Predicción de la precipitación diaria en Chicago usando modelos estadísticos

Proyecto investigación Estadística II

Marco Antonio Mejía Elizondo Amel Cáceres Cruz Esteban Loría Salas

Resumen

El presente trabajo pretende estudiar la relación entre variables climáticas en la ciudad de Chicago, Illinois, Estados Unidos. Específicamente, se plantea estudiar la capacidad de predecir precipitación líquida en Chicago observando otras variables climatológicas y utilizando tres modelos de regresión/clasificación: regresión lineal (cuantitativa), máquinas de soporte vectorial, árboles de decisión (cualitativa). Adicionalmente se muestra un caso ilustrativo de implementación de análisis de discriminante lineal.

Introducción

La lluvia es tal vez la variable del clima más relevante: afecta los cultivos y las actividades económicas y sociales humanas, es fuente de agua para todos los seres vivos y puede ser la causante de desastres con pérdidas económicas cuantiosas. Naturalmente surge la pregunta si es posible predecir cuándo va a llover y con qué magnitud se puede manifestar, pero también si es posible determinar esas incógnitas a través de observar los datos pasados del clima en una región específica. Aquí entra la estadística, la cual ofrece múltiples modelos de predicción de diferente naturaleza y con diferentes fundamentos matemáticos y supuestos. Este trabajo se inspira en querer poner a prueba algunos de esos modelos para un caso en concreto y observar los resultados.

Para poder obtener buenos resultados estadísticos es necesario contar con datos con cierto nivel de calidad. Sin embargo, los datos del clima que son reportados por instituciones o centros especializados no siempre son de fácil acceso al público. Por esta razón para este trabajo se decidió seleccionar una ciudad poblada y famosa de Estados Unidos, como lo es Chicago, con la esperanza de que los datos reportados y que son de acceso fácil sean fieles a los eventos ocurridos.

De esta manera el objetivo general del trabajo es evaluar la posibilidad de predecir la precipitación diaria que cae en la ciudad de Chicago mediante la utilización de tres modelos de regresión/clasificación: un modelo cuantitativo, regresión lineal; y dos modelos cualitativos, máquina de soporte vectorial y árboles de decisión.

Para llevar a cabo la valoración de los modelos primero se examina la relación observada en los datos entre la variable precipitación y otras variables climatológicas (temperatura, viento, presión atmosférica...) en Chicago. Luego se procede a seleccionar y calibrar modelos de regresión lineal, máquina de soporte vectorial y árboles de decisión utilizando los datos recolectados en la ciudad de Chicago durante el periodo 01/05/2008 - 01/05/2018. Seguidamente se ponen a prueba los modelos calibrados usando los datos del periodo 01/05/2018 - 01/05/2019 y se muestran los resultados dependiendo del modelo. Para el modelo de regresión lineal se indica el error cuadrático obtenido y para los modelos cuantitativos se comparan los resultados usando estadísticos como la exactitud y la curva ROC. Por último, se agrega un caso de implementación de análisis de discriminante lineal con un propósito ilustrativo y para comparar con los demás modelos cuantitativos.

En la literatura se pueden encontrar bastantes trabajos académicos que enfrentan el problema de pronosticar variables climáticas con modelos estadísticos y series de tiempo. Por ejemplo, en el trabajo de Cramer et al. (2017) se realiza una valoración de siete métodos de machine learning que incluye programación genética y redes neuronales, para la predicción de lluvia dentro del contexto de instrumentos de inversión derivados del clima. En Wilks (1999) se investiga la capacidad de representar variabilidad anual de precipitación con varios modelos estocásticos incluyendo cadenas de primer orden de Markov y modelos de distribución gamma. En

Pérez-Vega et al. (2016) se propone un modelo de máquinas de soporte vectorial para realizar pronóstico de temperatura.

Marco teórico

El área de estudio del trabajo es el clima y este engloba las condiciones de la atmósfera sobre una región y en un período de tiempo determinado, indicando también su variabilidad ("Clima" 1988). El objeto central es la lluvia o el agua que procede de la atmósfera en forma líquida o sólida y se deposita en la tierra ("Lluvia" 1988). Ya que el trabajo se basa en la existencia de relación o inclusive causalidad entre la lluvia y otras variables climatológicas es necesario explicar dichas variables.

La temperatura es una medida de la energía promedio cinética o velocidad de las moléculas en un cuerpo o sustancia (Bramer, Wojtowicz, and Halls 2010). En el estudio del clima se analiza la temperatura del aire, del océano y del suelo.

El viento es la corriente atmosférica de aire que se mueve en dirección determinada y que se origina por las diferencias de la temperatura de la atmósfera en distintos puntos de la superficie terrestre. La velocidad de viento sostenida es la velocidad del viento determinada por el promedio de valores observados en períodos de uno o dos minutos (National Weather Service, n.d.a).

Una ráfaga de viento es una fluctuación rápida en la velocidad del viento con una variación de diez o más nudos (5.14 metros por segundo) entre un incremento y decremento momentáneos. La velocidad de la ráfaga es la máxima velocidad instantánea del viento (National Weather Service, n.d.a).

El punto de rocío indica la cantidad de humedad en el aire y es la temperatura a la cual el aire se debe enfriar (con presión constante) para que alcance saturación. Un estado de saturación se da cuando el aire contiene la máxima cantidad de vapor de agua posible. Un punto de rocío mayor implica mayor humedad en el aire, dado una temperatura y presión. En condiciones normales el punto de rocío no debería ser mayor que la temperatura del aire (Bramer, Wojtowicz, and Halls 2010).

La presión atmosférica es la presión ejercida por la atmósfera en un punto como consecuencia de la atracción gravitacional que afecta la columna de aire sobre dicho punto (National Weather Service, n.d.b). La presión al nivel del mar es la presión atmosférica al nivel promedio de mar (National Weather Service, n.d.b). La presión estacional es la presión que se observa a una elevación específica (National Weather Service, n.d.a).

La visibilidad es la distancia a la cual un objeto dado puede ser visto e identificado usando solamente el ojo (National Weather Service, n.d.a). Este fenómeno se puede relacionar con la presencia de neblina.

La nieve es la precipitación en forma de cristales de hielo formada por el congelamiento del vapor de agua en el aire (National Weather Service, n.d.a). Es importante considerar que dada una cierta condición atmosférica es posible observar precipitación de nieve parcial y de agua líquida al mismo tiempo.

El muestreo de estas y otras variables se lleva a cabo en múltiples estaciones de medición que emplean instrumentos como el pluviómetro, termómetro y anemómetro.

Para entender el mecanismo de generación de precipitación se estudia el proceso de colisión-coalescencia ("Condensation nucleus" 2017). En este proceso la lluvia se forma ante la presencia de núcleos de condensación de diferentes tamaños. Núcleos de condensación son partículas diminutas suspendidas en las cuales el vapor de agua se condensa en la atmósfera ("Condensation nucleus" 2017). Las gotitas de agua formadas por núcleos de condensación grandes caen y colisionan con las gotitas formadas por los núcleos pequeños lo que provoca que se fusionen formando gotas más grandes. Cuando las gotas caen, la resistencia en el aire "rompe" la unión de aquellas que no tienen un tamaño suficiente, pero aquellas formadas por una gran cantidad de colisiones se precipitan hasta el suelo. Este modelo aplica para nubes cálidas.

En el estudio de pronóstico del clima se desarrollan, en su mayoría, modelos numérico-estadísticos que emplean como base una serie de ecuaciones diferenciales derivadas a partir de propiedades físicas (Pu and Kalnay 2019). Estas ecuaciones son: la segunda ley de Newton y la ecuación del movimiento, la ecuación de conservación de la masa del aire, la ecuación del estado ideal de los gases, la primera ley de la termodinámica

o conservación de la energía y la ecuación de conservación de la masa de agua. Dichas ecuaciones tienen la siguiente forma (Pu and Kalnay 2019):

$$\frac{d\overrightarrow{V}}{dt} = -\alpha\overrightarrow{\nabla}p - \overrightarrow{\nabla}\Phi + \overrightarrow{F} - 2\Omega \text{ x } \overrightarrow{V} \tag{1}$$

$$p\alpha = RT \tag{2}$$

$$\frac{dq}{dt} = Q_E - Q_C \tag{3}$$

$$\frac{\partial\rho}{\partial t} = -\overrightarrow{\nabla}\cdot(\overrightarrow{V}) \tag{4}$$

$$c_v\rho\frac{dT}{dt} + p\overrightarrow{\nabla}\cdot\overrightarrow{V} = Q_H + Q_D \tag{5}$$

Donde:

- $\overrightarrow{V}$ es la velocidad del aire en tres dimensiones (x, y, z)
- α es un volumen específico
- $\overrightarrow{\nabla}$ es la divergencia (u operador nabla)
- Φ es la altura geopotencial (aproxima la altura de una superficie de presión sobre el nivel del mar (American Meteorological Society 2017))
- $\overrightarrow{F}$ es la fuerza de fricción
- Ω x $\overrightarrow{V}$ es la fuerza de Coriolis (aparente desviación del aire por rotación de la Tierra (Schultz, Lomas, and Mulqueen, n.d.))
- p es presión
- R es una constante que varía según el gas
- T es temperatura
- q es la cantidad de agua en un pequeño volumen de aire
- Q_E es vapor de agua en el aire
- Q_C es vapor de agua condensado en el aire
- c_v es la capacidad calorífica (cantidad de calor necesaria para producir un cambio de una unidad de temperatura en una masa (Li, H. 2016)).
- Q_H es el flujo de calor de la superficie de la Tierra a la atmósfera
- Q_D es el calentamiento diabático causado por la condensación y la evaporación
- ρ es densidad
- t es el tiempo

La ecuación de movimiento (1) dice que el cambio en la velocidad tres-dimensional del aire más la fuerza de Coriolis es igual a la fuerza de fricción menos la divergencia de la presión por la densidad y menos la divergencia de la altura geopotencial.

La ley ideal del gas (2) describe la relación entre la presión, la densidad y la temperatura del aire en la atmósfera (Schultz, Lomas, and Mulqueen, n.d.).

La ley de conservación de la masa de agua (3) establece que el cambio en la cantidad de agua en un volumen pequeño de aire varía dependiendo si el vapor de aire es evaporado dentro del aire o condensado fuera de él (Schultz, Lomas, and Mulqueen, n.d.).

La ley de conservación de masa del viento (4) describe el cambio en la densidad de aire como resultado de la divergencia de aire en tres dimensiones (Schultz, Lomas, and Mulqueen, n.d.).

La primera ley de la termodinámica 5 describe cómo cambia la temperatura en la atmósfera en función de los flujos de calor de la superficie y el calor diabático más la divergencia del aire que causa movimientos verticales que calientan o enfrían el aire (Schultz, Lomas, and Mulqueen, n.d.).

Estas ecuaciones establecen un marco que identifica causalidad y relaciones entre el comportamiento de variables como la temperatura, el vapor de agua, la presión, el aire y el viento.

Como referencia para el estudio de la correlación entre ciertas variables se usa el trabajo de Huang and Dool (1993). En dicho trabajo se ha identificado, a través de observación experimental, una correlación negativa entre la lluvia y la temperatura en todas las temporadas y todas las áreas de Estados Unidos. El trabajo plantea que dicho fenómeno se explica generalmente por la presencia de nubes que enfrían el ambiente y también por la presencia de humedad en el suelo.

Metodología implementada

La idea central del trabajo es dividir el conjunto de datos en dos grupos: una para entrenamiento y ajustes de los modelos y el otro para probar los modelos. Se ajustaron tres modelos de regresión/clasificación utilizando la precipitación diaria siempre como la variable dependiente. Los modelos son: regresión lineal, máquina de soporte vectorial y árbol de decisión. Adicionalmente se agregó un caso de análisis de discriminante lineal.

Para el caso de regresión lineal se decidió aplicar una transformación continua a la variable precipitación dada por: $f(x) = log(1 + x)$, donde x es la precipitación observada para un día específico. El propósito de esta decisión es disminuir el efecto de días con precipitación extrema reportada, a la vez que los días con cero precipitación se mantienen sin cambio y todas las observaciones se mantienen no negativas.

Para el caso de los modelos cualitativos se decidió factorizar la variable precipitación diaria de manera que esta toma el valor de 1 si en ese día se reportó cualquier cantidad de lluvia mayor a cero y de lo contrario, precipitación diaria toma el valor de 0. El propósito de esta decisión es que los modelos cualitativos den una predicción de la probabilidad de lluvia ese día.

A continuación, una breve descripción de cada modelo:

Regresión Lineal

Es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente, una o varias variables independientes y un término de error. Así que tenemos la siguiente ecuación:

$$Y_t = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon_t$$

Donde: $Y :=$ es variable dependiente. $X_1, ..., X_n :=$ son variables independiente, pueden ser cualitativas o cuantitativas. $\beta_1, ..., \beta_n :=$ son parámetros que miden el nivel de influencia de los X_i. $\beta_0 :=$ intercepto de los parámetros β_i. $\epsilon :=$ es la perturbación o error cometido al hacer la aproximación. Entre los supuestos de las regresiones lineales tenemos que: la esperanza matemática de ϵ_t es nula, hay homocedasticidad para cada ϵ_t, normalidad de las perturbaciones, además para cualquier ϵ_i y ϵ_j tenemos que no están correlacionados (Wasserman 2006).

Máquinas de soporte vectorial

La idea de máquinas de soporte vectorial (SVM en inglés) es separar instancias de los datos utilizando hiperplanos con los márgenes más amplios posibles determinados por valores en los bordes llamados vectores de soporte. En el proceso de escoger el mejor hiperplano se considera si se permite o no que se den infracciones dentro de los márgenes y con qué nivel de tolerancia. Este parámetro juega un papel importante en la flexibilidad del modelo.

En el caso de un separador lineal SVM el problema de optimización está dado por:

$$\underset{w,b,\zeta}{\text{minimizar}} \quad \frac{1}{2}w^T w + \zeta \sum_{i=1}^{m} \zeta^{(i)}$$

$$\text{sujeto a} \quad t^{(i)}(w^T x^{(i)} + b) \geq 1 - \zeta^{(i)} \quad \text{y} \quad \zeta^{(i)} \geq 0 \quad \text{para } i = 1, 2, \cdots, m$$

Donde $\zeta^{(i)}$ mide la tolerancia de la instancia i para traspasar el margen, w es el vector de pesos que determinan la clase de una nueva instancia, b es el término de sesgo y $t^{(i)}$ indica si la instancia i pertenece a la clase (separación) o no. Los w estimados del problema de minimización son los pesos que se usan la clase de una nueva observación.

En el caso de que los datos no sean separables linealmente se puede utilizar el truco del kernel. Este consiste en utilizar una función que calcule el producto punto de $\phi(a)^T\phi(b)$ usando solo los vectores a y b sin aplicar la transformación ϕ. Los Kernel's más comunes son: lineal, polinomial, radial gaussiano (RBF) y sigmoidal (Géron 2017).

Árboles de decisión

Un árbol de decisión se utiliza como una herramienta visual y analítica, para predecir los valores objetivo. Es un clasificador expresado como una partición recursiva del espacio de instancia. El árbol de decisión consta de nodos que forman un árbol enraizado, lo que significa que es un árbol dirigido con un nodo llamado "raíz", que representa una prueba en un atributo (campo o parámetro), que no tiene bordes entrantes. En un árbol de decisión, cada nodo interno divide el espacio de la instancia en dos o más subespacios de acuerdo con una determinada función discreta de los valores de los atributos de entrada. Por lo general, el objetivo es encontrar el árbol de decisión óptimo minimizando el error de generalización. Sin embargo, otras funciones de destino también pueden definirse, por ejemplo, minimizando el número de nodos o minimizando la profundidad promedio (Rokach and Maimon 2005).

Análisis de discriminante lineal

Similar a SVM, el propósito de análisis de discriminante lineal (LDA) es encontrar el subespacio que acumula las observaciones de la misma clase (según la variable dependiente) y al mismo tiempo amplia el margen que separa los agrupamientos de observaciones de diferentes clases. En este modelo se asume que la función de densidad condicionada a la clase k (donde k = 1 es lluvia) de las variables dependientes es una gaussiana multivariada; es decir:

$$f(x|Y = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right\}, \quad k = 0, 1$$

donde Y es la precipitación, Σ_k es la matriz de covarianza de la clase k, μ_k es la media de la clase k y d es el número de variables explicativas. Luego se asume que las matrices de covarianzas son iguales entre las clases y entonces, a partir de la regla de Bayes, el problema de clasificación se reduce a estimar:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k \Sigma^{-1} + log(\pi_k) \ , k = 1, 2$$

donde π_k es la densidad previa de la clase k. El modelo clasifica una nueva observación con base en cuál $\delta_k(x)$ es mayor (Wasserman 2004).

PCA

Los modelos cualitativos tienen un mejor desempeño cuando los datos son fácilmente separables; es decir, cuando es posible observar agrupamientos (clusters) de los datos que pueden ser separados o clasificados según el valor que tome la variable dependiente (en este caso la precipitación). Para estudiar dicha separabilidad se realizó un análisis de componentes principales (PCA). Este método permite reducir la dimensión de los datos al calcular los componentes que aportan mayor varianza (componentes principales). A partir de los resultados obtenidos por PCA se graficó la información de las dos dimensiones con mayor varianza y se procedió a observar la existencia o no de clusters.

Metodología para selección de modelos

Para seleccionar el mejor modelo de regresión lineal se implementó el algoritmo de Forward Stepwise donde se compararon el R^2 ajustado, los valores p de la prueba t de regresión y el criterio de información de Akaike (AIC); de manera que entre dos modelos se escogió el que tuviera mayor R^2 ajustado, menores valores p y menor AIC.

Para seleccionar el mejor modelo de máquinas de soporte vectorial se implementó el algoritmo de Backward Stepwise donde se comparó la norma de la resta entre los pesos del modelo completo y los pesos del modelo sin una variable. De esta manera, la variable que generara un menor cambio en los pesos de decisión sería eliminada. Este proceso se continuó hasta que las variables restantes generaran una pérdida en los pesos de decisión similar. Una vez seleccionado el mejor modelo, se aplicó validación cruzada de k subconjuntos con diferentes kernels: lineal, polinomial y radial gaussiano; para elegir el mejor kernel y sus mejores parámetros. El método de selección se realizó a partir de comparar la exactitud y el Kappa de Cohen (entre mayor mejor). Para la implementación de validación cruzada se usó el paquete caret de R.

Para seleccionar el mejor modelo de árboles de decisión se implementó el algoritmo de Forward Stepwise donde se compararon el criterio de información de Akaike (AIC) y el criterio de información Bayesiano (BIC), de manera que entre dos modelos se escogió el que tuviera menor AIC y menor BIC. Una vez seleccionado el modelo se utilizó el paquete GPLTR de R para reducir el tamaño del árbol al eliminar secciones que no contribuyen a la clasificación (pruning).

Como el caso de análisis de discriminante lineal tiene un propósito solamente ilustrativo, se procedió a utilizar el mejor modelo obtenido del proceso de selección de máquina de soporte vectorial.

Para todos los modelos se consideró la posibilidad de incluir interacciones de variables explicativas si se encontraba alguna relación con la teoría del clima expuesta en el marco teórico y si el modelo lo permitía.

Evaluación de modelos

Una vez que los mejores modelos fueron seleccionados y calibrados se procedió a utilizar los datos observados del periodo 01/05/2018 - 01/05/2019 para probar las predicciones de los modelos. En el caso del modelo lineal se calculó el error cuadrático total que está dado por:

$$\sum_{k=1}^{N}(Y_k - \hat{Y}_k)^2$$

donde Y_k es el valor de la precipitación reportado para el día k, $\hat{Y}_k$ es el valor estimado de precipitación por el modelo para el día k y N es el total de días del periodo de prueba. Además, se calcularon los extremos de los intervalos de predicción (I.P.) para los datos de prueba y se muestra el número de eventos de precipitación cuyo valor sí se encuentra dentro de dichos intervalos.

En el caso de los modelos cualitativos se construyó la matriz de confusión, donde el evento positivo es lluvia y el negativo es no lluvia, y se calcularon los estadísticos de:

- Exactitud: total de escenarios de lluvia o no lluvia correctamente predichos entre el total de eventos.
- Kappa de Cohen: mide cuánto aporta el modelo ajustado sobre un modelo aleatorio construido usando la información de la matriz de confusión.
- Sensibilidad: número de eventos de lluvia correctamente predichos entre el total de eventos de lluvia.
- Especificidad: número de eventos de no lluvia correctamente predichos entre el total de eventos de no lluvia.

Finalmente se graficó la curva ROC que grafica la tasa de falsos positivos versus la sensibilidad del modelo y se calculó el área bajo la curva.

Como parte de preprocesamiento de los datos se procedió a dividir los datos por estaciones para incluir el efecto de periodos secos o de mucha lluvia siguiendo el siguiente cuadro:

Tabla 1: Fechas de las estaciones en Chicago

Estación	Período
Invierno	21 de diciembre al 21 de marzo
Primavera	21 de marzo al 21 de junio
Verano	21 de junio al 21 de setiembre
Otoño	21 de setiembre al 21 de diciembre

Fuente: Elaboración propia usando datos de: `https://seasonsyear.com/USA/Illinois/Chicago`

Como parte del análisis descriptivo se decidió estimar la densidad de la variable precipitación utilizando el método no paramétrico de kernels y calculando con un ancho de banda común dado por la regla normal.

Para la implementación de todos los modelos se utilizó el software estadístico R.

Análisis y descripción de los datos

Para la selección de datos se escogió el "Resumen Diario" reportado por estaciones climatológicas de Chicago y para el período que abarca desde el primero de mayo del 2008 hasta al primero de mayo del 2019. Dicho resumen está compuesto por observaciones diarias de diferentes fuentes y que son sujetos a un proceso de control de calidad (NOAA, n.d.). El "Resumen Diario" fue obtenido gratuitamente en la página Centro Nacional de Información Ambiental: Administración Oceánica y Atmosférica Nacional (NOAA), Estados Unidos (`https://www.ncdc.noaa.gov/cdo-web/datasets`). Los datos están en formato CSV y cada valor representa una observación de una variable climatológica para un día y una estación climatológica (NOAA, n.d.). Los valores observados están en medidas imperiales (pulgada, millas,...) por lo que se les aplicó la respectiva conversión al sistema métrico.

Tabla 2: Variables climatológicas disponibles en la tabla de datos

Variable	Descripción	Unidad de medida
TEMP	Temperatura promedio	Grados Celsius
DEWP	Punto de rocío promedio	Grados Celsius
SLP	Presión del nivel del mar promedio	Pascales
STP	Presión de estación promedio	Pascales
VISIB	Visibilidad promedio	Kilómetros
WDSP	Velocidad promedio del viento	Metros por segundo
MXSPD	Velocidad del viento máxima sostenida	Metros por segundo
GUST	Máxima ráfaga de viento	Metros por segundo
MAX	Temperatura máxima reportada	Grados Celsius
MIN	Temperatura mínima reportada	Grados Celsius
PRCP	Total de precipitación	milímetros
SNDP	Profundidad de nieve	milímetros
FRSHTT	Indicador de ocurrencia de un fenómeno	No aplica

Fuente: Elaboración propia usando datos de:
`https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/GHCND_documentation.pdf`

NOTA: La variable ***FRSHTT*** es una variable indicadora (1 ó 0) y reporta la ocurrencia de: Neblina (primer dígito), lluvia o llovizna (segundo dígito), nieve (tercer dígito), granizo (cuarto dígito), trueno (quinto dígito) y tornado (sexto dígito) (NOAA 2006).

Para el proceso de selección de estaciones climatológicas se buscaron aquellas con la mayor cobertura de datos para el período a estudiar. Para resumir los datos reportados por las estaciones meteorológicas y así obtener un único valor para una fecha observada se utilizó el método de polígonos de Thiessen (Arias 2001). De esta manera el valor observado de una variable para un día es el promedio ponderado de las observaciones reportadas por las estaciones para ese día donde los pesos son determinados por su área de influencia. Para

calcular las áreas de influencia se utilizó la latitud y longitud de cada estación como punto de localización y se utilizó el paquete **deldir** de R.

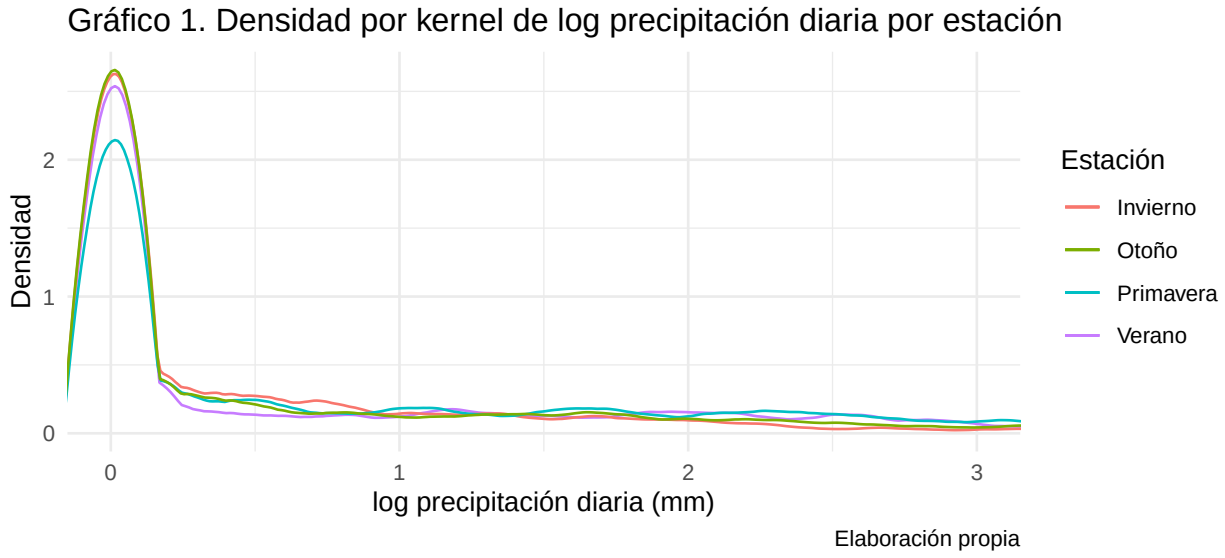
Tabla 3: Estaciones climatológicas de Chicago seleccionadas y su peso según Thiessen

Estación	Latitud	Longitud	Peso
CHICAGO NORTHERLY ISLAND, IL US	41.856	-87.609	0.030
CHICAGO BOTANIC GARDEN, IL US	42.140	-87.785	0.107
CHICAGO MIDWAY AIRPORT, IL US	41.786	-87.752	0.044
CROWN POINT 1.1 N, IN US	41.439	-87.354	0.063
DYER 1.0 WNW, IN US	41.507	-87.525	0.145
BRIDGEVIEW 1.3 NNW, IL US	41.755	-87.817	0.213
CHICAGO 4.7 NE, IL US	41.886	-87.621	0.139
CHICAGO 5.5 ESE, IL US	41.801	-87.590	0.106
ELK GROVE VILLAGE 2.2 WSW, IL US	41.995	-88.053	0.084
CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	41.960	-87.932	0.069

Fuente: Elaboración propia y usando datos de:
`https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/GHCND_documentation.pdf`

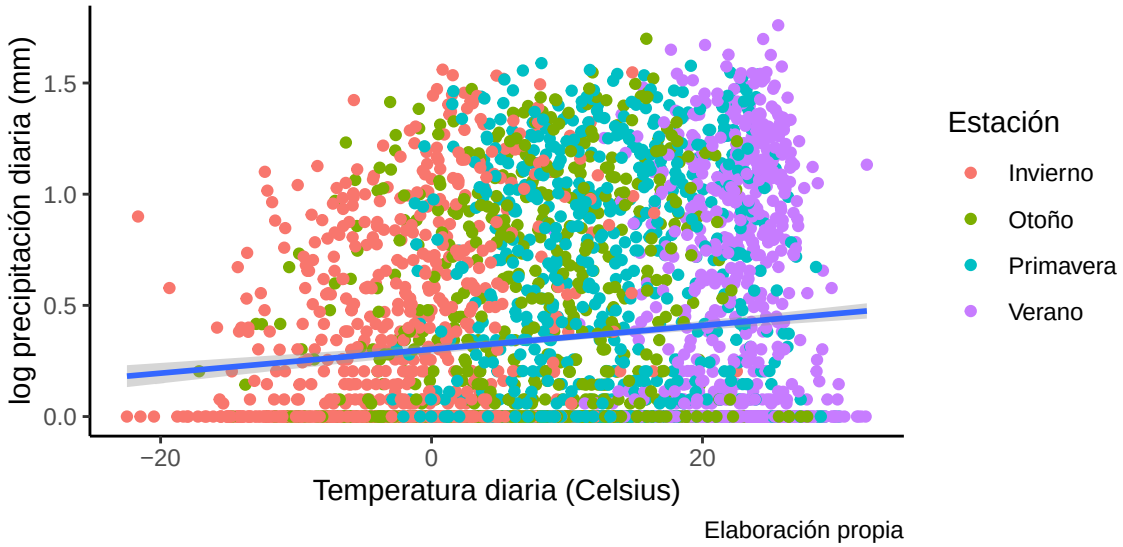
La tabla cuenta con 3392 valores ausentes para la variable SNDP (nieve) y 339 valores ausentes para la variable GUST (ráfaga de viento). Se puede pensar que el número de valores ausentes de la variable nieve es muy significativo (el total de datos es 4018), pero de acuerdo con la documentación esto se puede deber a que la mayoría de las estaciones climatológicas no reportan 0 en días sin nieve en el suelo (NOAA 2006). Tomando esto en consideración se procedió a llenar los valores ausentes de la variable nieve con ceros. Los valores ausentes de la variable ráfaga de viento se reemplazaron utilizando el paquete mice de R y el método de muestreo aleatorio.

Gráficos de resultados descriptivos



Implementando un kernel epanechnikov y tomando un ancho de banda común para las estaciones se puede observar en el gráfico 1 que la densidad no paramétrica de la log lluvia se comporta muy similar para todas las estaciones con una gran acumulación alrededor de 0. Se puede destacar que invierno presenta la mayor concentración y primavera la menor.

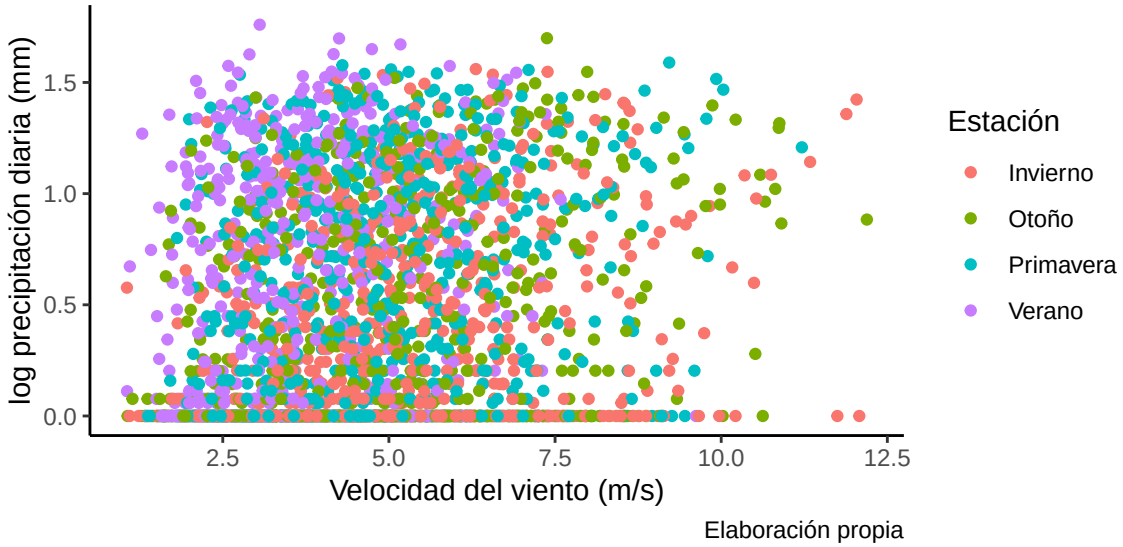
Gráfico 2. Dispersión temperatura y log precipitación diaria



Elaboración propia

Los mayores niveles de log precipitación se acumulan para las temperaturas más altas, dicha acumulación deja ver un comportamiento creciente de la lluvia con respecto a la temperatura, es decir, que conforme aumenta la temperatura se logran acumular una mayor cantidad de precipitación. Además, se observa que las temperaturas son características de las estaciones, por lo que se logra observar una agrupación de puntos para las distintas estaciones.

Gráfico 3. Dispersión velocidad del viento y log precipitación diaria

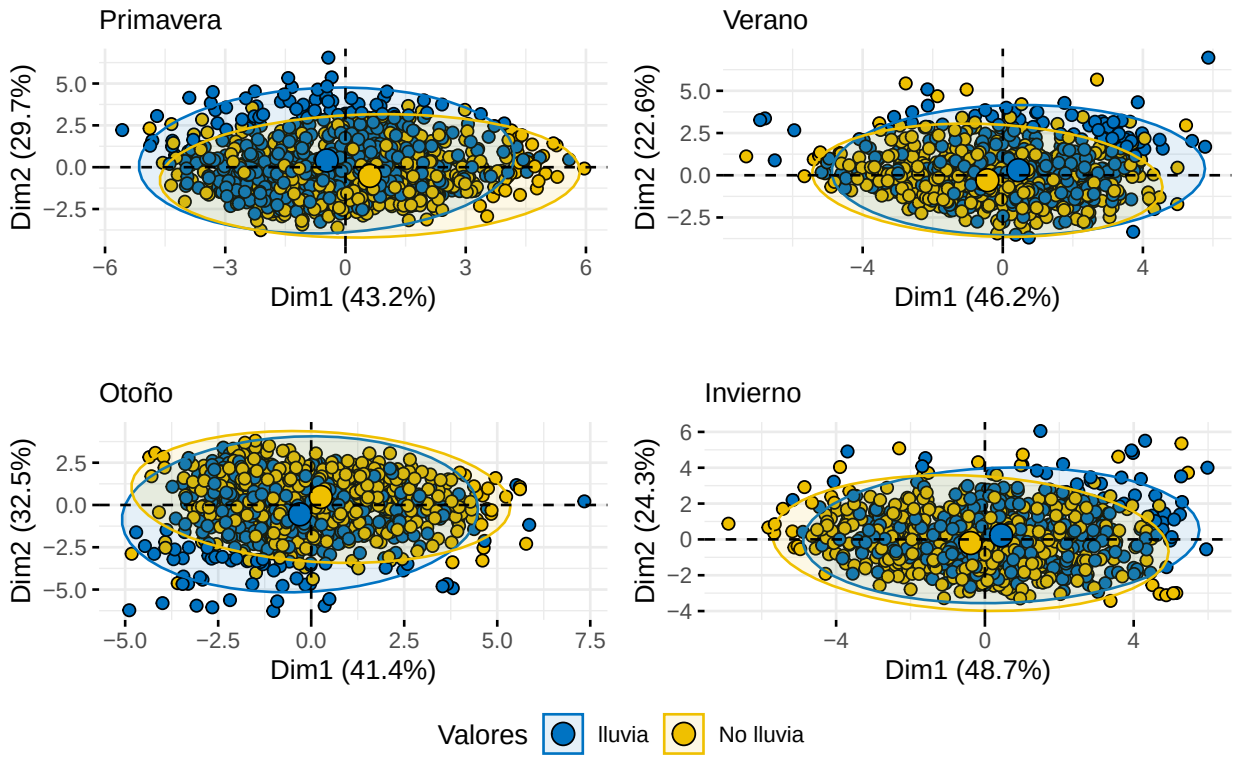


Elaboración propia

Cuando analizamos los niveles de log precipitación diarios con la velocidad del viento, se logra observar una dispersión de los distintos puntos observados para cada estación. Las mayores acumulaciones de lluvia, en este caso, están vinculadas en mayor grado con una velocidad menor de viento.

Resultados implementación PCA

Gráfico 4. PCA con dos dimensiones por estación



Fuente:Elaboración propia

En el gráfico 4 no se observa una separación clara de los datos por estación, pero sí se observa que la estación primavera presenta cierta separabilidad donde las observaciones de no lluvia se extienden para valores positivos del segundo componente principal. Se espera que dicha estación tenga los mejores resultados en los modelos cualitativos.

Resultados

Ahora se presentan todos los resultados de calibrar los modelos y realizar las pruebas de predicción y al final se incluye una discusión general de los valores obtenidos.

Regresión lineal

Modelos seleccionados

Estos son los modelos seleccionados por el método de Forward Stepwise:

- **Primavera**: $\log(1+\text{PRCP}) = \beta_0 + \beta_1\text{VISIB} + \beta_2\text{GUST} + \beta_3\text{WDSP} + \beta_4\text{SLP} + \beta_5\text{SLP}\cdot\text{DEWP} + \beta_6\text{DEWP} + \beta_7\text{TEMP} + \beta_8\text{STP}\cdot\text{TEMP} + \beta_9\text{STP}$
- **Verano**: $\log(1+\text{PRCP}) = \beta_0 + \beta_1\text{VISIB} + \beta_2\text{MXSPD} + \beta_3\text{MIN} + \beta_4\text{MAX} + \beta_5\text{SLP} + \beta_6\text{SLP}\cdot\text{DEWP} + \beta_7\text{DEWP} + \beta_8\text{TEMP} + \beta_9\text{STP}\cdot\text{TEMP} + \beta_{10}\text{STP}$
- **Otoño**: $\log(1+\text{PRCP}) = \beta_0 + \beta_1\text{SLP} + \beta_2\text{SLP}\cdot\text{DEWP} + \beta_3\text{DEWP} + \beta_4\text{TEMP} + \beta_5\text{STP}\cdot\text{TEMP} + \beta_6\text{DEWP}\cdot\text{STP} + \beta_7\text{STP} + \beta_8\text{VISIB} + \beta_9\text{VISIB}\cdot\text{WDSP} + \beta_{10}\text{WDSP}$

- **Invierno**: $\log(1 + \text{PRCP}) = \beta_0 + \beta_1 \text{SLP} + \beta_2 \text{SLP} \cdot \text{DEWP} + \beta_3 \text{DEWP} + \beta_4 \text{TEMP} + \beta_5 \text{STP} \cdot \text{TEMP} + \beta_6 \text{TEMP} \cdot \text{SLP} + \beta_7 \text{STP} + \beta_8 \text{VISIB} + \beta_9 \text{VISIB} \cdot \text{WDSP} + \beta_{10} \text{WDSP} + \beta_{11} \text{GUST} + \beta_{12} \text{SNDP}$

Donde: PRCP es la precipitación diaria, DEWP es el punto de rocío, TEMP es la temperatura promedio, MAX y MIN son el máximo y mínimo de temperatura, SLP es la presión al nivel del mar, STP es la presión de estación, VISIB es la visibilidad, WDSP es la velocidad promedio del viento, GUST es ráfaga de viento y SNDP es la nieve.

Estadísticos de selección

Tabla 4: Valor de estadísticos de selección regresión lineal

Estadístico	Primavera	Verano	Otoño	Invierno
R cuadrado	0.444	0.380	0.455	0.432
R cuadrado ajustado	0.439	0.374	0.448	0.425
AIC	2149.867	2282.150	1798.148	1609.850

Fuente: Elaboración propia

A continuación, se muestran los coeficientes calculados para la estación de invierno.

Tabla 5: Valor de los coeficientes y valores p (prueba t) de invierno regresión lineal

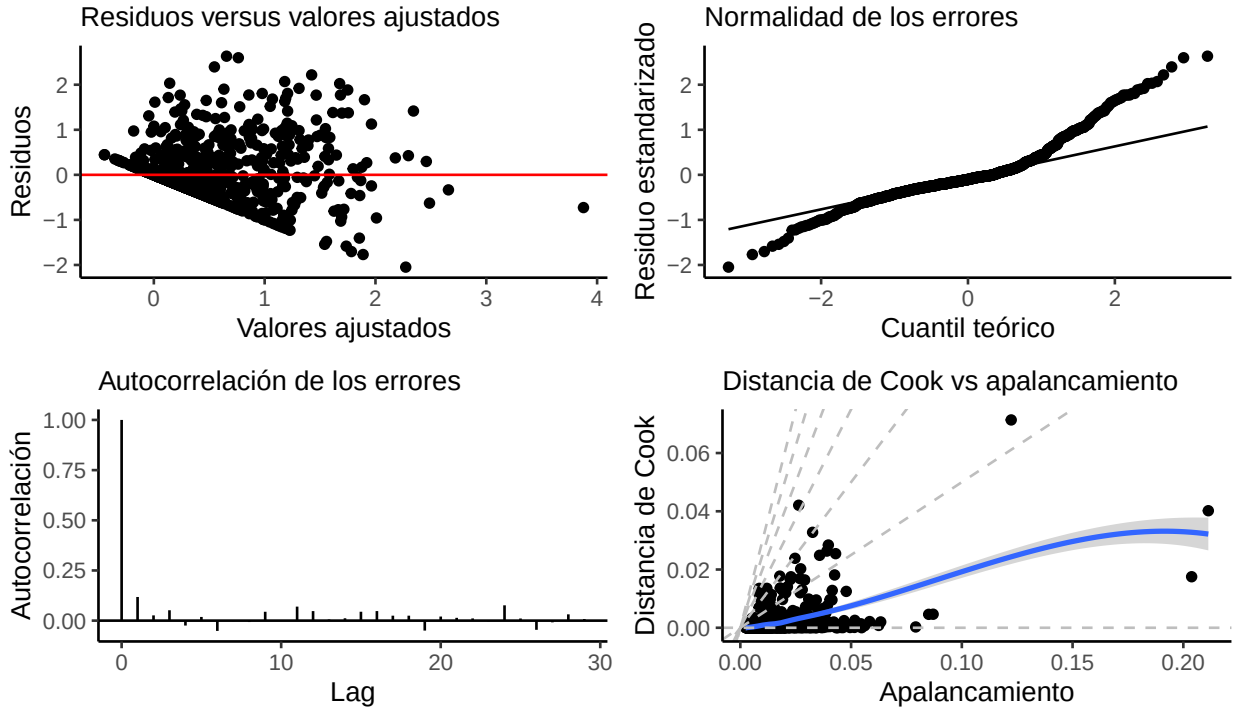
Variable	Coeficiente	Valor p
Intercepto	33.7253	11.96×10^{-9}
GUST	0.0247	3.427×10^{-3}
SLP	0.0018	880.6×10^{-9}
DEWP	6.1056	143.9×10^{-9}
TEMP	-4.4767	1.123×10^{-3}
STP	-0.0022	4.867×10^{-9}
VISIB	0.0503	8.061×10^{-3}
WDSP	0.3496	7.249×10^{-12}
SNDP	0.0009	1.507×10^{-3}
SLP:DEWP	-0.0001	198.1×10^{-9}
SLP:TEMP	0.0002	407.4×10^{-9}
TEMP:STP	-0.0001	355.0×10^{-6}
VISIB:WDSP	-0.0195	81.27×10^{-9}

Fuente: Elaboración propia

Gráficos de chequeos de supuestos modelo regresión lineal

Se muestran los gráficos de chequeos para la estación de invierno.

Gráfico 5. Chequeo de supuestos modelo regresión lineal invierno



Fuente:Elaboración propia

Resultados de la predicción

Tabla 6: Resultado de las predicciones modelo regresión lineal

Variable	Primavera	Verano	Otoño	Invierno
Error cuadrático total	72.236	64.802	50.503	27.230
Valores contenidos en I.P.	84.000	85.000	85.000	84.000
Tasa de contenidos en I.P.	0.903	0.924	0.934	0.933

Máquinas de soporte vectorial

El modelo seleccionado por Backward Stepwise es el mismo para todas las estaciones y está dado por: PRCP (lluvia) ~ DEWP(punto de rocío) + WDSP (velocidad viento) + TEMP (temperatura) + SLP (presión nivel del mar) + MIN (mínimo de temperatura).

A partir de validación cruzada se seleccionó el kernel Gaussiano Radial Basis (RBF): $K(a, b) = \exp(-\gamma||a-b||^2)$, donde γ es un parámetro libre positivo. Los mejores resultados se obtuvieron con los siguientes parámetros:

Tabla 7: Mejores parámetros validación cruzada para SVM Radial

Parámetro	Primavera	Verano	Otoño	Invierno
zeta	1.000	0.500	1.000	1.000
gamma	0.218	0.231	0.228	0.231
Vectores de soporte	548.000	628.000	611.000	563.000

Ahora se muestra la matriz de confusión de la estación verano:

Tabla 8: Matriz de confusión modelo svm radial estación verano

	No lluvia referencia	Lluvia referencia
No lluvia predicción	39	11
Lluvia predicción	6	36

Y los resultados generales de la predicción:

Tabla 9: Estadísticos matriz de confusión modelo svm Radial

Estadístico	Primavera	Verano	Otoño	Invierno
Exactitud	0.8172	0.8152	0.6374	0.7444
Kappa de Cohen	0.6206	0.6311	0.2807	0.4993
Sensibilidad	0.8704	0.7660	0.5106	0.6538
Especificidad	0.7436	0.8667	0.7727	0.8684
Área bajo curva ROC	0.8913	0.8965	0.7911	0.7854

Árboles de decisión

El modelo seleccionado por Forward Stepwise y AIC es el mismo para todas las estaciones y está dado por: PRCP (lluvia) ~ DEWP(punto de rocío) + WDSP (velocidad viento) + TEMp (temperatura) + GUST (ráfaga de viento) + VISIB (visibilidad) + MAX (temperatura máxima) + STP (presión de estación) + SLP (presión nivel del mar)

Estos son los parámetros resultantes:

Tabla 10: Parámetros resultantes del proceso de Forward stepwise y pruning

Parámetro	Otoño	Primavera	Invierno	Verano
AIC	957.3	845.2	825.7	1065
Total nodos	9.0	11.0	19.0	19
Nodos terminales	5.0	6.0	10.0	10
Profundidad	4.0	5.0	4.0	5

Ahora se muestra la matriz de confusión de la estación verano:

Tabla 11: Matriz de confusión modelo árbol de decisión estación verano

	No lluvia referencia	Lluvia referencia
No lluvia predicción	39	16
Lluvia predicción	6	31

Ahora se muestran los resultados generales de la predicción:

Tabla 12: Estadísticos matriz confusión modelo árbol de decisión

Estadístico	Verano	Primavera	Otoño	Invierno
Exactitud	0.761	0.763	0.637	0.711
Kappa de Cohen	0.524	0.521	0.281	0.440

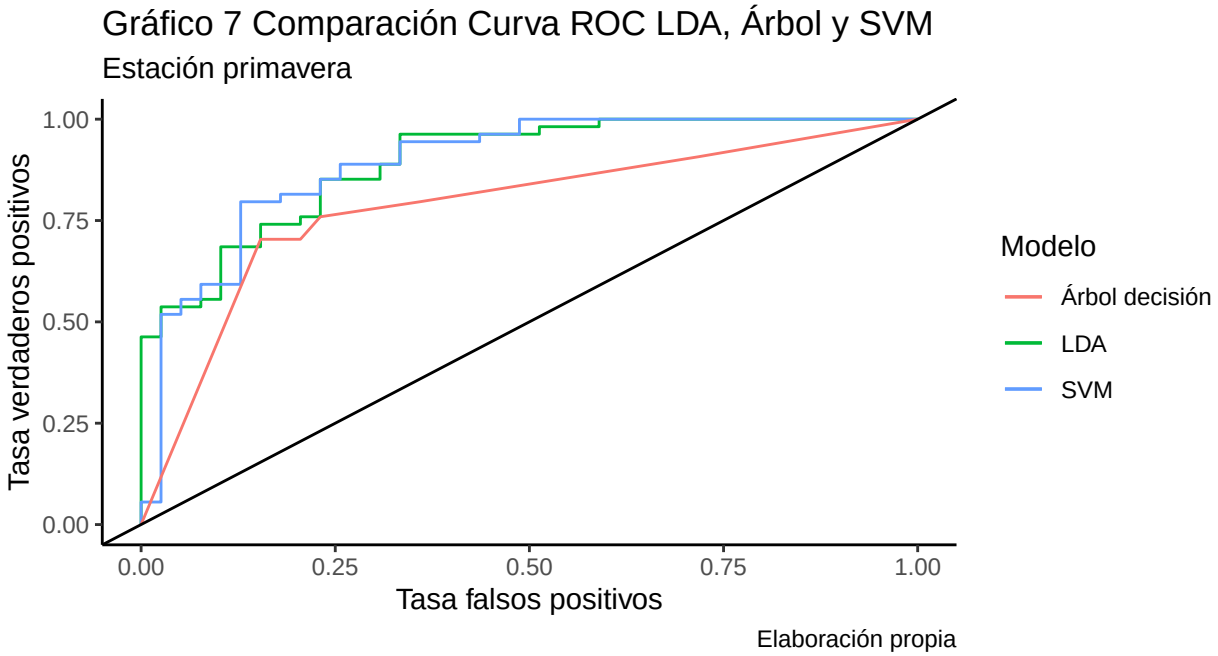
Estadístico	Verano	Primavera	Otoño	Invierno
Sensibilidad	0.660	0.759	0.511	0.596
Especificidad	0.867	0.769	0.773	0.868
Área bajo curva ROC	0.814	0.783	0.752	0.715

Caso análisis de discriminante lineal

Se muestra los resultados de implementar análisis de discriminante lineal a la estación primavera usando el mismo modelo que máquina de soporte vectorial

Tabla 13: Estadísticos modelo análisis discriminante lineal primavera

Estadístico	Valor
Exactitud	0.7957
Kappa de Cohen	0.5760
Sensibilidad	0.7179
Especificidad	0.8519
AUC	0.8932



Análisis de resultados

En general todos los modelos tuvieran dificultad para ajustarse bien a los datos. El modelo de regresión lineal no pudo superar el umbral del 50% de R^2 para ninguna estación. En el gráfico 5 de chequeos de supuestos se puede observar que para la estación de invierno los errores tienen poca correlación y que la distancia de Cook y el apalancamiento se mantienen en un rango aceptable. Por otro lado, los residuos difieren bastante de una distribución normal y no parecen estar centrados en cero, esto se puede deber a la transformación logarítmica que se le aplico a la lluvia y a la acumulación en cero resultante que se observó en el gráfico 1. Si bien se intentó trabajar con los datos de lluvia sin transformar, aplicarle el logaritmo a la lluvia siempre mejoró los resultados de R^2 y AIC. Es interesante que en la Tabla 5 se confirma lo expuesto en el marco

teórico referente a que se espera que la temperatura tenga una correlación negativa con respecto a la lluvia. El modelo de SVM generó muchas infracciones de margen (vectores de soporte) lo que refleja lo difícil de poder separar bien los datos, algo observado desde el gráfico de PCA. En particular la estación de verano tiene aproximadamente el 70% de las observaciones como vectores de soporte lo que se espera que dificulte que el modelo pueda realizar una clasificación clara. Para el modelo de árbol de decisión los parámetros quedaron similares para todas las estaciones. Tal vez es valioso destacar que el modelo LDA presentó buenos resultados a pesar de no tener un procedimiento de selección de variables propio. Sería oportuno en el futuro realizar un análisis más exhaustivo de este modelo.

Por parte de la predicción, los modelos cualitativos mostraron resultados balanceados entre verdaderos positivos y verdaderos negativos, además de resultados similares de exactitud entre estaciones de un mismo modelo. Particularmente se observa que la estación de otoño tuvo los peores resultados de exactitud y sensibilidad para SVM y para árboles de decisión. Esto se puede deber a que otoño es una estación de transición entre verano e invierno y presenta un mayor rango en sus variables como se puede observar en el gráfico 2 de temperatura. Por otro lado, para ambos modelos los mejores resultados generales se presentaron en la estación de verano y primavera. También cabe destacar que el modelo SVM tuvo ventajas generales en exactitud, kappa de Cohen y sensibilidad comparado con el modelo de árbol de decisión. Por otro lado, el modelo de árbol de decisión tuvo mejores resultados para clasificar correctamente días de no lluvia en primavera. Una posible razón de la superioridad general de SVM sobre árboles puede deberse a que el árbol esté sobreajustando (overfitting) los datos de entrenamiento lo que disminuye su capacidad de generalizar y empeora sus resultados de clasificación. Es posible que la rigidez del modelo de SVM haya resultado en una mejor generalización.

Por parte del modelo de regresión lineal se observó que el menor error cuadrático total se presentó en las estaciones de invierno y otoño. Esto puede ser resultado de lo observado en el Gráfico 1 de densidad donde invierno y otoño presentan la mayor acumulación de lluvia alrededor de 0. Además, los intervalos de predicción atraparon la gran mayoría de las nuevas observaciones de precipitación para todas las estaciones. Sin embargo, esto no es necesariamente un indicador sólido pues hay que recordar que la lluvia está transformada de manera logarítmica lo que comprime sus valores y los intervalos de predicción son por lo general amplios.

Conclusiones y mejoras

A pesar de que la predicción del clima es un proceso complejo que debe considerar múltiples variables, escenarios y factores; los modelos estadísticos obtuvieron resultados relativamente buenos lo que permite pensar que a través de un análisis y transformaciones más complejas de los datos se podría obtener mejores predicciones. Se observó que trabajar los datos de las variables climatológicas con las alteraciones comunes como la normalización no es suficiente para obtener los mejores resultados. Por ejemplo, se especula que debe haber una transformación más adecuada para la lluvia que la logarítmica y esa sería una mejora para explorar.

Ante la pregunta de que si es posible realizar un pronóstico de la lluvia solo utilizando modelos estadísticos y el historial del clima de una región parece ser que la respuesta es un depende de la rigurosidad o importancia del resultado de la predicción. Por lo menos de este trabajo no se puede concluir que un modelo estadístico simple sea capaz de sustituir un modelo complejo numérico-estadístico que se actualiza con nueva información de manera constante. Además, este trabajo no consideró variables exógenas y otros fenómenos como huracanes y el Fenómeno del Niño de la Niña.

Una mejora general es implementar una selección más adecuada de los modelos cuantitativos y cualitativos, tal vez consultando a personas con más experiencia en el tema (por ejemplo, meteorólogos). Lo mismo aplica para la selección de variables y para la obtención de los datos.

Referencias

Arias, O. 2001. "Estudio Hidrometeorológico de la cuenca del río Tempisque, Guanacaste." https://repositoriotec.tec.ac.cr/bitstream/handle/2238/2616/Informe%7B/_%7DFinal.pdf?sequence=1%7B/&%7DisAllowed=y.

Bramer, D, D Wojtowicz, and Halls. 2010. "Observed temperature." http://ww2010.atmos.uiuc.edu/(Gh)/guides/maps/sfcobs/tmp.rxml.

"Clima." 1988. Oceano.

"Condensation nucleus." 2017. Oceano. https://www.britannica.com/science/condensation-nucleus.

Cramer, S, M. Kampouridis, A. Freitas, and A. Alexandridis. 2017. "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives." *ELSEVIER*, 1–27.

Géron, Aurelien. 2017. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly Media, Inc.

Huang, Jin, and Huug M. van den Dool. 1993. "Monthly Precipitation-Temperature Relations and Temperature Prediction over the United States." *Journal of Climate* 6 (6): 1111–32. https://doi.org/10.1175/1520-0442(1993)006%7B/%%7D3C1111:MPTRAT%7B/%%7D3E2.0.CO;2.

"Lluvia." 1988. Oceano.

National Weather Service. n.d.a. "National Weather Service Glossary." http://ww2010.atmos.uiuc.edu/(Gh)/guides/maps/sfcobs/dwp.rxml.

———. n.d.b. "Pressure Definitions." https://www.weather.gov/bou/pressure%7B/_%7Ddefinitions.

NOAA. 2006. "GLOBAL SURFACE SUMMARY OF DAY DATA." https://www7.ncdc.noaa.gov/CDO/GSOD%7B/_%7DDESC.txt.

———. n.d. "GHCN (Global Historical Climatology Network) – Daily Documentation." https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/GHCND%7B/_%7Ddocumentation.pdf.

Pérez-Vega, A, C M Travieso, J G Hernández-Travieso, J B Alonso, M K Dutta, and A Singh. 2016. "Forecast of temperature using support vector machines." In *2016 International Conference on Computing, Communication and Automation (Iccca)*, 388–92. https://doi.org/10.1109/CCAA.2016.7813752.

Pu, Zhaoxia, and Eugenia Kalnay. 2019. "Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation." *Springer Nature*, 67–97.

Rokach, Lior, and Oded Maimon. 2005. *Decision Trees*. Springer US. https://doi.org/10.1007/0-387-25465-X%7B/_%7D9.

Schultz, D, F Lomas, and Katy Mulqueen. n.d. "The Laws of Physics for the Atmosphere – The Computer Model." http://manunicast.seaes.manchester.ac.uk/how/physics.html.

Wasserman, Larry. 2004. *All of Statistics A Concise Course in Statistical Inference*. Springer. https://doi.org/https://doi.org/10.1007/978-0-387-21736-9.

———. 2006. *All of Nonparametric Statistics*. Springer-Verlag New York. https://doi.org/10.1007/0-387-30623-4.

Wilks, D S. 1999. "Interannual variability and extreme-value characteristics of several stochastic daily precipitation models." *Agricultural and Forest Meteorology* 93 (3): 153–69. https://doi.org/https://doi.org/10.1016/S0168-1923(98)00125-7.

Anexos

Código implementado

Sección Regresión lineal

```
#load("Datos_finales.RData")

# PRIMAVERA
Primavera <- Tabla_entrenamiento %>% filter(Estacion == "Primavera")

RegLin_Pri <- lm(data=Primavera, formula = PRCP ~ VISIB + GUST + WDSP + SLP*DEWP + TEMP*STP)
summary(RegLin_Pri)
summary(glm(data=Primavera, formula = PRCP ~ VISIB + GUST + WDSP + SLP*DEWP + TEMP*STP))
# # 0.4386 2149.9


Pri_newdata <- Tabla_testeo %>%
  filter(Estacion == "Primavera") %>% select(c("VISIB", "GUST", "WDSP", "SLP", "DEWP",
"TEMP", "STP"))

predict(object = RegLin_Pri, newdata = Pri_newdata, interval = "confidence")

predict(object = RegLin_Pri, newdata = Pri_newdata, interval = "prediction")


# Usando Forward Stepwise Selection, R^2 y AIC
# RegLin_Pri <- lm(data=Primavera, formula = PRCP ~ VISIB + STP + GUST + DEWP + TEMP + WDSP
#+ SLP + SNDP)
# summary(RegLin_Pri)
# summary(glm(data=Primavera, formula = PRCP ~ VISIB + STP + GUST + DEWP + TEMP + WDSP
#+ SLP + SNDP))
# # 0.4264 2168.6


Tabla_valores <-
  tibble(Variable = names(RegLin_Pri$coefficients),
           Coeficiente = RegLin_Pri$coefficients,
         Valores_p = summary(RegLin_Pri)$coefficients[,4])

Tabla_estadisticos <- tibble(Estadistico = c("R cuadrado",
                                             "R cuadrado ajustado",
                                             "F",
                                             "Grados de libertad"),
                             Valor = c(summary(RegLin_Pri)$r.squared,
                                       summary(RegLin_Pri)$adj.r.squared,
                                       summary(RegLin_Pri)$fstatistic[1],
                                       summary(RegLin_Pri)$df[2]))
```

```
## Gráficos de chequeos

# ggplot(RegLin_Inv, aes(x = .fitted, y = .resid)) + geom_point()+
#   geom_hline(yintercept = 0, colour = "red")+
#   labs(title = "Gráfico 7. Residuos vs valores ajustados",
```

```
#        subtitle = "Modelo regresión lineal primavera ", y = "Residuos", x = "Valores ajustados",
#        caption="Fuente: Elaboración propia")+
#   theme_classic()
#
#
#
#
# tibble( x = residuals(RegLin_Pri)) %>%
#   ggplot(aes(sample = x)) +
#   geom_qq(distribution = qnorm)+
#   stat_qq_line()+
#   labs(title = "Gráfico 8. Normalidad de los errores",
#        subtitle = "Modelo regresión lineal primavera ", y = "Residuo estandarizado", x = "Cuantil teó
#        caption="Fuente: Elaboración propia")+
#   theme_classic()
#
#
# homocedasticidad <-  acf(residuals(RegLin_Pri), plot = F)
#
# tibble(x = homocedasticidad$lag, y = homocedasticidad$acf) %>%
# ggplot(aes(x, y)) +
#        geom_hline(aes(yintercept = 0)) +
#        geom_segment(mapping = aes(xend = x, yend = 0))+
#   labs(title = "Gráfico 9. Autocorrelación de los errores",
#        subtitle = "Modelo regresión lineal primavera ", y = "Autocorrelación", x = "Lag",
#        caption="Fuente: Elaboración propia")+
#   theme_classic()
#
# Grafico8 <-ggplot(RegLin_Pri, aes(.hat, .cooksd))+
#   geom_point(na.rm=TRUE)+stat_smooth(method="loess", na.rm=TRUE)+
#   labs(subtitle = "Distancia de Cook vs apalancamiento",
#        x="Apalancamiento",
#        y = "Distancia de Cook")+
#   geom_abline(slope=seq(0,3,0.5), color="gray",
#               linetype="dashed")+
#   theme_classic()
#
# figure2 <- ggarrange(Grafico5, Grafico6, Grafico7, Grafico8 ,
#                      ncol=2,
#                      nrow = 2)
```

```
## Verano
Verano <- Tabla_entrenamiento %>% filter(Estacion == "Verano")

RegLin_Ver <- lm(data=Verano, formula = PRCP ~ VISIB + MXSPD + MIN + MAX + DEWP*SLP + TEMP*STP)
summary(RegLin_Ver)
summary(glm(data=Verano, formula = PRCP ~ VISIB + MXSPD + MIN + MAX + DEWP*SLP + TEMP*STP))
# # 0.3681 2476.6


Ver_newdata <- Tabla_testeo %>%
  filter(Estacion == "Verano") %>%
  select(c("VISIB", "MXSPD", "MIN", "MAX", "SLP", "DEWP", "TEMP", "STP"))
```

```r
predict(object = RegLin_Ver, newdata = Ver_newdata, interval = "confidence")

predict(object = RegLin_Ver, newdata = Ver_newdata, interval = "prediction")



# Usando Forward Stepwise Selection, R^2 y AIC
# RegLin_Ver <- lm(data=Verano, formula = PRCP ~ STP + VISIB + MXSPD + DEWP + TEMP + SLP
#+ MIN + MAX)
# summary(RegLin_Ver)
# summary(glm(data=Verano, formula = PRCP ~ STP + VISIB + MXSPD + DEWP + TEMP + SLP
#+ MIN + MAX))
# # 0.3511 2501.4
```

```r
#OTOÑO
Otonho <- Tabla_entrenamiento %>% filter(Estacion == "Otoño")

RegLin_Otonho <- lm(data=Otonho, formula = PRCP ~ SLP*DEWP + TEMP*STP + VISIB*WDSP +
                      DEWP*STP + SNDP)
summary(RegLin_Otonho)
summary(glm(data=Otonho, formula = PRCP ~ SLP*DEWP +
              TEMP*STP + VISIB*WDSP + DEWP*STP + SNDP))
# # 0.4398 1850.7


Oto_newdata <- Tabla_testeo %>%
  filter(Estacion == "Otoño") %>%
  select(c("VISIB", "WDSP", "SNDP", "SLP", "DEWP", "TEMP", "STP"))

predict(object = RegLin_Otonho, newdata = Oto_newdata, interval = "confidence")

predict(object = RegLin_Otonho, newdata = Oto_newdata, interval = "prediction")


# Usando Forward Stepwise Selection, R^2 y AIC
# RegLin_Otonho <- lm(data=Otonho, formula=PRCP~STP+MXSPD+DEWP+TEMP+SLP+WDSP)
# summary(RegLin_Otonho)
# summary(glm(data=Otonho, formula=PRCP~STP+MXSPD+DEWP+TEMP+SLP+WDSP))
# # 0.3877 1927.7
```

```r
## Invierno
Invierno <- Tabla_entrenamiento %>% filter(Estacion == "Invierno")

RegLin_Inv <- lm(data=Invierno, formula = PRCP ~ GUST + SLP*DEWP + TEMP*SLP +
                   TEMP*STP + VISIB*WDSP + SNDP)
summary(RegLin_Inv)
summary(glm(data=Invierno, formula = PRCP ~ GUST + SLP*DEWP + TEMP*SLP + TEMP*STP +
              VISIB*WDSP + SNDP))
# # 0.434 1382.1


Inv_newdata <- Tabla_testeo %>%
  filter(Estacion == "Invierno") %>%
  select(c("GUST", "VISIB", "WDSP", "SNDP", "SLP", "DEWP", "TEMP", "STP"))
```

```r
predict(object = RegLin_Inv, newdata = Inv_newdata, interval = "confidence")

predict(object = RegLin_Inv, newdata = Inv_newdata, interval = "prediction")



# Usando Forward Stepwise Selection, R^2 y AIC
# RegLin_Inv <- lm(data=Invierno, formula=PRCP~VISIB+WDSP+DEWP+TEMP+SNDP+GUST)
# summary(RegLin_Inv)
# summary(glm(data=Invierno, formula=PRCP~VISIB+WDSP+DEWP+TEMP+SNDP+GUST))
# # 0.3443 1499.5
```

```r
## GENERAL

RegLin_Com <- lm(data = Tabla_entrenamiento,
                 formula = PRCP ~ SLP*DEWP + TEMP*SLP +
      TEMP*STP + VISIB*WDSP + SNDP + MAX + MIN + GUST + MXSPD)
summary(RegLin_Com)
summary(glm(data = Tabla_entrenamiento,
                 formula = PRCP ~ SLP*DEWP + TEMP*SLP +
        TEMP*STP + VISIB*WDSP + SNDP + MAX + MIN + GUST + MXSPD))
# # 0.406 8084.2



Com_newdata <- Tabla_testeo %>%
  select(c("GUST", "VISIB", "WDSP", "SNDP", "SLP", "DEWP", "TEMP", "STP", "MXSPD", "MAX",
           "MIN"))

predict(object = RegLin_Com, newdata = Com_newdata, interval = "confidence")
predict(object = RegLin_Com, newdata = Com_newdata, interval = "prediction")



# Usando Forward Stepwise Selection, R^2 y AIC
# RegLin_Com <- lm(data=Tabla_entrenamiento,
#                  formula = PRCP ~ STP + VISIB + DEWP + TEMP + MXSPD + SLP + GUST
#+ MAX + MIN + SNDP)
# summary(RegLin_Com)
# summary(glm(data=Tabla_entrenamiento,
#             formula = PRCP ~ STP + VISIB + DEWP + TEMP + MXSPD + SLP + GUST +
#MAX + MIN + SNDP))
# # 0.3555 8377.2
```

Sección SVM

```r
#load("Datos_finales.RData")
#load("SVM.RData")

## Selección de modelos y validación cruzada

# df_svm_pri <- Tabla_entrenamiento %>%
#   filter(Estacion == "Primavera") %>%
```

```
#   mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
#   mutate(PRCP = as.factor(PRCP)) %>%
#   select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")
# df_sum_ver <- Tabla_entrenamiento %>%
#   filter(Estacion == "Verano") %>%
#   mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
#   mutate(PRCP = as.factor(PRCP)) %>%
#   select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")
# df_sum_oto <- Tabla_entrenamiento %>%
#   filter(Estacion == "Otoño") %>%
#   mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
#   mutate(PRCP = as.factor(PRCP)) %>%
#   select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")
# df_sum_inv <- Tabla_entrenamiento %>%
#   filter(Estacion == "Invierno") %>%
#   mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
#   mutate(PRCP = as.factor(PRCP)) %>%
#   select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")
#
#
# SVM_prim <- train(x = df_sum_pri %>% select(-PRCP),
#       y = df_sum_pri$PRCP, method = "svmRadial",
#       preProcess = "scale", probability = TRUE)
#
#
# SVM_ver <- train(x = df_sum_ver %>% select(-PRCP),
#       y = df_sum_ver$PRCP, method = "svmRadial",
#       preProcess = "scale")
#
# SVM_oto <- train(x = df_sum_oto %>% select(-PRCP),
#       y = df_sum_oto$PRCP, method = "svmRadial",
#       preProcess = "scale")
#
# SVM_inv <- train(x = df_sum_inv %>% select(-PRCP),
#       y = df_sum_inv$PRCP, method = "svmRadial",
#       preProcess = "scale")

#Selección de modelos manual backwise stepward
#svm_lineal <- svm(formula = PRCP ~ ., data = df_svm ,
                #  kernel = "linear")
#completo <- svm_lineal$decision.values

#norm(completo - svm_lineal$decision.values)

# Modelo primavera
Prim <- svm(x = df_svm_pri %>% select(-PRCP),
      y = df_svm_pri$PRCP,kernel = "radial",
    cost = as.numeric(SVM_prim$bestTune[2]), gamma = as.numeric(SVM_prim$bestTune[1]),
    probability = T)

## Modelo Verano
Ver <- svm(x = df_svm_ver %>% select(-PRCP),
      y = df_svm_ver$PRCP,kernel = "radial",
```

```
    cost = as.numeric(SVM_ver$bestTune[2]), gamma = as.numeric(SVM_ver$bestTune[1]),
    probability = T)

## Modelo otoño

Oto <- svm(x = df_svm_oto %>% select(-PRCP),
      y = df_svm_oto$PRCP,kernel = "radial",
    cost = as.numeric(SVM_oto$bestTune[2]), gamma = as.numeric(SVM_oto$bestTune[1]),
    probability = T)

## Modelo invierno

Inv <- svm(x = df_svm_inv %>% select(-PRCP),
      y = df_svm_inv$PRCP,kernel = "radial",
    cost = as.numeric(SVM_inv$bestTune[2]), gamma = as.numeric(SVM_inv$bestTune[1]),
    probability = T)


df_svmRa <- tibble(Parametro = c("zeta", "gamma",
                                 "Vectores de soporte" ),
    Primavera = c(as.numeric(c(SVM_prim$bestTune[2],
              SVM_prim$bestTune[1])), Prim$tot.nSV),
                  Verano = c(as.numeric(c(SVM_ver$bestTune[2],
                  SVM_ver$bestTune[1])), Ver$tot.nSV),
                  Otonho = c(as.numeric(c(SVM_oto$bestTune[2],
                  SVM_oto$bestTune[1])), Oto$tot.nSV),
                  Inv = c(as.numeric(c(SVM_inv$bestTune[2],
                  SVM_inv$bestTune[1])), Inv$tot.nSV ))
```

```
svm_test_pri <- Tabla_testeo %>%
  filter(Estacion == "Primavera") %>%
  mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
  mutate(PRCP = as.factor(PRCP)) %>%
  select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")

Prediction_pri <- predict(Prim, svm_test_pri %>% select(-PRCP), type = "prob",
                          probability = T)

svm_test_ver <- Tabla_testeo %>%
  filter(Estacion == "Verano") %>%
  mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
  mutate(PRCP = as.factor(PRCP)) %>%
  select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")

Prediction_ver <- predict(Ver, svm_test_ver %>% select(-PRCP),type = "prob",
                          probability = T)


svm_test_oto <- Tabla_testeo %>%
  filter(Estacion == "Otoño") %>%
  mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
  mutate(PRCP = as.factor(PRCP)) %>%
  select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")
```

```
Prediction_oto <- predict(Oto, svm_test_oto %>% select(-PRCP),
                          type = "prob",
                          probability = T)


svm_test_inv <- Tabla_testeo %>%
  filter(Estacion == "Invierno") %>%
  mutate(PRCP = ifelse(PRCP> 0 ,1,0))%>%
  mutate(PRCP = as.factor(PRCP)) %>%
  select("DEWP", "WDSP", "TEMP", "PRCP", "SLP", "VISIB", "MIN")

Prediction_inv <- predict(Inv, svm_test_inv %>% select(-PRCP),
                          type = "prob",
                          probability = T)

Matriz_prima <- confusionMatrix(Prediction_pri, reference = svm_test_pri$PRCP, positive = "1")

Matriz_ver <- confusionMatrix(Prediction_ver, reference = svm_test_ver$PRCP, positive = "1")

Matriz_oto <- confusionMatrix(Prediction_oto, reference = svm_test_oto$PRCP, positive = "1")

Matriz_inv <- confusionMatrix(Prediction_inv, reference = svm_test_inv$PRCP, positive = "1")


CM_estadisticos <-
  tibble(Estadistico = c("Accuracy" , "Kappa", "Sensitivity", "Specificity"),
         Primavera = c(Matriz_prima$overall[1:2], Matriz_prima$byClass[1:2] ),
          Verano = c(Matriz_ver$overall[1:2], Matriz_ver$byClass[1:2] ),
          Otonho = c(Matriz_oto$overall[1:2], Matriz_oto$byClass[1:2] ),
         Inv = c(Matriz_inv$overall[1:2], Matriz_inv$byClass[1:2]))
```

```
## ROC y área bajo la curva

Pred_prim <- prediction(attr(Prediction_pri, "probabilities")[,2], svm_test_pri$PRCP)

roc_prim <- performance(Pred_prim, "tpr","fpr")

Pred_ver <- prediction(attr(Prediction_ver, "probabilities")[,2], svm_test_ver$PRCP)

#roc_ver <- performance(Pred_ver, "tpr","fpr")

Pred_oto <- prediction(attr(Prediction_oto, "probabilities")[,2], svm_test_oto$PRCP)

#roc_oto <- performance(Pred_oto, "tpr","fpr")

Pred_inv <- prediction(attr(Prediction_inv, "probabilities")[,2], svm_test_inv$PRCP)

#roc_inv <- performance(Pred_inv, "tpr","fpr")


aucPrim <- performance(Pred_prim, measure = "auc")@y.values
aucVer <- performance(Pred_ver, measure = "auc")@y.values
aucInv <- performance(Pred_inv, measure = "auc")@y.values
aucOto <- performance(Pred_oto, measure = "auc")@y.values
```

```
Tabla_auc <- tibble(Estacion = c("Primavera", "Verano", "Invierno", "Otoño"),
        AUC = as.numeric(c(aucPrim, 1 - aucVer[[1]], aucInv, 1 - aucOto[[1]])))
```

Sección árboles

Preparación de los datos de Otoño

```
Tabla_entremaniento2<-Tabla_entrenamiento %>% filter(Estacion=="Otoño")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Tabla_entremaniento2$PRCP<-factor(Tabla_entremaniento2$PRCP)

OtonhoTest<-Tabla_testeo %>% filter(Estacion=="Otoño") %>%
  select("WDSP","PRCP","TEMP","GUST","DEWP","MXSPD","STP","MAX","MIN","VISIB","SLP")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
OtonhoTest$PRCP<-factor(OtonhoTest$PRCP)

Lluvia_Primavera<-Tabla_testeo %>% filter(Estacion=="Otoño") %>%
  select("PRCP") %>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Lluvia_Primavera$PRCP<-as.factor(Lluvia_Primavera$PRCP)
```

construccion del arbol

```
tree5<-rpart(PRCP~WDSP+TEMP+GUST+VISIB+MAX+DEWP+STP+SLP,Tabla_entremaniento2)

#plotcp(tree5)
```

optimizacion del arbol con best.tree.AIC.BIC

```
temp5<-best.tree.BIC.AIC(tree5, Tabla_entremaniento2, Y.name="PRCP",
                         X.names=c("TEMP","DEWP","STP","VISIB","WDSP","MXSPD","GUST","MAX","MIN","SLP")
                         family = "binomial", verbose = TRUE)
```

Evalución de resultados

```
Prediccion5 <- predict(temp5$tree$AIC,OtonhoTest)
Prediccion5<- ifelse(Prediccion5[,1] > Prediccion5[,2], 0 ,1 )
Prediccion5<-as.factor(Prediccion5)

matriz_Otonho<-confusionMatrix(Prediccion5, Lluvia_Primavera$PRCP)
```

Preparación de los datos de Primavera

```
Tabla_entremaniento2<-Tabla_entrenamiento %>% filter(Estacion=="Primavera")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Tabla_entremaniento2$PRCP<-factor(Tabla_entremaniento2$PRCP)

PrimaveraTest<-Tabla_testeo %>% filter(Estacion=="Primavera") %>%
  select("WDSP","PRCP","TEMP","GUST","DEWP","MXSPD","STP","MAX","MIN","VISIB","SLP")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
```

```
PrimaveraTest$PRCP<-factor(PrimaveraTest$PRCP)

Lluvia_Primavera<-Tabla_testeo %>% filter(Estacion=="Primavera") %>% select("PRCP") %>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Lluvia_Primavera$PRCP<-as.factor(Lluvia_Primavera$PRCP)
```

construccion del arbol

```
tree5<-rpart(PRCP~WDSP+TEMP+GUST+VISIB+MAX+DEWP+STP+SLP,Tabla_entremaniento2)

#plotcp(tree5)
```

optimizacion del arbol con best.tree.AIC.BIC

```
temp5<-best.tree.BIC.AIC(tree5, Tabla_entremaniento2, Y.name="PRCP",
                         X.names=c("TEMP","DEWP","STP","VISIB","WDSP","MXSPD","GUST","MAX","MIN","SLP")
                         family = "binomial", verbose = TRUE)
```

Evalución de resultados

```
Prediccion5 <- predict(temp5$tree$AIC,PrimaveraTest)
Prediccion5<- ifelse(Prediccion5[,1] > Prediccion5[,2], 0 ,1 )
Prediccion5<-as.factor(Prediccion5)

matriz_Primavera<-confusionMatrix(Prediccion5, Lluvia_Primavera$PRCP)
```

```
matriz_Primavera$overall
```

Preparación de los datos de Invierno

```
Tabla_entremaniento2<-Tabla_entrenamiento %>% filter(Estacion=="Invierno")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Tabla_entremaniento2$PRCP<-factor(Tabla_entremaniento2$PRCP)

InviernoTest<-Tabla_testeo %>% filter(Estacion=="Invierno") %>%
  select("WDSP","PRCP","TEMP","GUST","DEWP","MXSPD","STP","MAX","MIN","VISIB","SLP")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
InviernoTest$PRCP<-factor(InviernoTest$PRCP)

Lluvia_Invierno<-Tabla_testeo %>% filter(Estacion=="Invierno") %>%
  select("PRCP") %>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Lluvia_Invierno$PRCP<-as.factor(Lluvia_Invierno$PRCP)
```

construccion del arbol

```
tree5<-rpart(PRCP~WDSP+TEMP+GUST+VISIB+MAX+DEWP+STP+SLP,Tabla_entremaniento2)

#plotcp(tree5)
```

optimizacion del arbol con best.tree.AIC.BIC

```
temp5<-best.tree.BIC.AIC(tree5, Tabla_entremaniento2, Y.name="PRCP",
                         X.names=c("TEMP","DEWP","STP","VISIB","WDSP","MXSPD","GUST","MAX","MIN","SLP")
                         family = "binomial", verbose = TRUE)
```

Evalución de resultados

```
Prediccion5 <- predict(temp5$tree$AIC,InviernoTest)
Prediccion5<- ifelse(Prediccion5[,1] > Prediccion5[,2], 0 ,1 )
Prediccion5<-as.factor(Prediccion5)

matriz_Invierno<-confusionMatrix(Prediccion5, Lluvia_Invierno$PRCP)
```

```
matriz_Invierno$overall
```

Preparación de los datos de Verano

```
Tabla_entremaniento2<-Tabla_entrenamiento %>% filter(Estacion=="Verano")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Tabla_entremaniento2$PRCP<-factor(Tabla_entremaniento2$PRCP)

VeranoTest<-Tabla_testeo %>% filter(Estacion=="Verano") %>%
  select("WDSP","PRCP","TEMP","GUST","DEWP","MXSPD","STP","MAX","MIN","VISIB","SLP")%>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
VeranoTest$PRCP<-factor(VeranoTest$PRCP)

Lluvia_Verano<-Tabla_testeo %>%
  filter(Estacion=="Verano") %>%
  select("PRCP") %>%
  mutate(PRCP=ifelse(PRCP>0.00000000001,1,0))
Lluvia_Verano$PRCP<-as.factor(Lluvia_Verano$PRCP)
```

construccion del arbol

```
tree5<-rpart(PRCP~WDSP+TEMP+GUST+VISIB+MAX+DEWP+STP+SLP,Tabla_entremaniento2)

#plotcp(tree5)
```

optimizacion del arbol con best.tree.AIC.BIC

```
temp5<-best.tree.BIC.AIC(tree5, Tabla_entremaniento2, Y.name="PRCP",
                         X.names=c("TEMP","DEWP","STP","VISIB","WDSP","MXSPD","GUST","MAX","MIN","SLP")
                         family = "binomial", verbose = TRUE)
```

Evalución de resultados

```
Prediccion5 <- predict(temp5$tree$AIC,VeranoTest)
Prediccion5<- ifelse(Prediccion5[,1] > Prediccion5[,2], 0 ,1 )
Prediccion5<-as.factor(Prediccion5)

matriz_Verano<-confusionMatrix(Prediccion5, Lluvia_Verano$PRCP)
```

```
matriz_Verano$byClass
```

```
matriz_Otonho$byClass
matriz_Primavera$byClass
matriz_Invierno$byClass
matriz_Verano$byClass

matriz_Otonho$byClass
```

```
oto<-data.frame(Estacion="Otonho",Accuracy=matriz_Otonho$overall[1],Kappa=matriz_Otonho$overall[2],
        Sensitivity=matriz_Otonho$byClass[1],Specificity=matriz_Otonho$byClass[2],AIC=957.3,
        Parametro1=c("Temperatura máxima"),Parametro2=("Máxima rafaga de viento"))

inv<-data.frame(Estacion="Invierno",Accuracy=matriz_Invierno$overall[1],Kappa=matriz_Invierno$overall[2]
                Sensitivity=matriz_Invierno$byClass[1],
                Specificity=matriz_Invierno$byClass[2],AIC=825.7)

vera<-data.frame(Estacion="Verano",Accuracy=matriz_Verano$overall[1],Kappa=matriz_Verano$overall[2],
                 Sensitivity=matriz_Verano$byClass[1],
                 Specificity=matriz_Verano$byClass[2], AIC=1065)

prima<-data.frame(Estacion="Primavera",Accuracy=matriz_Primavera$overall[1],Kappa=matriz_Primavera$overa
                  Sensitivity=matriz_Primavera$byClass[1],
                  Specificity=matriz_Primavera$byClass[2],AIC=845.2)
```

Sección LDA

```
## Ajuste del modelo

library(MASS, quietly = T)

lda_prima <- lda(PRCP~., df_svm_pri)

detach("package:MASS", unload = TRUE)

LDA_pri_pre <- predict(lda_prima, svm_test_pri %>% select(-PRCP))

Matrix_LDA_pri <- confusionMatrix(LDA_pri_pre$class, svm_test_pri$PRCP)

Pred_lda_prim <- prediction(LDA_pri_pre$posterior[,2], svm_test_pri$PRCP)



Matrix_LDA__cuadro <- tibble(Estadistico =
  c("Accuracy" , "Kappa", "Sensitivity", "Specificity", "AUC"),
  Primavera = c(Matrix_LDA_pri$overall[1:2],
                Matrix_LDA_pri$byClass[1:2],
                auc_prim_lda@y.values[[1]]))


roc_prim_lda <- performance(Pred_lda_prim, "tpr","fpr")
```

```
SVM_Prim <- tibble(x = roc_prim@x.values[[1]],
                   y = roc_prim@y.values[[1]])
```

```
#Gráfico curva roc

tibble(x = roc_prim_lda@x.values[[1]], y = roc_prim_lda@y.values[[1]]) %>% ggplot(aes(x , y, colour = "
  geom_line()+
  geom_line(data = SVM_Prim, aes(x = x, y = y, colour = "SVM") )+
  labs(title = "Gráfico 7 Comparación Curva ROC LDA y SVM",
       subtitle = "Estación primavera",
       x = "Tasa falsos positivos", y = "Tasa verdaderos positivos", caption = "Elaboración propia", co
  theme_classic()
```