza una mejora en los pronósticos que realiza el modelo. Por tanto, la definición de las entradas como de los hiperparámetros para la red neuronal deben ser analizados cuidadosamente.

La tasa de aprendizaje debe ser el primer hiperparámetro a analizar para un modelo, debido a que impacta fuertemente en la curva de aprendizaje del modelo. Si una tasa de aprendizaje produce oscilaciones (rebotes) en la curva, esta situación se mantendrá en mayor o menor medida al modificar el resto de los hiperparámetros.

En particular, sobre el modelo final se puede concluir que presenta una exactitud del 92% y es capaz de pronosticar el 73% de los casos de heladas, de los cuales alrededor el 41% es con una temperatura menor a la real. Esto es importante debido a que, si bien la planificación de medidas de mitigación del daño para una helada más intensa que lo real, trae como consecuencia el dispendio de recursos, pero garantiza la protección del cultivo. En el caso contrario, si la temperatura pronosticada es mayor a la real, se corre el riesgo de que los recursos sean insuficientes y finalmente el cultivo resulte perjudicado.

Respecto de las heladas tardías, el modelo desarrollado es capaz de pronosticar el 72% de ellas, prácticamente el mismo valor que para los casos de heladas en general.

Así, se ha desarrollado y analizado un conjunto de variantes en cuanto a datos de entrada, hiperparámetros y modelos, a través de un proceso enriquecedor. Se ha obtenido un modelo final que ha arrojado resultados satisfactorios. Este es un buen comienzo para dar solución al problema del daño que causa el fenómeno meteorológico de la helada en los cultivos en la zona sur de la provincia de San Juan.

Finalmente, cabe aclarar que a partir de esta investigación se han realizado publicaciones y exposiciones en distintos eventos científicos.

El trabajo “Entorno web de visualización de información meteorológica para el uso agrícola y de generación de alertas ante eventos climáticos” en el marco del XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019) [42].

El artículo “Procesamiento de Datos Meteorológicos para Determinar la Ocurrencia de Heladas en la Agricultura” en el marco del XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021) [43].

El trabajo “Técnicas de balanceo de datos para predecir la ocurrencia del fenómeno meteorológico de la helada” en la XIX Reunión de Trabajo en Procesamiento de la Información y Control (RPIC’2021) [25].

El trabajo “Procesamiento de datos meteorológicos para determinar la ocurrencia, intensidad y duración de heladas” en las 51º Jornadas Argentinas de Informática – JAIIO.

El trabajo “Pronóstico de heladas meteorológicas con datos balanceados en redes neuronales LSTM” en el marco del 1° Congreso Latinoamericano de Ciencia, Tecnología y Sociedad.

7.1. Trabajo futuro

La continuidad de este trabajo puede darse desde distintas perspectivas. Ya sea a través de mejoras al modelo, modificaciones al modelo, predicción de otras variables, entre otras.

Investigar posibles mejoras en la precisión en las predicciones del modelo que se ha logrado. Continuando el análisis de los hiperparámetros, así como otras variables meteorológicas como el viento; y no meteorológicas como la fecha y hora, entre otras. Además de otras arquitecturas u otro tipo de redes.

Considerar horizontes de predicción más lejanos, que permita al productor el abastecimiento de recursos para las medidas de protección.

Por otra parte, es de interés de pronosticar la duración de la helada, además de la intensidad. Estas dos variables están directamente ligadas al daño de la helada, la

magnitud del daño es distinta ante una helada de poca intensidad y larga duración, como una helada de mucha intensidad y corta duración.

La recopilación de datos, análisis, investigación y desarrollo de un clasificador con capacidad de identificar el posible tipo de la helada que ocurrirá.

Es relevante analizar de forma particular las heladas tardías; el pronóstico y contexto de ocurrencia. Si bien estos fenómenos son escasos, resultan muy dañinos para los cultivos por encontrarse en estado de floración o cuaje del fruto. Como se ha constatado en este trabajo la cantidad de días con heladas tardía en un período de heladas es muy baja, por lo tanto, este futuro estudio debe prever que deberá afrontar el problema de la escasez en la cantidad de casos.

La predicción de heladas es amplia y permite continuar con distintas investigaciones.

Referencias

[1] "Instituto Nacional de Tecnología Agropecuaria | Argentina.gob.ar." https://www.argentina.gob.ar/inta (accessed Oct. 20, 2019).

[2] J. L. F. Yagüe, *Iniciación a la meteorología y climatología*. España, 2012.

[3] P. Möller-Acuña, R. Ahumada-García, and J. Reyes-Suárez, "Predicción de Episodios de Heladas Basado en Información Agrometeorológica y Técnicas de Aprendizaje Automático," Dec. 2016, doi: 10.1109/ICA-ACCA.2016.7778386.

[4] V. L. Castro and S. E. Alday, "Reporte sobre el efecto de helada tardía en el rendimiento de la variedad de almendra Guara durante el ciclo productivo 2016-2017 en el departamento de Pocito provincia de San Juan | Instituto Nacional de Tecnología Agropecuaria," San Juan, Argentina, 2018. Accessed: Mar. 26, 2020. [Online]. Available: https://inta.gob.ar/documentos/reporte-sobre-el-efecto-de-helada-tardia-en-el-rendimiento-de-la-variedad-de-almendra-guara-durante-el-ciclo-productivo-2016-2017-en-el-departamento-de-pocito-provincia-de-san-juan.

[5] M. H. Jorenoosh and A. R. Sepaskhah, "Prediction of frost occurrence by estimating daily minimum temperature in semi-arid areas in Iran," vol. 37, no. 1, pp. 19–32, 2018, doi: 10.22099/IAR.2018.4689.

[6] M. Ángel, O. Chong, S. De Gobernación, L. Felipe, and P. Espinosa, "SERIE Fascículos - Heladas," Mexico, 2014.

[7] R. M. A. Latif, S. B. Brahim, S. Saeed, L. B. Imran, M. Sadiq, and M. Farhan, "Integration of Google Play Content and Frost Prediction Using CNN: Scalable IoT Framework for Big Data," *IEEE Access*, vol. 8, pp. 6890–6900, 2020, doi: 10.1109/ACCESS.2019.2963590.

[8] A. L. Diedrichs, F. Bromberg, D. Dujovne, K. Brun-Laguna, and T. Watteyne, "Prediction of frost events using Bayesian networks and Random Forest," 2018.

[9] M. Fuentes, C. Campos, and S. García-Loyola, "Application of artificial neural networks to frost detection in central chile using the next day minimum air temperature forecast," *Chil. J. Agric. Res.*, vol. 78, no. 3, pp. 327–338, Sep. 2018, doi: 10.4067/S0718-58392018000300327.

[10] J. R. Rozante, E. R. Gutierrez, P. L. da Silva Dias, A. de Almeida Fernandes, D. S. Alvim, and V. M. Silva, "Development of an index for frost prediction: Technique and validation," *Meteorol. Appl.*, 2019, doi: 10.1002/met.1807.

[11] "Los daños por heladas tardías alcanzaron las 30 mil hectáreas de vid y 16 mil de frutales en Mendoza en 2020," 2021. https://www.infocampo.com.ar/los-danos-por-heladas-tardias-alcanzaron-las-30-mil-hectareas-de-vid-y-16-mil-de-frutales-en-mendoza-en-2020/ (accessed Apr. 14, 2022).

[12] "Alerta entre los productores de Cuyo por las heladas tardías - Revista InterNos," 2021.

[13] R. L. Snyder, J. P. de Melo-Abreu, and J. M. Villar-Mir, "Protección contra las heladas: fundamentos, práctica y economía," *Ser. FAO Sobre el Medioambiente y la Gestión los Recur. Nat.*, vol. 1, p. 257, 2010, Accessed: Mar. 30, 2020. [Online]. Available: http://www.fao.org.

[14] J. M. Raigón and S. Silva, "La producción de almendros en San Juan y su vinculación con el viento Zonda y las heladas. | Instituto Nacional de Tecnología Agropecuaria," San Juan, Argentina, 2012. Accessed: Mar. 26, 2020. [Online]. Available:

https://inta.gob.ar/documentos/la-produccion-de-almendros-en-san-juan-y-su-vinculacion-con-el-viento-zonda-y-las-heladas.

[15] E. Abrahamsen, O. M. Brastein, and B. Lie, “Machine Learning in Python for Weather Forecast based on Freely Available Weather Data,” in *Proceedings of The 59th Conference on imulation and Modelling (SIMS 59), 26-28 September 2018, Oslo Metropolitan University, Norway*, Nov. 2018, vol. 153, pp. 169–176, doi: 10.3384/ecp18153169.

[16] M. A. Tomkowicz and A. O. Schmitt, “Frost Prediction in Apple Orchards Based upon Time Series Models,” in *Data Analysis and Applications 1*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2019, pp. 181–194.

[17] S. Lee, Y.-S. Lee, and Y. Son, “Forecasting Daily Temperatures with Different Time Interval Data Using Deep Neural Networks,” *Appl. Sci.*, vol. 10, no. 5, p. 1609, Feb. 2020, doi: 10.3390/app10051609.

[18] P. Hewage, M. Trovati, E. Pereira, and A. Behera, “Deep learning-based effective fine-grained weather forecasting model,” *Pattern Anal. Appl.*, vol. 24, no. 1, pp. 343–366, Feb. 2021, doi: 10.1007/S10044-020-00898-1/FIGURES/12.

[19] A. Castañeda-Miranda and V. M. Castaño, “Smart frost control in greenhouses by neural networks models,” *Comput. Electron. Agric.*, vol. 137, pp. 102–114, May 2017, doi: 10.1016/j.compag.2017.03.024.

[20] C. Talsma, K. C. Solander, M. K. Mudunuru, B. Crawford, and M. Powell, “Frost Prediction Using Machine Learning and Deep Neural Network Models for Use on Iot Sensors,” *SSRN Electron. J.*, Feb. 2022, doi: 10.2139/SSRN.4032447.

[21] L. Igual and S. Seguí, *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*. Barcelona, España, 2017.

[22] S. Ozdemir, *Principles of data science : learn the techniques and math you need to start making sense of your data*. 2016.

[23] F. Cady, *The Data Science Handbook*, 1st. United States of America, 2017.

[24] L. V. Gutiérrez, M. O. Ortega, M. I. Masanet, and F. De La Jara, “Técnicas de Análisis y Visualización de Minería de Datos para la Reducción de Dimensiones en Tablas de Datos,” *An. del Congr. Int. Ciencias la Comput. y Sist. Inf. 2019*, 2019.

[25] M. I. Masanet, R. Klenzi, and F. Capraro, “Técnicas de balanceo de datos para predecir la ocurrencia del fenómeno meteorológico de la helada,” *Actas la XIX Reun. Trab. en Proces. la Inf. y Control. RPIC’2021*, pp. 511–516, 2021, [Online]. Available: https://drive.google.com/file/d/1byaIS-ssvJP-SMHtKP9ahu8LQuoshyq6/view.

[26] F. Chollet, *Deep Learning with Python*, 1st ed. USA: Manning Publications Co., 2017.

[27] C. C. Aggarwal, *Neural networks and deep learning : a textbook*, 1st ed. Cham, Switzerland: Springer Nature Switzerland AG, 2018.

[28] Z. Zhang, *Multivariate Time Series Analysis in Climate and Environmental Research*, 1st ed. Springer International Publishing, 2018.

[29] S. Siaamimi-Nni, N. Tavakoli, and A. Siami Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series,” in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, Jan. 2019, pp. 1394–1401, doi: 10.1109/ICMLA.2018.00227.

[30] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd Editio. United States of America: O'Reilly Media, Inc., 2019.

[31] A. Konar and D. Bhattacharya, *Time-Series Prediction and Applications*, vol. 127. 2017.

[32] M. S. Mayuri Shelke, P. R. Deshmukh, and V. K. Shandilya, "A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique," doi: 10.23883/IJRTER.2017.3168.0UWXM.

[33] I. Pisa, I. Santín, J. L. Vicario, A. Morell, and R. Vilanova, "Data preprocessing for ANN-based industrial time-series forecasting with imbalanced data," in *European Signal Processing Conference*, Sep. 2019, vol. 2019-September, doi: 10.23919/EUSIPCO.2019.8902682.

[34] L. G. Matías Ramirez, Ó. A. Fuentes Mariles, and F. García Jiménez, *Fascículo Heladas*, 1a. Edició. México, 2001.

[35] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, pp. 243–248, Apr. 2020, doi: 10.1109/ICICS49469.2020.239556.

[36] "NumPy." https://numpy.org/ (accessed Mar. 26, 2023).

[37] "Scikit-Learn : Descubre la biblioteca de Python dedicada al Machine Learning." https://datascientest.com/es/scikit-learn-decubre-la-biblioteca-python (accessed Mar. 26, 2023).

[38] "TensorFlow Core | Aprendizaje automático para principiantes y expertos." https://www.tensorflow.org/overview?hl=es-419 (accessed Mar. 26, 2023).

[39] "Matplotlib — Visualization with Python." https://matplotlib.org/ (accessed Mar. 26, 2023).

[40] "mamasanet/Prediccion-heladas: Predicción de heladas usando redes neuronales." https://github.com/mamasanet/Prediccion-heladas/tree/main (accessed Jun. 08, 2023).

[41] "tf.keras.regularizers.Regularizer | TensorFlow Core v2.9.1." https://www.tensorflow.org/api_docs/python/tf/keras/regularizers/Regularizer (accessed Aug. 06, 2022).

[42] M. Masanet, F. Capraro, R. Klenzi, M. Muñoz, and C. Suarez, "ENTORNO WEB DE VISUALIZACIÓN DE INFORMACIÓN METEOROLÓGICA PARA EL USO AGRÍCOLA Y DE GENERACIÓN DE ALERTAS ANTE EVENTOS CLIMÁTICOS," San Juan, Argentina, 2019.

[43] M. Masanet, F. Capraro, R. Klenzi, and M. Muñoz, "PROCESAMIENTO DE DATOS METEOROLÓGICOS PARA DETERMINAR LA OCURRENCIA DE HELADAS EN LA AGRICULTURA," *WICC 2019*, 2019.

Acrónimos, siglas y abreviaturas

ANN	Red Neuronal Artificial (Artificial Neural Network)
AR-ANN	Red Neuronal Artificial Autorregresiva
ARIMA	Promedio móvil integrado autorregresivo (Autoregressive Integrated Moving Average)
ARX	Autoregresivo con entrada exógena (AutoRegressive eXogenous)
CNN	Red Neuronal Convolucional (Convolutional Neural Network
DNN	Red Neuronal Profunda (Deep Neural Network)
EEA	Estación Experimental Agropecuaria
ELU	Unidad Lineal Exponencial (Exponential Linear Unit)
FN	Falsos negativos
FP	Falsos positivos
INTA	Instituto Nacional de Tecnología Agropecuaria
IG	Índice de Geada
IoT	Internet de las cosas (Internet of things)
LOOCV	(Leave One-Out Cross Validation)
LSTM	Long Short-Term Memory
MAE	Error Absolute Medio (Mean Absolute Error)
MAPE	Error Porcentual Absoluto Medio
ME	Error Medio (Medium Error)
MLP	Multicapa Perceptrón
MSE	Error Cuadrático Medio (Mean Squared Error)
NRMSE	Error Cuadrático Medio Normal (Normalized Root Mean Square Error)
R^2	Coeficiente de determinación
RELU	Unidad Lineal Rectificada (Rectified Linear Unit)
RMSE	Raíz del Error Cuadrático Medio (Root Mean Squared Error)
SGD	Descenso de Gradiente Estocástico (Stochastic gradient descent)
SMOTE	Técnica de Sobremuestreo de Minorías Sintéticas (Synthetic Minority Oversampling Technique)
TIC	Tecnologías de Información y Comunicación
TN	Verdaderos Negativos
TP	Verdaderos Positivos