# Procesamiento de datos meteorológicos para determinar la ocurrencia, intensidad y duración de heladas.

Joaquin Cortez, María Masanet, Raul Klenzi, Manuel Ortega

Instituto de Informática / Departamento de Informática /Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan

{joaquincortez1999, mimasanet, rauloscarklenzi, Manuel.ortega}@gmail.com

**Resumen.** En este trabajo se abordará el procesamiento de datos meteorológicos provenientes de agro sensores situados en la provincia de San Juan, Argentina. Se analizará el impacto de diferentes variables meteorológicas en la ocurrencia, intensidad y duración de las heladas. El objetivo que se persigue es predecir si en un día determinado ocurrirá una helada, que tan intensa será y cuál será su duración para poder alertar al productor agropecuario acerca de este fenómeno, a fin de activar los mecanismos de mitigación correspondientes. Se comienza el escrito haciendo una pequeña introducción y destacando el contexto desde donde surge la presente propuesta, a continuación se describen brevemente los datos, qué enfoque se utilizó para agruparlos, los desafíos encontrados, y posteriormente se presentará un análisis exploratorio para determinar la relación entre las diferentes variables y su impacto en las heladas. Finalmente se aplicarán distintos modelos de aprendizaje de máquina, trabajando las predicciones mediante clasificadores y regresores para poder llegar a anticipar la ocurrencia y las características de la helada.

**Keywords:** Machine Learning, Random Forest, Agricultura Inteligente.

## 1 Introducción y contexto

La presente propuesta surge en el marco de trabajo compuesto por los proyectos de investigación llevados adelante en el ámbito del Departamento e Instituto de Informática de la Facultad de Ciencias Exactas Físicas y Naturales DI-IdeI-FCEFN y en el Instituto de Automática de la Facultad de Ingeniería INAUT-FI de la Universidad Nacional de San Juan UNSJ.

Uno de los proyectos involucrados “Evaluación de visualizaciones eficientes en Ciencia de Datos” comienza su tercer año de ejecución en el ámbito del Laboratorio de Sistemas Inteligentes para Extracción de Conocimiento en Datos Masivos DI-IdeI-FCEFyN vinculando sinérgicamente el accionar de docentes investigadores y alumnos pertenecientes a las carreras de la Licenciatura en Sistemas de Información LSI y Licenciatura en Ciencias de la Computación LCC y cuenta con el aval y subsidio del Consejo de Investigaciones Científicas y Técnicas y de Creación Artística CICITCA-UNSJ. La propuesta cuenta con antecedentes logrados en el tema conforme a sucesivos

proyectos aprobados y subsidiados por el ente mencionado en los que el grupo, conformado por docentes responsables de diferentes asignaturas que dan soporte al área de la Ciencia de Datos (Data Science-DS), viene trabajando desde el año 2010. Así mismo el proyecto llevado adelante en el ámbito del INAUT-FI y con el cual se han desarrollado tareas conjuntas es el PIO84 "Telemetría Agrícola" y desde su interacción con el medio agroproductivo sanjuanino se han obtenido y se están procesando datos de estaciones meteorológicas de la provincia correspondientes a la estación INTA Pocito y a la del establecimiento privado San Francisco en el departamento Sarmiento distantes ambos 37Km. Estos datos son representados como series temporales y sobre este tipo de datos se trabaja en la presente propuesta.

Concretamente una de las problemáticas asociadas con el clima tiene que ver con la predicción de heladas y particularmente las heladas tardías en meses de octubre y noviembre, que afectan a los cultivos en plena floración llegando a ocasionar pérdidas totales. Poder predecir esta condición climática y con ello alertar al productor a efectos de que se tomen las medidas de mitigación y así atenuar o evitar pérdidas, es un aporte relevante en que está trabajando el grupo de investigación.

En el caso de la predicción de heladas es importante destacar el momento en que se producirá y la duración e intensidad de la misma a efectos de ajustar adecuadamente las diferentes instancias de mitigación.

Los datos que se disponen son los referidos entre otros a: Temperatura, Humedad, Velocidad del viento, Radiación Solar, Punto de Rocío, desde el año 2013 al 2019 que las estaciones meteorológicas capturan cada 10 minutos.

Atento a que hay cultivos que no son factibles de producir en invernaderos, la problemática esencialmente apunta a tratar de lograr buenos modelos de predicción de intensidad y duración del fenómeno de la helada, objetivo de este trabajo. Respecto al momento en el cual estimativamente se producirá la helada se realiza conforme lo publicado en [1], como para que el productor active los elementos de mitigación correspondiente (riego, calentadores etc…) y atenuar así, los posibles daños en los cultivos [2]. Para ello la bondad del modelo debe ser tal que los Falsos Positivos (predicción de heladas que no se producen) o Falsos Negativos (predecir que no habrá heladas cuando en realidad se producen) se reduzcan cuanto sea posible. El saber la intensidad y duración de la helada tiene que ver finalmente, con que el productor administre adecuadamente los recursos asociados a las instancias de mitigación como depósitos de agua o combustible.

## 2 Características de los datos

### 2.1 Procedencia de los datos

Los datos provienen de agro sensores ubicados en el establecimiento privado San Francisco, departamento de Sarmiento, en la provincia de San Juan, Argentina, desde el 11 de abril de 2013 hasta el 21 de marzo de 2019.

### 2.2 Atributos

Los datos son captados cada diez minutos y contienen las siguientes variables: *Fecha, Hora, Temp. Ext., Temp. Max., Temp. Min., Hum. Ext.,Pto. Rocio, Vel. Viento, Dir. Viento, Rec. Viento, Vel. Max.,Dir. Max., Sens. Term., Ind. Calor, Indice THW , Indice THSW , Bar , Lluvia, Int. Lluvia, Rad. Solar, Energia Solar, Max. Rad. Solar , Grad.D. Calor , Grad.D. Frio , Temp. Int., Hum. Int., Rocio Int., In.Cal. Int., ET , Muest Viento, Tx Viento, Recep. ISS* . Estas hacen referencia a fecha y hora, temperatura, humedad, punto de rocío, radiación solar, lluvia, dirección del viento, velocidad del viento, entre otros. Posteriormente se profundizará en las diferentes columnas del dataset.

### 2.3 Tamaño

El conjunto de datos tiene un total de 299671 registros y 42 columnas.

### 2.4 Datos faltantes

Se observaron datos faltantes en las siguientes variables: dirección del viento y dirección de viento máxima, debiendo destacar que las mismas se obtienen de la dirección dominante en los últimos 30 minutos de los atributos velocidad de viento y velocidad máxima de viento respectivamente, de todas maneras por ser variables de tipo categórico fueron filtradas atendiendo al procesamiento de los datos de tipo numérico a modelar; ;para índice THSW (una forma de medir la sensación térmica) y el punto de rocío, aún siendo variables numéricas no se tuvieron en cuenta para el modelo.

Por otro lado, se pudo obtener que hay periodos donde no se captaron datos. Se encontraron siete “huecos”, de una duración promedio de 12 días y 16 horas. Posteriormente, se profundizará en cómo se abordó este problema.

## 3 Preprocesamiento de los datos

### 3.1 Selección de variables relevantes

Los datos captados contienen información redundante para los objetivos de este trabajo, la cual se eliminó a fin de lograr modelos de predicción más eficientes. Por ejemplo, se registra la temperatura exterior además, la máxima y mínima producida en el intervalo de 10 minutos, obteniendo tres mediciones, que en la mayoría de los casos son iguales. Algo similar sucede con otras variables como la humedad. Las mediciones redundantes para estas variables fueron eliminadas, seleccionando una única medición para las mismas.

Se realizó un análisis de correlación para determinar cuáles variables son redundantes conforme su alto grado de dependencia y de las cuales se puede prescindir. Todas las variables con un índice de correlación mayor a 0.9 fueron eliminadas. Como una alternativa de rápida visualización de interdependencia entre variables, la Fig1 presenta una matriz de calor con rojo intenso especificando una correlación de 1 entre

variables, en este caso, cuando una variable crece la otra lo hace en la misma proporción. En tanto en color azul intenso cuando permite representar cuando la correlación es inversa -1 de modo que ante un aumento en una variable la otra decrece en la misma proporción.

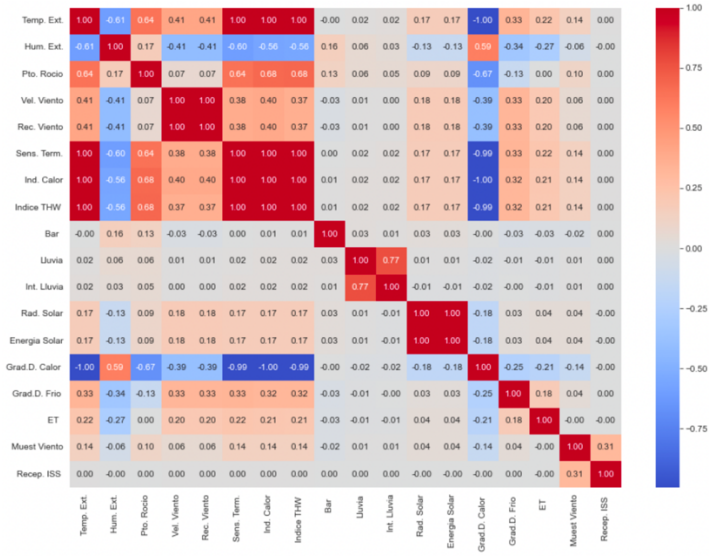


**Fig. 1.** Matriz de calor de la correlación entre las distintas variables del conjunto de datos, indicando el valor de esta correlación.

### 3.2 Abordaje de datos faltantes

Como se mencionó en la sección 1.4 existieron tanto registros faltantes para algunas variables, así como también períodos en los que no se capturaron datos producto de fallas en los equipos sensores de registración. Para estos casos, se empleó una interpolación lineal hacia adelante para obtener estos valores.

Para ejemplificar esta interpolación, se muestra en la figura 2 el hueco en la serie temporal correspondiente a la temperatura más grande de todo el conjunto de datos. En la figura 3, se muestra la gráfica correspondiente a los datos posteriormente a realizar la interpolación. Se puede observar que claramente hay una pérdida de información, sin embargo, esto ocurrió en un periodo de tiempo que no consideramos relevante, dado que no se ubica en épocas en las que se pueda producir una helada. Por lo tanto, estos datos interpolados no se utilizarán como entrada de los modelos predictivos que se plantean posteriormente en este trabajo.

Puede parecer innecesario realizar la interpolación de datos dentro de un periodo de tiempo que no será considerado, sin embargo es necesario dado que la interpolación es uno de los primeros pasos del preprocesamiento que se realiza [7], luego de haber filtrado las columnas. Esto se hace para facilitar el procesamiento posterior de la

información, ya que las librerías utilizadas para el abordaje de series temporales asumen que las muestras se encuetran en periodos igualmente espaciados de tiempo.

### 3.3 Abordaje del desbalance en los datos

Dado que el objetivo de este trabajo es abordar los casos en donde suceda helada, nos encontramos con el problema del desbalance de datos. Del total de 299671 registros, tan sólo 13069 corresponden a periodos de helada, es decir, poco más del 4%. Para abordar este problema se consideraron únicamente los períodos en los que ocurren heladas.

Realizando este filtrado de los datos, resultan un total de 66342 registros, de los cuales 12475 corresponden a heladas. Si bien las heladas ahora suponen un 18% de los registros, aún persiste cierto desbalance entre las dos categorías a predecir. Sin embargo, en trabajos realizados anteriormente [2] se observó que distintas técnicas de sobremuestreo no suponen una mejora significativa en los modelos para este conjunto de datos.

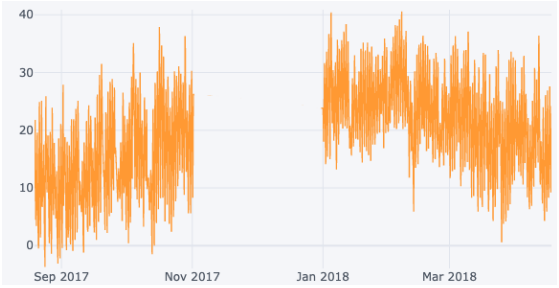


**Fig. 2.** Gráfica correspondiente a los registros de temperatura entre Septiembre de 2017 y Abril de 2018, antes de realizar la interpolación de los datos.

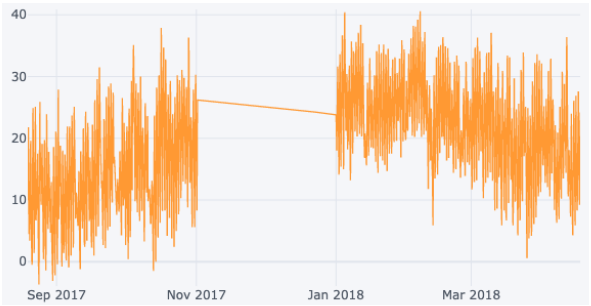


**Fig. 3.** Gráfica correspondiente a los registros de temperatura entre Septiembre de 2017 y Abril de 2018, luego de realizar la interpolación de los datos.

**Table 1.** Fechas de primera y última helada por año

| Año | Primera helada | Última helada |
|---|---|---|
| 2013 | 2013-05-15 | 2013-09-29 |
| 2014 | 2014-05-18 | 2014-09-12 |
| 2015 | 2015-05-03 | 2015-09-17 |
| 2016 | 2016-04-26 | 2016-10-21 |
| 2017 | 2017-05-20 | 2017-10-13 |
| 2018 | 2018-05-20 | 2018-10-02 |

# 4 Procesamiento de los datos

## 4.1 Procesamiento en trabajos anteriores

En trabajos anteriores, para predecir la ocurrencia de heladas se utilizaron ventanas temporales de 3 horas, lo que significa que el modelo toma como entrada los 18 registros anteriores - 3 horas multiplicadas por un registro cada 10 minutos- y en base a eso predecirá si ocurrirá o no helada en un horizonte de 3 horas hacia delante[1]. tiempo en cual el productor puede tomar medidas de mitigación correspondiente.

Para este trabajo, donde además de la ocurrencia de la helada también se quiere determinar su duración y su intensidad, se utilizó un enfoque distinto en el abordaje previo del procesamiento de los datos [3].

## 4.2 Agrupación de los datos

Los datos fueron procesados de la siguiente forma:

Se agruparon los datos en “días lógicos”. Un día lógico inicia a las 10 AM de un día real y termina a las 9:59 AM del día real siguiente. Luego, se separó este conjunto de datos agrupados, en dos conjuntos distintos.

Un primer conjunto con los datos en horarios en los que no ocurren heladas, que son de 10 a 19:59, el cual posee el promedio para las distintas variables meteorológicas durante esas horas [3]. Estos son los datos que el modelo utilizará como entrada.

El segundo conjunto, con los datos en horarios en los que ocurren heladas, es decir, entre las 20 y las 9:59, que contendrá la temperatura mínima en ese rango de tiempo, la intensidad de la helada y la duración de la helada. Estos valores son los que el modelo intentará predecir.

Se utilizaron las temperaturas medias para los horarios de día con el objetivo de lograr un sistema que, basado en la media de las variables meteorológicas de un determinado día, llegada las 19:00 hs, sea capaz de predecir si ocurrirá una helada, que tan intensa será y que duración tendrá. Por otro lado, la razón por la que se tomó la temperatura mínima para los horarios de helada, es dado que el valor mínimo para dicha variable en ese periodo de tiempo determinará si ocurrirá helada y que intensidad tendrá [8].

En cuanto a la intensidad de la helada, la misma está dada por la temperatura mínima que alcanza la helada. La duración, por otro lado, se obtuvo contando la cantidad de registros entre la primera y la última temperatura bajo cero en ese rango de tiempo y atento a que los datos se registran cada 10 minutos, se multiplicó por diez para obtener la duración.

Si bien existen casos en los que la temperatura cae por debajo de los cero grados y luego sube nuevamente, se consideró oportuno despreciar estos casos, debido a que en su mayoría lo que sucedía era que la temperatura subía unas pocas décimas por arriba de cero durante una escasa cantidad de registros y luego bajaba nuevamente. Desde el punto de vista del agricultor, carece de sentido práctico detener por completo todos los mecanismos de mitigación ante estos casos.

La Fig.4 presenta la cantidad de registros con heladas por hora a lo largo de los años relevados. Así a la 0hs durante esos años, algo más de 500 fueron los registros con heladas, en tanto a las 7hs, superaron los 1.750. Se aprecia también que no se contabilizaron registros con heladas entre las 11hs y las 18hs.

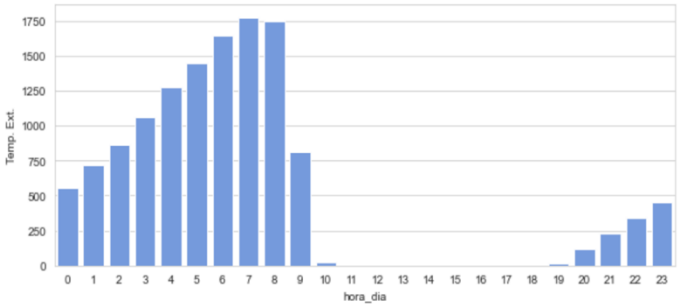


**Fig. 4.** Cantidad de registros de heladas por hora

### 4.3 Modelos utilizados

Para la predicción se utilizaron tanto modelos regresores como clasificadores. Dado que los modelos clasificadores requieren de la existencia de clases o etiquetas no continuas o discretas[4], en las cuales se enmarcarán los resultados, nace la necesidad de crear distintas clases considerando los objetivos de predecir ocurrencia, intensidad y duración de la helada. Por otro lado, los modelos regresores arrojarán un resultado numérico acerca de cuánto durará la helada, que temperatura mínima alcanzará, determinando así la ocurrencia o no, así como también su intensidad. Los algoritmos de los modelos utilizados en conjunto con sus hiperparámetros serán descriptos en secciones posteriores.

**Temperaturas: ocurrencia de helada.**

Para la clasificación respecto a la ocurrencia de helada, es claro que existen dos clases posibles: ocurre helada o no ocurre helada. Para esto se utilizarán dos clases discretas (Verdadero, Falso) como se muestra en la Tabla 2 para la salida.

Se observó que a partir de esta clasificación, ambas clases se encuentran balanceadas, donde los datos correspondientes a registros de heladas representan un 43% del total.

**Table 2.** Clasificación de helada y cantidad de registros para cada clase

| Clase | Valores de temperatura | Frecuencia |
|---|---|---|
| Verdadero | < 0 | 383 |
| Falso | ≥ 0 | 480 |

**Duración de las heladas.**

. Similarmente a lo que ocurre con las temperaturas, para las estrategias mediante clasificadores es necesaria la existencia de clases. Considerando las necesidades del productor agropecuario que le interesará saber que tan extensa será la helada, surge una clasificación en tres tipos: breve, media y extensa, y a su vez para aquellos registros que no corresponden a helada, se determina clase nula, que significará duración 0. El criterio que se utilizó para establecer la cantidad de etiquetas asociadas al clasificador de la duración, fue que cada uno de estos grupos mencionados contengan aproximadamente la misma cantidad de registros, con la finalidad de que el conjunto de datos se encuentre balanceado y así también poder establecer valores numéricos límite que impliquen qué duración corresponde a qué clase, especificado en la tabla 3. Para lograr esto se analizó la distribución de las duraciones de las heladas, la cual se evidencia en la figura 5.

**Table 3.** Clasificación de duración con valores expresados en minutos.

| Clase | Desde | Hasta | Frecuencia |
|---|---|---|---|
| BREVE | 10 | 209 | 130 |
| MEDIA | 210 | 449 | 117 |
| EXTENSA | 450 | 840 (máx.) | 116 |

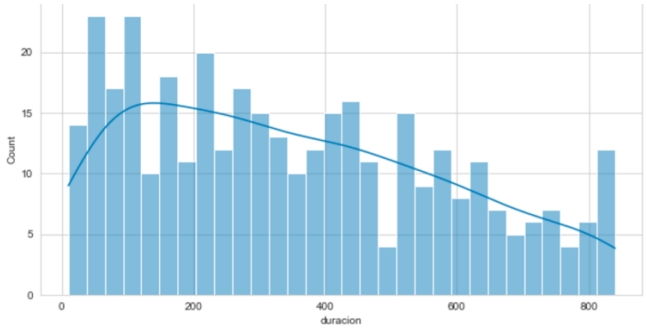


**Fig. 5.** Distribución de la duración de las heladas.

**Intensidad de las heladas.**
Continuando lo mencionado para la clasificación de la duración de las heladas, surge la clasificación para la intensidad de la helada, considerando esta vez las temperaturas mínimas alcanzadas. Atendiendo tanto a la necesidad de los modelos clasificadores de un conjunto finito de etiquetas dentro las cuales proyectar sus resultados, como a la necesidad del productor agropecuario de establecer distintos niveles de intensidad para la helada, se proponen las siguientes clases: baja, media y alta, evidenciadas en la tabla 4, y como sucedía para la duración, aquellos registros en los que no hubiera helada, también se utilizará la clase nula, en este caso indicando que la temperatura fue mayor a cero. Como sucedía anteriormente, para esto se analizó la distribución de las temperaturas para aquellos registros en los que se presentara helada, y a partir de esto se determinaron las cotas que enmarcan en qué clase de intensidad corresponde una helada, esta distribución se muestra en la figura 6.

**Table 4.** Clasificación de la intensidad de heladas, con valores expresados en grados celcius.

| Clase | Desde | Hasta | Frecuencia |
|---|---|---|---|
| BAJA | -0.1 | -1.4 | 119 |
| MEDIA | -1.5 | -3.4 | 123 |
| ALTA | -3.5 | -9.4 (mín.) | 121 |

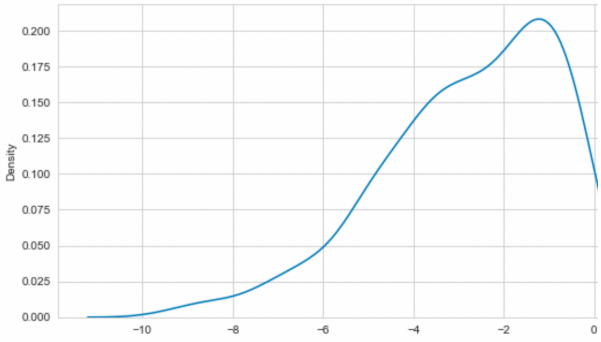


**Fig. 6.** Distribución de las temperaturas para los registros correspondientes a heladas.

# 5 Modelos predictivos

## 5.1 Separación de datos de entrenamiento y pruebas

Se utilizaron los datos correspondientes a los años 2013 al 2016 para entrenar los modelos y los datos de 2017 y 2018 para probarlos.

Los datos para entrenamiento suponen un total de 576 registros, mientras que los datos de prueba un total de 279. Esto implica que un 67% de los datos se utilizarán para entrenamiento y un 33% para pruebas.

### 5.2 Modelos e hiper parámetros

Tal como se mencionó anteriormente, se abordaron las predicciones como problemas de regresión y de clasificación. En ambos casos se utilizó como algoritmo de modelación *Random Forest,* dada su buena expresividad y performance mostrada en trabajos anteriores [2] y también porque es un modelo relativamente simple si es comparado con, por ejemplo, redes neuronales recurrentes de tipo LSTM, y dada la baja cantidad de datos luego de haber realizado el preprocesamiento, se prefiere utilizar un algoritmo de las características de Random Forest para evitar un sobreajuste [6].

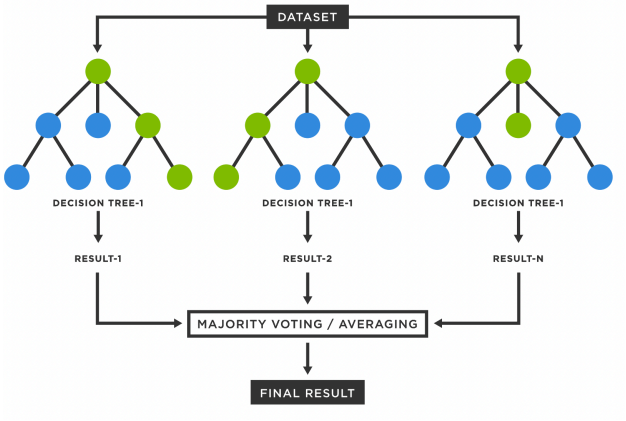


**Fig. 7.** Esquema de funcionamiento de la estrategia Random Forest [12].

Los hiper parámetros utilizados fueron:

- Número de árboles: 1.000.
- Criterio: gini para los problemas de clasificación y error cuadrático para los problemas de regresión.
- Cantidad mínima de hojas: 2.

### 5.3 Resultados de los modelos

**Predicción de temperaturas y ocurrencia de heladas.**
El modelo arrojó un $R^2$ de 0.60 y un error medio cuadrático de 9.63.
Clasificando los resultados en “helada” y “no helada”, según si la temperatura predicha es menor o mayor a cero, respectivamente.

Como se puede apreciar en la tabla 5, observando particularmente el valor obtenido para exhaustividad, que nos permite ver cuál es el comportamiento de los modelos respecto a los falsos negativos, que es de nuestro máximo interés reducir, es claro que el modelo hace un muy buen trabajo.

**Table 5.** Métricas para el modelo de regresión a través de random forest, clasificando las temperaturas obtenidas.

| Clase | Precisión | Exhaustividad (recall) | F1 Score | Exactitud |
|---|---|---|---|---|
| No helada | 0.84 | 0.80 | 0.82 | 0.80 |
| Helada | 0.76 | 0.79 | 0.77 | |

**Table 6.** Métricas para el modelo de clasificación a través de random forest.

| Clase | Precisión | Exhaustividad (recall) | F1 Score | Exactitud |
|---|---|---|---|---|
| No helada | 0.82 | 0.83 | 0.83 | 0.81 |
| Helada | 0.80 | 0.79 | 0.79 | |

Se puede observar de las tablas 5 y 6 que los resultados son prácticamente idénticos, usando regresión y clasificación.

**Predicción de intensidad de heladas**

. Haciendo una clasificación en base a la intensidad de las temperaturas obtenidas por el algoritmo de Random Forest regresor, se obtuvieron las siguientes métricas observadas en la tabla 7.

**Table 7.** Métricas para la predicción de temperaturas por regresor Random Forest, clasificadas en base a la intensidad.

| Clase | Precisión | Exhaustividad (recall) | F1 Score | Exactitud |
|---|---|---|---|---|
| NULA | 0.84 | 0.80 | 0.82 | 0.60 |
| BAJA | 0.34 | 0.28 | 0.31 | |
| MEDIA | 0.36 | 0.21 | 0.27 | |
| ALTA | 0.27 | 0.83 | 0.41 | |

Se puede ver que, la determinación en clases tiene margen de mejora atento a que la exactitud del modelo no es suficientemente buena,. Sin embargo, se aprecia que la exhaustividad es elevada para las heladas de alta intensidad, lo cual es positivo a la hora de identificar heladas graves con anticipación.

Por otro lado, las métricas obtenidas por Random Forest clasificador se encuentran en la tabla 8.

**Table 8.** Métricas para la predicción de temperaturas por clasificador Random Forest.

| Clase | Precisión | Exhaustividad (recall) | F1 Score | Exactitud |
|---|---|---|---|---|
| NULA | 0.87 | 0.78 | 0.82 | 0.66 |
| BAJA | 0.08 | 0.18 | 0.11 | |
| MEDIA | 0.27 | 0.26 | 0.26 | |
| ALTA | 0.71 | 0.70 | 0.71 | |

En este caso, se observa que la exactitud del modelo sube, y existe una notable variabilidad en las distintas métricas para cada clase respecto al modelo regresor.

**Predicción de duración de heladas.**
Utilizando un algoritmo de Random Forest regresor, este arrojó un $R^2$ de 0.16 y un error medio absoluto de 116 minutos. Los resultados de la clasificación de los valores obtenidos se encuentran en la tabla 9.

**Table 9.** Métricas obtenidas para la predicción de la duración por regresor Random Forest.

| Clase | Precisión | Exhaustividad (recall) | F1 Score | Exactitud |
|---|---|---|---|---|
| NULA | 0.07 | 1.00 | 0.12 | 0.27 |
| BREVE | 0.75 | 0.16 | 0.26 | |
| MEDIA | 0.44 | 0.27 | 0.33 | |
| EXTENSA | 0.46 | 0.65 | 0.54 | |

Por otro lado, utilizando un Random Forest clasificador se obtuvieron las métricas presentes en la tabla 10.

**Table 10.** Métricas obtenidas para la predicción de la duración por clasificador Random Forest.

| Clase | Precisión | Exhaustividad (recall) | F1 Score | Exactitud |
|---|---|---|---|---|
| NULA | 0.86 | 0.77 | 0.81 | 0.63 |
| BREVE | 0.08 | 0.23 | 0.12 | |
| MEDIA | 0.33 | 0.30 | 0.31 | |
| EXTENSA | 0.86 | 0.77 | 0.81 | |

Como se puede observar de las tablas 9 y 10, abordando la predicción de la duración como un problema de clasificación, la exactitud del modelo mejora notablemente. Y en ambos modelos se mantuvo la tendencia de que mientras más extensa es la helada, mejor es la predicción. Esto puede deberse a que para el caso de las heladas breves, se deban a heladas de intensidad baja y de corta duración, las cuales se clasificaron erróneamente como nulas.
Tanto a nivel intensidad como de duración se destaca el buen rendimiento del modelo. Esto se debe a que la dinámica de los sistemas térmicos ante una helada intensa su posterior retorno a temperatura sobre 0 grados requiere de un mayor tiempo de duración.

## 5.4 Combinación de modelos

Tal como se mencionó en secciones anteriores, en trabajos previos se utilizó un esquema de ventanas de tiempo, para poder predecir la hora en la que sucederá la helada. Por otra parte, el modelo de este trabajo permite indicar si ocurrirará una helada y que características tendrá, pero no determina la hora a la que ocurrirá, lo cual es algo necesario para el agro productor a fín de establecer los mecanismos de mitigación en el

horario correspondiente a la helada, ya que considerando únicamente el modelo que predice que puede ocurrir una helada en un rango de 14hs, estos mecanismos deberían estar activos durante todo este rango de tiempo, lo cual no es eficiente. Es por esto que se propone complementar el modelo presentado en el trabajo actual con el modelo presentado en trabajos anteriores [2]. De esta forma se tienen las ventajas del modelo actual, donde la exhaustividad es más alta, esto es, posee una menor cantidad de falsos negativos, a la vez que se puede obtener el horario específico en el que ocurrirá la helada. La Fig. 8 muestra la aplicación conjunta de ambos modelos.

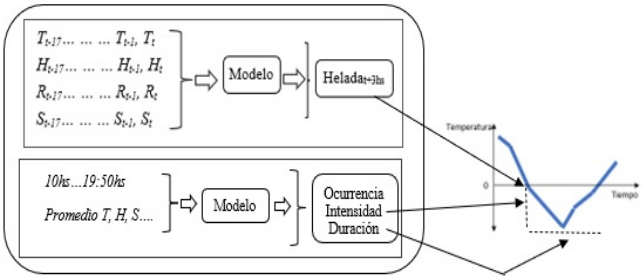


**Fig. 8.** Implementación de modelos predictores de heladas

## 6 Conclusiones y trabajos a futuro

En este trabajo se presentó un método de procesar los datos para poder realizar un análisis predictivo acerca de la ocurrencia, intensidad y duración del fenómeno climático de la helada.

Para esto se mostró el formato de los datos, la cantidad y su distribución. Así mismo, se filtraron las columnas redundantes mediante un análisis de correlación, se realizó una interpolación de los datos faltantes y se abordó el problema del desbalance del conjunto de datos original considerando únicamente los períodos de helada.

Posteriormente se propuso y se llevó a cabo un enfoque de procesamiento de datos que consistió en agrupar los mismos por día, con promedio de las diferentes variables meteorológicas durante las horas en las que no ocurren helada y la temperatura mínima registrada para las horas en las que ocurren heladas, así como también la duración, en caso de ocurrir.

También se realizó un análisis sobre los datos ya procesados, para conocer la distribución de las variables que los modelos intentarán predecir y determinar clases asociadas.

Finalmente, se entrenaron modelos de Random Forest, abordando la predicción tanto como clasificación, como regresión, para determinar la ocurrencia, intensidad y duración de la helada. Se obtuvo que el modelo tiene una performance aceptable determinando la ocurrencia de heladas en ambos modelos, los cuales arrojaron métricas esencialmente iguales. Respecto a la predicción de la intensidad ambos modelos tuvieron un rendimiento regular donde el clasificador fue ligeramente mejor al regresor.

Y en cuanto a la predicción de la duración, el modelo de regresión tuvo una mala performance, mientras que el modelo clasificador arrojó resultados mucho mejores. A su vez, se planteó la combinación del modelo predictivo desarrollado en este trabajo con modelos utilizados en trabajos anteriores.
Todo este análisis indica que esta propuesta introductoria es perfectible al mejorar las métricas que miden la performance de los modelos, sea a través de otros algortimos y/o la incorporación de nuevos registros asociados a los últimos años calendario.
A su vez, como trabajo a futuro se plantea implementar estrategias de balanceo de datos para ser utilizados en este modelo y la implementación de la combinación de modelos planteada.

## 7 Referencias bibliográfica


1. 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.
2. 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.
3. 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.
4. L. Igual and S. Seguí, *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*. Barcelona, España, 2017.
5. C. O'Neil and R. Schutt, *Doing Data Science*. O'Reilly Media, Inc., 2013.
6. F. Chollet, *Deep Learning with Python*, 1st ed. USA: Manning Publications Co., 2017.
7. J. Luengo, F. Herrera, S. Garcia, Data Preprocessing in Data Mining. 1st edn. Springer Cham (2014).
8. Verdes, Granitto, Navone, and Ceccatto: Frost Prediction With Machine Learning Techniques. pp 3-4. (2000).
9. Sitio web de la biblioteca Pandas, https://pandas.pydata.org, accedida por última vez 21 de junio de 2022.
10. Adams, Char (2014) Learning Python Data Visualization. PacktPublishing. ISBN 978-1-78355-333-4.
11. Aureì lien Geὶ ron: Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2nd edn. O'Reilly. (2019)
12. What is a Random Forest?, https://www.tibco.com/es/reference-center/what-is-a-random-forest. , accedida por última vez 21 de junio de 2022.