CACiC 2023
CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN

Ciencia de Datos para el Desarrollo de un Modelo Predictivo de Heladas

María Isabel Masanet[1], Raúl Orcar Klenzi[1]

[1] Facultad de Ciencias Exactas, Físicas y Naturales,
Universidad Nacional de San Juan, San Juan, Argentina.
mimasanet@gmail.com, rauloscarklenzi@unsj-cuim.edu.ar

Resumen. En este trabajo se aplica Ciencia de Datos para el desarrollo de un modelo que predice el fenómeno meteorológico de la helada en la Provincia de San Juan, Argentina. A través del proceso sistemático que propone la Ciencia de Datos se ha logrado un modelo basado en una red neuronal recurrente de memoria de corto-largo plazo que, a partir de los valores de la temperatura y la humedad relativa censados cada diez minutos durante tres horas predice la temperatura hacia un horizonte de tres horas. La red fue entrenada, validada y testeada a partir de datos climáticos registrados por dos estaciones meteorológicas instaladas en la provincia de San Juan. Estos datos fueron preprocesados con un enfoque orientado hacia el análisis del fenómeno de la helada; además se aplicó la técnica de remuestreo SMOTEEN. Los resultados del modelo se analizaron con métricas de regresión y de clasificación.

Palabras clave: Ciencia de datos, Machine Learning, Red LSTM, Heladas.

1. Introducción

La Ciencia de Datos comúnmente se define como una metodología [1] mediante la cual se puede extraer conocimiento a partir de los datos, con el fin de utilizar este conocimiento para predecir eventos, comprender el pasado o presente, crear nuevos productos, entre otros usos. Tiene un objetivo ambicioso, pretendiendo generar opiniones basadas en los datos para que sean usadas en la toma de decisiones [2].

El proceso de la Ciencia de Datos 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. Luego, la etapa de exploración de los datos, para extraer características significativas que son usadas por herramientas de modelado y análisis, para finalmente obtener y presentar los resultados obtenidos [3]. 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.

Los proyectos de Ciencia de Datos suelen involucrar algún algoritmo de aprendizaje automático o Machine Learning (ML) [3]. El ML es un área de la

Inteligencia Artificial o Artificial Intelligence (AI) que busca aprender automáticamente relaciones y patrones significativos a partir de ejemplos y observaciones. Es decir, los algoritmos aprenden iterativamente de los datos de entrenamiento específicos del problema, lo que permite encontrar información oculta y patrones complejos sin ser programados explícitamente. Al aprender de cálculos anteriores y extraer regularidades de conjuntos de

datos masivos, ayuda a producir decisiones confiables y repetibles. Por esta razón, los algoritmos de ML se han aplicado con éxito en diversas áreas [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 [5]. 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.

La helada es un fenómeno meteorológico localizado que puede ocasionar daño en diferentes niveles en los cultivos. Puede destruir toda la producción en cuestión de horas, provocando pérdidas de la cosecha de un año entero y comprometiendo ingresos del siguiente. 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 [6].

Durante el año 2.020, en la provincia de Mendoza, el daño de las heladas tardías en el cultivo de vid fue estimado en un 24%, mientras que en el resto de las frutas fue del 84% [7]. En 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, ocasionando daños en los parrales [8].

En este trabajo se presenta el desarrollo de un modelo de predicción del fenómeno meteorológico de la helada a través de en una red neuronal recurrente de memoria a corto-largo plazo aplicando el proceso de Ciencia de Datos.

La organización del documento es la siguiente: la Sección 1 introduce los conceptos generales que explican el contexto y el propósito del trabajo; la Sección 2 presenta los trabajos relacionados a la predicción de heladas con ML; la Sección 3 describe el desarrollo del estudio y los resultados obtenidos. Finalmente, la Sección 4 presenta las conclusiones y sugiere posibles trabajos futuros.

2. Trabajos Relacionados

Diversos estudios se han llevado a cabo para el análisis y pronóstico de la temperatura. Los más recientes, aplican algoritmos de aprendizaje automático como redes neuronales de distinto tipo. En la mayoría de los casos, los datos son obtenidos de estaciones meteorológicas; aplicando técnicas de remuestreo cuando se presenta el inconveniente de escasez en la cantidad de datos reales.

Para predicciones hacia un horizonte a corto plazo, como es por hora y por día, se han usado 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), con datos meteorológicos de tres regiones de Corea del Sur, para el período de 2.009 a 2.018. Concluyendo 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[9].

Un análisis comparativo de rendimiento de modelos para el pronóstico del clima con redes de memoria a largo-corto plazo (LSTM) y redes convolucionales temporales

(TCN) con enfoques clásicos (Regresión estándar, ARIMA, Random Forest, entre otros) presentan los autores de [10]; 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.

Castañeda-Miranda y Castaño [11] desarrollaron un modelo autorregresivo con entrada externa (ARX) y otro con una red neuronal artificial (ANN) perceptrón multicapa, ambos capaces de pronosticar la temperatura interior de un invernadero. La conclusión fue que el modelo basado en ANN producía mejores resultados, con un nivel de confianza de 95%.

En Chile, con el objeto de predecir la temperatura mínima a partir de la información recopilada por diez estaciones meteorológicas desarrollaron una red neuronal artificial (ANN), con un número diferente de neuronas en la capa oculta (1, 5, 10, 15, 20, 25 y 30 neuronas). Los autores concluyeron que la ANN con 25 neuronas en la capa oculta fue el modelo que proporcionó los mejores resultados [5].

La técnica de sobremuestreo sintético de la clase minoritaria o Synthetic Minority Over-sampling Technique (SMOTE) fue aplicada en [6] para sobremuestrear los datos de cinco estaciones agrometeorológicas de la provincia de Mendoza, Argentina, a partir de los cuales trataron la predicción de heladas mediante un modelo desarrollado con redes Bayesianas y Random Forest. Al igual que la investigación de MöllerAcuña et all. [12], donde se generan datos a través de SMOTE.

En [13] el conjunto de datos obtenido de las estaciones meteorológicas es remuestreado con los métodos SMOTE, ADASYN, SMOTETomek y SMOTEEN. Cada conjunto de datos es entrada de un modelo basado en el algoritmo Random Forest que predice la ocurrencia o no ocurrencia de la helada. De la comparación de los resultados obtenidos concluyeron que SMOTEENN produce mejores resultados.

3. Desarrollo

El estudio se desarrolló a partir del 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 Fig. 1 resume el flujo del trabajo realizado. Iniciando con el entendimiento del contexto del problema a través de información brindada por expertos. Se obtuvieron los datos desde dos estaciones meteorológicas, los cuales fueron sometidos a las tareas de análisis exploratorio. Luego prepararon en una estructura de ventana deslizante, adecuada para el procesamiento con algoritmos a usar. Posteriormente, en la fase de modelado se desarrolló una red neuronal de tipo LSTM. Finalmente, la evaluación del modelo se realizó a través de las métricas (recall, F1-Score y exactitud) y de la representación gráfica de las predicciones realizadas.

Fig. 1. - Proceso de Ciencia de Datos aplicado en la investigación. Fuente: Autora

3.1 Entendimiento del Contexto del Problema

Para el entendimiento del contexto del problema se solicitó información a expertos. 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 [14]. Se produce cuando la superficie terrestre y el aire que se siente sobre ella alcanza una temperatura por debajo de los 0°C [15].

El pronóstico del fenómeno meteorológico de la helada a través de algoritmos de ML requiere de un conjunto de datos. Para este estudio se usaron los datos provistos por el Instituto de Automática (INAUT) de la Facultad de Ingeniería de la Universidad Nacional de San Juan, que correspondían a una estación meteorológica situada en el Establecimiento San Francisco S.A. (explotación privada), en la localidad de Cañada Honda, departamento Sarmiento, provincia de San Juan, Argentina. Otro conjunto de datos fue cedido por el Servicio de Agrometeorología de la Estación Experimental Agropecuaria San Juan dependiente del Instituto Nacional de Tecnología Agropecuaria (INTA) del departamento Pocito, donde se encuentra instalada otra estación meteorológica. Existiendo entre ellas una distancia de 37 km aproximadamente.

Ambas estaciones registran 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. Realizan la medición y almacenamiento de los datos de las variables cada 10 minutos.

3.2 Análisis Exploratorio de los Datos

De la estación situada en INTA se consideraron los 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. De San Francisco se disponían un total de 299.671 registros, de los cuales 13.872 corresponden a heladas (aproximadamente el 4,63%). Para la estación INTA, el

total de registros es 301.899 que incluyen 3.854 casos de helada (aproximadamente el 1,28%).

Los datos registrados por las estaciones presentan una estructura tabular. La Fig. 2 muestra el preprocesamiento realizado a los datos.

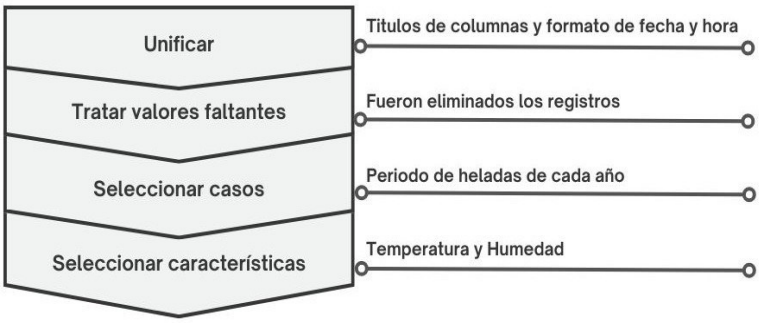

Fig. 2. Tareas de preprocesamiento de los datos

Unificar. Si bien ambas estaciones miden a igual intervalo de tiempo y las mismas variables, existieron diferencias en los conjuntos de datos que se debieron unificar, como nombre de columnas y formato de fecha y hora.

Tratar valores faltantes. 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 [16]. En este caso se optó por la primera opción.

Seleccionar casos. Por cada año se seleccionaron los datos correspondientes al periodo comprendido entre la fecha de la primera helada y la última helada. Las heladas son frecuentes entre las 00 horas y 8 horas, esto también se tuvo en cuenta para la selección de casos.

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

La Tabla 1 muestra la cantidad de casos que se contiene el conjunto original y la cantidad de casos que resultaron luego de las tareas de selección, discriminando cantidad de heladas y no heladas.

Tabla 1. Cantidad total, de heladas y no heladas del conjunto original y el conjunto resultante de la selección de casos para cada estación.

Estación	Casos	Originales	Seleccionados
San Francisco	Total	299.671	82.975
	Heladas	13.872	13.814
	No Heladas	285.799	69.161
INTA	Total	301.899	49.973
	Heladas	3.854	3.830
	No Heladas	298.045	46.143

3.3 Preparación de los Datos

Los datos se estructuraron con la técnica de ventana deslizante, tomando los registros de las variables meteorológicas durante un periodo de 3 horas, este valor fue obtenido a través de un trabajo empírico desarrollado previamente.

El horizonte de predicción es de 3 horas, este el tiempo mínimo necesario para que el productor pueda desplegar las medidas de mitigación contra el fenómeno. La Fig. 3 representa la estructura de los datos.

Fig. 3. Ventana de entrada de 18 valores de cada variable meteorológica y salida con horizonte de 3 horas.

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 [17]. 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. Cada conjunto queda conformado por la cantidad de casos indicados en la Tabla 2.

Tabla 2. Cantidad de casos para cada conjunto de datos requerido para el modelado.

	Entrenamiento	Validación	Test
Total de casos	44.203	8.404	31.613
Casos de Heladas	9.838	2.843	3.760
Casos de No Heladas	34.365	5.561	27.853

En esta tarea de preparación de datos no fue posible la conformación de la ventana de 3 horas para algunos casos, debido a la falta de registro de datos ocasionada por problemas en la estación meteorológica.

Se puede observar claramente en la Tabla 2 que los conjuntos presentan un desbalance, superando notablemente la cantidad de casos de no heladas a los casos de heladas, siendo estos últimos el interés de la predicción. Este sesgo en el conjunto de datos de entrenamiento suele reflejarse en el rendimiento del modelo, siendo más preciso para los casos mayoritarios, que para los casos que son minoría.

En esta experimentación, el conjunto de datos de entrenamiento ha sido balanceado con la técnica de remuestreo SMOTEEN [18]. En un estudio previo concluyó que esta técnica arroja buenos resultados al aplicarla sobre estos datos para ser usados en el algoritmo de clasificación Random Forest [13]. Luego del remuestreo, la cantidad de casos que conforman el conjunto de entrenamiento y de validación usados para entrenar la red se detalla en la Tabla 3.

Tabla 3. Cantidad de casos para los conjuntos remuestreados.

	Entrenamiento	Validación
Total de casos	65.643	21.868
Casos de Heladas	34.000	10.810
Casos de No Heladas	31.643	11.058

3.4 Modelado

La arquitectura de red neuronal recurrente LSTM desarrollada consta de la capa de entrada con 18 valores de las variables temperatura y humedad relativa (36 valores en total); una capa oculta con función de activación RELU, por la facilidad que presenta esta función para el entrenamiento de las redes neuronales [19]; la función de pérdida (loss) el MSE por tratarse de un problema de regresión; y el algoritmo de optimización es Adam, con un valor para la tasa de aprendizaje de 8e-7. La capa de salida posee una neurona con función de activación lineal. La Fig. 4 representa la arquitectura de la red.

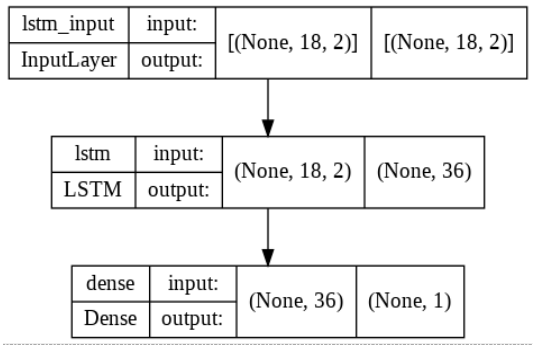

Fig. 4. Arquitectura de la red neuronal.

El entrenamiento de la red se realizó para 80 épocas (epochs). La curva de aprendizaje del modelo se muestra en la Fig. 5.

El modelo entrega como resultado el pronóstico de la temperatura a un horizonte de 3 horas. Luego, a partir del valor predicho para cada caso del conjunto de test, se realiza el etiquetado del caso con 0 o 1, según corresponda a una temperatura de helada o no helada respectivamente, transformando los resultados en un caso de clasificación binaria.

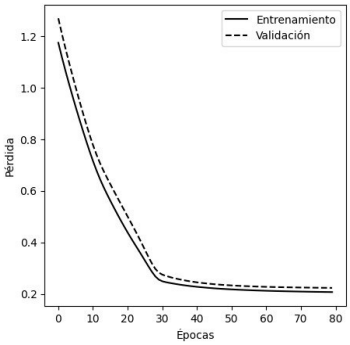

Fig. 5. Curva de aprendizaje del entrenamiento de la red neuronal.

3.5 Evaluación y Visualización

Para evaluar el modelo, los resultados arrojados por la red analizaron como un problema de regresión, a través del valor del MSE, RMSE y el coeficiente R^2. Luego del etiquetado y transformación a clasificación, se construyó la matriz de correlación y se analizó el valor recall y el F1-Score para los casos de heladas. Los resultados obtenidos por el modelo se presentan en la Tabla 4.

Tabla 4. Resultados de las métricas analizadas para el modelo desarrollado.

Métrica	Valor
R^2 entrenamiento	0,80
R^2 test	0,66
RMSE entrenamiento	2,52
RMSE test	2,60
Recall (Helada)	0.73
F1-Score (Helada)	0.69
Recall (No helada)	0.95
F1-Score (No helada)	0,96
Exactitud (Accuracy) en test	0.92

Las métricas de regresión (R^2 y RMSE) no presentan diferencias muy significativas, evidenciando que en el funcionamiento del modelo no hay sobreajuste. Se puede observar que el recall para los casos de heladas es de 0,73; lo que indica que el 73% de

las heladas son predichas correctamente. Y la exactitud del modelo es 0,92 contra 1 que es el valor ideal.

Además de analizar los valores de las métricas, se analizó la representación gráfica de la predicción realizada con el conjunto de test (la Fig. 6 muestra un recorte de la gráfica) donde se visualizó que la mayoría de los casos de heladas reales fueron predichos por el modelo.

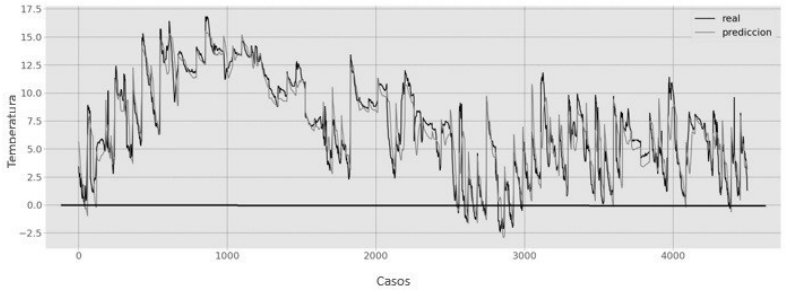

Fig. 6. Recorte de la representación de los valores reales del conjunto de test y sus correspondientes predicciones.

4. Conclusiones

Ante la necesidad de pronosticar el fenómeno meteorológico de la helada en una zona de la provincia de San Juan, Argentina, se llevó adelante este estudio que aplica el proceso de Ciencia de Datos para satisfacer el requerimiento.

Se logró desarrollar 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 o recall).

El proceso aplicado condujo a que los datos obtenidos desde las estaciones meteorológicas sean sometidos a una intensa tarea de preprocesamiento, para lo que ha planteado un flujo de trabajo que comienza con la unificación de los nombres de columnas y de formato de los datos, luego el tratamiento de valores faltantes, seguido de la selección de casos relevantes para el problema a tratar, y finalmente la identificación de las características de interés.

El método de remuestreo SMOTEEN aplicado a los datos estructurados en una ventana deslizante de 3 horas (18 valores) ha permitido generar conjuntos de datos equilibrados con los que fue entrenada la red neuronal.

Se puede concluir que el proceso de Ciencia de Datos ha permitido cumplir con el objetivo de desarrollar de forma sistemática y disciplinada un modelo de predicción del fenómeno de la helada con resultados aceptables.

En trabajos futuros incorporando datos desde otras estaciones meteorológicas y/o manteniendo los datos analizados, se pueden plantear mejoras al modelo, predicción de otras variables o considerar horizontes de predicción más lejanos.

Referencias

[1] L. Igual and S. Seguí, Introduction to Data Science. A Python Approach to Concepts,
[2] Techniques and Applications. Barcelona, España, 2017.
[3] S. Ozdemir, Principles of data science: learn the techniques and math you need to start making sense of your data. 2016.
[4] F. Cady, The Data Science Handbook, 1st. United States of America, 2017.
[5] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electron. Mark., vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/S12525-02100475-2/TABLES/2.
[6] 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.
[7] L. Diedrichs, F. Bromberg, D. Dujovne, K. Brun-Laguna, and T. Watteyne, "Prediction of frost events using Bayesian networks and Random Forest," 2018.
[8] "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-porheladas-tardias-alcanzaron-las-30-mil-hectareas-de-vid-y-16-mil-de-frutales-enmendoza-en-2020/ (accessed Apr. 14, 2022).
[9] "Alerta entre los productores de Cuyo por las heladas tardías - Revista InterNos," 2021.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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].
[15] Available: https://drive.google.com/file/d/1byaIS-ssvJPSMHtKP9ahu8LQuoshyq6/view.
[16] 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.
[17] J. L. F. Yagüe, Iniciación a la meteorología y climatología. España, 2012.
[18] 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.
[19] F. Chollet, Deep Learning with Python, 1st ed. USA: Manning Publications Co., 2017.
[20] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Comput. Surv., vol. 52, no. 4, Aug. 2019, doi: 10.1145/3343440.
[21] C. Aggarwal, Neural networks and deep learning: a textbook, 1st ed. Cham, Switzerland: Springer Nature Switzerland AG, 2018.