Doble Grado en
Matemáticas y Estadística

TRABAJO FIN DE GRADO

Estudio matemático-computacional de la producción oleícola en explotaciones agrarias

Autor: Manuel Bejarano Segado

Tutor: Luis Valencia Cabrera

Sevilla, Julio de 2023

Índice general

Prólogo

Este trabajo es fruto de cincos intensos años de carrera, que han culminado en seis meses de intenso estudio, tratamiento, desarrollo, investigación y redacción, con el fin de traer al mundo un trabajo que nos permita seguir ampliando nuestro conocimiento sobre la producción oleícola, el que considero un sector fundamental en la economía tanto agrícola como alimentaria, y de una indudable relevancia en nuestra región. Durante la etapa universitaria, la cual comenzó en septiembre de 2018, el autor siempre tuvo claro que este iba a ser el trabajo en el cual aplicar todas las herramientas y recursos que el doble grado en Matemáticas y Estadística le ofrecería. Hay que reconocer que desde ese entonces, el mundo y el autor de este trabajo, han cambiado enormemente, pero la motivación que le llevó a empezar este viaje, la de ser capaz de entender un poco más todo lo que nos rodea, se mantiene.

Quiero agradecer en primer lugar a mis padres, los cuales siempre han sido un pilar fundamental en mi vida, por su apoyo incondicional, por creer siempre en mí y por su eterna paciencia. También quisiera agradecer a dos personas que indudablemente marcaron mi vida estudiantil. El primero de ellos, Manolo, sin ti y sin tus consejos, no habría descubierto el mundo de las matemáticas. Fuiste un pilar fundamental en mi etapa anterior a la universidad y un excelente profesor de Matemáticas, pero sobre todo, una gran persona. El segundo, el tutor de este trabajo, Luis. Gracias por tu disposición a la hora de guiar este trabajo, por tus correcciones y recomendaciones, pero sobre todo gracias por ser, además de un excelente profesor, una gran persona. Gracias por siempre creer en mí y apoyarme en todas mis situaciones. Agradecer a también, a todos mis compañeros durante estos 5 años, haciendo especial mención a Alberto, Daniel y Miguel Ángel. Gracias por siempre estar a mi lado, tanto en los momentos buenos como en los malos. Sin cada uno de vosotros, no sería la persona que soy hoy en día.

Resumen

En este trabajo se ha llevado a cabo a través del lenguaje R, a través del entorno RStudio, un análisis exhaustivo de los datos de producción de aceite de oliva, utilizando métodos matemáticos y modelos computacionales avanzados. Se han aplicado técnicas de análisis de series temporales, modelización estadística y aprendizaje automático para comprender los factores que influyen en la producción y realizar predicciones precisas.

El enfoque multidisciplinario adoptado en este estudio ha permitido integrar conceptos y herramientas de las ciencias matemáticas y de la computación, en combinación con el conocimiento agrícola y los datos empíricos recopilados. Esto ha posibilitado una visión integral y enriquecedora de la producción oleícola, proporcionando nuevas perspectivas y oportunidades para mejorar los procesos y la toma de decisiones en el sector.

Se han explorado diferentes enfoques estadísticos y de aprendizaje automático para comprender la naturaleza de las variables, identificar interacciones significativas y desarrollar modelos predictivos precisos. Los resultados obtenidos han proporcionado un mayor conocimiento sobre el tema y han permitido realizar aportaciones relevantes al campo de estudio. Se han planteado las conclusiones derivadas del estudio y posibles trabajos futuros para ampliar la investigación en esta área.

Abstract

In this work, an exhaustive analysis of olive oil production data has been carried out using the R language, through the RStudio environment, using advanced mathematical methods and computational models. Time series analysis, statistical modeling and machine learning techniques have been applied to understand the factors influencing production and to make accurate predictions.

The multidisciplinary approach adopted in this study has allowed the integration of concepts and tools from mathematical and computational sciences, in combination with agricultural knowledge and empirical data collected. This has enabled a comprehensive and enriching vision of olive production, providing new perspectives and opportunities to improve processes and decision making in the sector.

Different statistical and machine learning approaches have been explored to understand the nature of variables, identify significant interactions and develop accurate predictive models. The results obtained have provided further knowledge on the subject and have allowed relevant contributions to be made to the field of study. Conclusions derived from the study and possible future work to expand research in this area have been presented.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. El cultivo del olivar en Arjonilla: tradición, excelencia y motor económico

El cultivo del olivar es una actividad agrícola de gran importancia en la región mediterránea, y la localidad de Arjonilla, situada en la provicia de Jaén, destaca por su larga tradición y excelencia en la producción de aceite de oliva.

Arjonilla, situada en pleno corazón de la comarca de La Campiña de Jaén, cuenta con un entorno geográfico y climático privilegiado para el desarrollo del olivar. La combinación de suelos fértiles, un clima mediterráneo con inviernos suaves y veranos calurosos, y una adecuada disponibilidad de agua, conforma condiciones ideales para el cultivo de olivares de alta calidad. Estas características han contribuido a que Arjonilla se haya ganado un reconocimiento destacado en el sector olivarero, siendo conocida por producir aceite de oliva de excelencia y por la dedicación de sus agricultores a la preservación de las técnicas tradicionales de cultivo.

En Arjonilla, se cultivan diferentes variedades de olivos, cada una con caraterísticas específicas que influyen en la calidad y el sabor del aceite de oliva producido. Estos tipos de olivar son el resultado de una selección cuidadosa y adaptada a las condiciones particulares de la zona, con el objetivo de obtener productos de alta calidad. Algunas de las variedades más destacadas presentes en Arjonilla son las siguientes:

1. Olivar de variedad “Picual”: esta variedad es la predominante en la región y en todo el territorio de Jaén. El olivar picual se caracteriza por su resistencia a condiciones climáticas adversas, como altas temperaturas y sequías, lo que lo convierte en una opción ideal para los terrenos situados en Arjonilla. El aceite de oliva elaborado a partir de las aceitunas de la variedad picual se distingue por su sabor afrutado y amargo, así como por su alta estabilidad y contenido de antioxidantes.

2. Olivar de variedad “Arberquina”: esta variedad, de origen catalán, ha encontrado un lugar destacado en los olivares de la zona. Estos producen aceitunas pequeñas y dulces, y su aceite de oliva se caracteriza por su suavidad y aroma frutal. La variedad Arberquina ha ganado popularidad debido a su adaptabilidad a diferentes condiciones de cultivo y a su creciente demanda en el mercado, tanto nacional como internacional.

3. Olivar de variedad “Hojiblanca”: esta variedad es reconocida por sus hojas de color plateado, de ahí su nombre. Destacan por la producción de aceitunas grandes y de alta productividad. El aceite de oliva otenido de esta variedad se caracteriza por su sabor suave y equilibrado, con notas a frutas maduras y un toque de amargor. Ha sido tradicionalmente cultivado en Srjonilla y sigue siendo una parte importante del patrimonio agrícola de la región.

El cultivo del olivar en Arjonilla no solo tiene un impacto económico significativo, sino que también es parte integral de la identidad y la cultura local. La tradición olivarera ha sido transmitida de generación en generación, y los agricultores de la zona han perfeccionado técnicas de cultivo, recolección y elaboración del aceite de oliva a lo largo de los años. La actividad olivarera en Arjonilla ha moldeado el paisaje, definiendo su estética y dotándolo de un encanto característico.

La producción de aceite de oliva en Arjonilla y en la provincia de Jaén en general es un motor económico fundamental. El sector genera empleo tanto en las labores de cultivo y recolección como en las actividades relacionadas con la transformación y comercialización del aceite de oliva. La cooperativa de Arjonilla desempeña un papel fundamental en la industria olivarera local, proporcionando un espacio de colaboración y apoyo a los agricultores y garantizando la calidad y trazabilidad de los productos.

1.2. Objetivos

El olivar es uno de los cultivos más emblemáticos de la región mediterránea y desempeña un papel crucial en la economía, la historia y la cultura de muchas localidades. España, en particular, es reconocida mundialmente como el mayor productor de aceite de oliva, y dentro de su territorio, la provincia de Jaén destaca por su tradición olivarera y la calidad de sus productos.

El objetivo de este trabajo de fin de grado es realizar un análisis exahustivo del sector olivarero, centrándonos en la cooperativa de Arjonilla como un caso de estudio representativo. La cooperativa de Arjonilla se ha posicionado como un referente en la producción y comercialización de aceite de oliva, y su éxito es testimonio de la importancia de la cooperación entre los agricultores y la gestión eficiente en la cadena de valor.

El olivar, además de ser una fuente de empleo y generador de riqueza, desempeña un papel vital en la preservación del medio ambiente y en el desarrollo sostenible de las zonas rurales. El cultivo del olivo se ha adaptado a lo largo de los siglos a las condiciones climáticas y geográficas de la región, convirtiéndose en un sitema agrícola resiliente y sostenible.

En este sentido, el presente trabajo de fin de grado abordará diversos aspectos relacionados con el olivar y la cooperativa de Arjonilla. Se analizarán las variables que afectan en la producción de aceituna y aceite de oliva, y se desarrollarán diferentes modelos de predicción basados en estas variables.

En el contexto del sector olivarero, comprender las variables que influyen en la producción de aceituna y aceite de oliva es esencial para optimizar los procesos agrícolas

y garantizar una mayor eficiencia en la obtención de productos de calidad. Estas variables pueden abarcar diversos aspectos, como condiciones climáticas, precipitaciones y prácticas agrícolas, entre otros.

El estudio de estas variables permitirá comprender cómo afectan a los rendimientos de la producción, así como a la calidad y cantidad de aceituna recolectada y el aceite de oliva obtenido. Además, se buscará identificar relaciones y patrones entre estas variables, para así desarrollar modelos de predicción que permitan estimar la producción de aceituna y aceite a partir de los valores de dichas variables.

Este enfoque tiene una relevancia significativa en el ámbito agrícola, ya que la capacidad de predecir la producción de aceituna y aceite de oliva permite a los productores y a la industria planificar de manera más precisa y tomar decisiones informadas sobre aspectos como la gestión de la cosecha, la planificación de la producción y la comercialización.

En este trabajo de fin de grado se recopilarán datos relevantes sobre estas variables, que incluirán registros climáticos, precipitaciones y datos de producción históricos de la cooperativa de Arjonilla.

1.3. Estructura del documento

Se ha dividido en 5 capítulos. En el primero de ellos, se hace una breve introducción de los objetivos y la motivación de este trabajo.

En el segundo capítulo, se realiza un exhaustivo análisis y descripción de todas las técnicas empleadas en el estudio. Se presentan de manera rigurosa y técnica los procedimientos y métodos utilizados, con el objetivo de proporcionar una comprensión profunda de las herramientas estadísticas y de análisis utilizadas en el estudio. Se abordan en detalle los fundamentos teóricos de cada técnica, se presentan las fórmulas matemáticas correspondientes y se explican las implicaciones y consideraciones relevantes en su aplicación.

En el tercer capítulo, se lleva a cabo un análisis descriptivo exhaustivo de las variables empleadas en la investigación. El objetivo principal de este análisis es comprender la naturaleza de estas variables, explorar sus interacciones y extraer conclusiones relevantes.

En el cuarto capítulo, se lleva a cabo la predicción del rendimiento y los kilogramos de aceituna recolectados mediante el uso de modelos predictivos. El objetivo principal de este capítulo es utilizar técnicas de modelado estadístico y de aprendizaje automático para desarrollar modelos que puedan estimar de manera precisa y fiable estas variables de interés.

En el quinto y último capítulo, se presentan las conclusiones derivadas del estudio realizado, donde se resumen los hallazgos más relevantes y se responden a los objetivos planteados inicialmente. Se analizan los resultados obtenidos a partir de las técnicas y metodologías aplicadas, y se discuten sus implicaciones en relación con el problema de investigación. Además de las conclusiones, se destacan las aportaciones realizadas por el estudio en el campo de estudio correspondiente. Se resaltan los aspectos innovadores, los conocimientos generados y las contribuciones específicas al cuerpo de conocimiento existente. Por último, se sugieren posibles trabajos futuros que podrían surgir a partir de las limitaciones identificadas durante el estudio. Se plantean nuevas preguntas de

investigación y se proponen áreas de estudio adicionales que podrían abordarse para profundizar en el tema y ampliar el alcance del conocimiento existente.

Capítulo 2

Preliminares

En esta sección, realizaremos una descripción detallada de las técnicas estadísticas y matemáticas que utilizaremos en nuestro estudio. Estas técnicas desempeñarán un papel fundamental en la comprensión y el análisis de los datos recopilados, permitiéndonos obtener información valiosa y realizar predicciones relevantes.

2.1. Análisis estadístico y transformaciones en estudios de distribuciones

En esta sección se abordan tres aspectos fundamentales en el análisis estadístico: los test de Shapiro-Wilk y Dickey-Fuller, utilizados para evaluar la normalidad de los datos y la existencia de raíces unitarias en series temporales, respectivamente, así como las transformaciones de Box-Cox, empleadas para corregir sesgos y mejorar la distribución de los datos

2.1.1. Test de Shapiro-Wilk

En el campo de la estadística, la prueba de Shapiro-Wilk se utiliza para evaluar si un conjunto de datos sigue una distribución normal. La hipótesis nula de esta prueba afirma que los datos provienen de una población con una distribución normal.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

donde:

- $x_{(i)}$ es el número que ocupa la i-ésima posición en la muestra (con la muestra ordenada de menor a mayor)
- $\bar{x}$ es la media muestral
- a_i $= \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}$

donde $m = (m_1, \ldots, m_n)^\top$ siendo $m_1, \ldots, m_n$ los valores medios del estadístico ordenado de variables aleatorias independientes e identicamente distribuidas, muestreadas de distribuciones normales y V denota la matriz de covarianzas de ese estadístico de orden.

El estadístico de la prueba se calcula utilizando la fórmula que se muestra anteriormente. Se compara la suma de los productos de los valores de la muestra ordenados con una función específica con la suma de las diferencias al cuadrado entre cada valor de la muestra y la media muestral. El resultado de este cálculo, denominado estadístico W, puede variar entre 0 y 1.

Para interpretar los resultados de la prueba, se compara el valor obtenido de W con los valores críticos correspondientes. Si el valor de W es demasiado pequeño en comparación con los valores críticos, se rechaza la hipótesis nula, lo que indica que los datos no siguen una distribución normal. Por el contrario, si el valor de W está cerca de 1 y no es significativamente diferente de los valores críticos, no se puede rechazar la hipótesis nula, lo que sugiere que los datos se ajustan a una distribución normal.

2.1.2. Transformación Box-Cox

Las transformaciones de Box y Cox son un conjunto de técnicas utilizadas en estadística para abordar diferentes problemas en el análisis de datos. Estas transformaciones se aplican con el objetivo de corregir sesgos en la distribución de errores, corregir varianzas desiguales y mejorar la linealidad en la relación entre variables.

La transformación potencial, utilizada en las transformaciones de Box y Cox, se define como una función continua que depende de un parámetro lambda (λ). Para aplicar esta transformación a un conjunto de datos $(Y_1, ..., Y_n)$, se realiza la transformación $Y_i' = Y_i^\lambda$ de la siguiente manera:

$$Y_i^{(\lambda)} = \begin{cases} K_1(Y_i^\lambda - 1) & \text{si } \lambda \neq 0, \\ K_2 \ln(Y_i) & \text{si } \lambda = 0 \end{cases}$$

Donde K_2 es la media geométrica de los valores $Y_1, ..., Y_n$.

$$K_2 = \left(\prod_{i=1}^{n} Y_i \right)^{1/n} = (Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n)^{1/n}$$

y K_1 es un parámetro que depende de k_2 y de λ, así:

$$K_1 = \frac{1}{\lambda \cdot K_2^{\lambda - 1}}$$

Para seleccionar el mejor valor de λ, primero se debe seleccionar un rango de valores de lambda λ de los cuales se quiere seleccionar el que logra que la transformación se acerque al máximo a los datos. Para cada valor de λ se realiza la transformación del paso anterior. Finalmente se sustituyen los valores de la o las variables explicativas en las diferentes funciones y se calculan los cuadrados de los residuales estadísticos. Aquella que tenga el menor valor de la suma de residuales será la mejor opción.

2.1.3. Test de Dickey-Fuller

La prueba de Dickey-Fuller es una prueba estadística utilizada para determinar la presencia o ausencia de raíces unitarias en una serie de tiempo. La hipótesis nula de esta prueba postula que la serie de tiempo tiene una raíz unitaria, lo que implica que es no estacionaria.

La prueba de Dickey-Fuller se basa en un modelo de regresión que considera la serie de tiempo y su diferencia. El modelo se puede expresar de la siguiente manera:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \cdots + \delta_p \Delta Y_{t-p} + \varepsilon_t$$

donde:

- Y_t representa la serie de tiempo original
- ΔY_t es la diferencia entre el valor de la serie de tiempo en el tiempo t y su valor en el tiempo t-1.
- α, β y γ son los coeficientes de la regresión
- $\delta_1, \delta_2, ..., \delta_p$ son los coeficientes que representan los efectos de las diferencias pasadas.
- ε_t es el término de error

El objetivo de la prueba es determinar si el coeficiente γ es significativamente diferente de cero. Si el coeficiente γ es cero, indica la presencia de una raíz unitaria y la serie de tiempo es no estacionaria. Por otro lado, si el coeficiente γ es significativamente diferente de cero, se rechaza la hipótesis nula de la presencia de una raíz unitaria y se concluye que la serie de tiempo es estacionaria.

La prueba de Dickey-Fuller utiliza estadísticos de prueba basados en la estimación de los coeficientes de regresión. El estadístico más comúnmente utilizado es el estadístico t, que se calcula dividiendo el coeficiente γ por su error estándar. Si el valor absoluto del estadístico t es mayor que el valor crítico correspondiente, se rechaza la hipótesis nula.

2.2. Modelos y técnicas aplicadas en el estudio

En esta sección, se presentan los modelos estadísticos y matemáticos que serán aplicados en el estudio. Estos modelos representan herramientas fundamentales para analizar y comprender los datos recopilados, así como para realizar predicciones y extraer conclusiones significativas.

2.2.1. Serie Temporal

Una serie temporal se define (ver con más detalle en el libro Series Temporales de González Velasco and Inés del Puerto García [2009]) como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados. Si los datos se recogen en instantes temporales de forma continua, se debe o bien digitalizar la serie, es decir, recoger sólo los valores

en instantes de tiempo equiespaciados, o bien acumular los valores sobre intervalos de tiempo.

Una serie temporal puede tener observaciones discretas o continuas, lo cual depende de la forma en que se registran los valores a lo largo del tiempo.

Si los valores de la serie temporal pueden ser predichos de manera precisa, se dice que la serie es determinística. Esto significa que los valores futuros se pueden calcular sin incertidumbre utilizando solo las observaciones pasadas.

Por otro lado, si los valores futuros de la serie temporal no pueden ser determinados de manera precisa y solo se pueden estimar parcialmente a partir de las observaciones pasadas, se considera que la serie es estocástica. En este caso, los valores futuros están sujetos a una distribución de probabilidad condicionada a los valores pasados. Esto implica que existe cierta incertidumbre en la predicción de los valores futuros.

El estudio descriptivo de series temporales se basa en descomponer la variación de la serie en diferentes componentes fundamentales. Este enfoque es especialmente útil cuando se observa una cierta tendencia o periodicidad en la serie. Es importante tener en cuenta que esta descomposición no es única, es decir, puede haber diferentes formas de identificar y separar estas componentes.

Las componentes o fuentes de variación más comunes son las siguientes:

- Tendencia: Se refiere a los cambios a largo plazo que ocurren en la serie en relación a su nivel medio o la evolución a largo plazo de la media. La tendencia se caracteriza por un movimiento suave y gradual de la serie a lo largo del tiempo.
- Efecto estacional: Muchas series temporales exhiben patrones periódicos o variaciones que se repiten en ciertos períodos, como anualmente o mensualmente. Estos efectos estacionales son fácilmente identificables y se pueden medir explícitamente o incluso eliminar de los datos originales mediante técnicas de desestacionalización.
- Componente aleatoria: Una vez que se han identificado y eliminado los componentes anteriores, lo que queda son valores residuales que se consideran aleatorios. El objetivo es estudiar el comportamiento de esta componente aleatoria, utilizando algún tipo de modelo probabilístico que describa su variabilidad.

De las tres componentes reseñadas, las dos primeras son componentes determinísticas, mientras que la última es aleatoria. Así, se puede denotar que:

$$X_t = T_t + E_t + I_t$$

donde:

T_t es la tendencia
E_t es la componente estacional
I_t es el ruido o parte aleatoria

Una vez que se ha completado el análisis descriptivo, es necesario desarrollar un modelo estadístico que describa de manera precisa el comportamiento estocástico del proceso

subyacente en la serie temporal. Este modelo debe ser capaz de capturar las características observadas en los datos y, al mismo tiempo, ser consistente con las implicaciones teóricas del fenómeno estudiado. Procedemos a introducir la familia de modelos ARIMA, los cuales son ampliamente utilizados en el análisis de series temporales. Estos modelos permiten modelar la dependencia temporal de los datos, teniendo en cuenta la tendencia, la estacionalidad y los componentes aleatorios presentes en la serie.

Definición 2.1. Un **modelo** para un proceso estocástico es cualquier conjunto de hipótesis bien definidas sobre las propiedades estadísticas de dicho proceso que usualmente se expresa mediante una ecuación de recurrencia.

En el análisis de series temporales, el modelo de medias móviles es un enfoque común utilizado para modelar datos univariantes. Este modelo establece una relación lineal entre la variable de estudio y los errores de predicción pasados.

Definición 2.2. Se dice que un proceso estocástico Y_t es un proceso $MA(q)$ media móvil de orden q si

$$Y_t = c + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \ldots + \theta_q\epsilon_{t-q}$$

$$\forall t = 0, \pm1, \pm2, \ldots \quad \text{y } \{\varepsilon_t\} \sim N(0, \sigma^2)$$

Los modelos autoregresivos buscan describir la variable de interés en el momento t como una combinación lineal de sus valores pasados. El término autoregresión hace referencia a que el modelo es una regresión lineal de la variable de estudio sobre sí misma.

Definición 2.3. Un proceso estocástico $\{Y_t\}$ describe un modelo AR(p) autorregresivo de orden p si

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$
$$\forall t = 0, \pm1, \pm2, \ldots \quad \text{y } \{\varepsilon_t\} \sim N(0, \sigma^2)$$

Definición 2.4. Un proceso estocástico es ruido blanco y se denota $\{Y_t\} \sim N(0, \sigma^2)$ si se trata de una secuencia de variables aleatorias idéntica e independientemente distribuidas tal que

$-E[Y_t] = 0 \quad \forall t = 0, \pm1, \pm2, \ldots$
$-Var[Y_t] = \sigma^2 \quad \forall t = 0, \pm1, \pm2, \ldots$
$-Cov[Y_t, Y_{t+h}] = 0 \quad \forall t = 0, \pm1, \pm2, \ldots$ y cualquier h $\neq 0$

Definición 2.5. Un proceso estocástico $\{Y_t\}$ es un proceso ARMA(p, q) si es débilmente estacionario y para cualquier $t = 0, \pm1, \pm2, \ldots$

$$Y_t - \phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p} = c + \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_q\epsilon_{t-q}$$

donde

$-\{\varepsilon_t\} \sim N(0, \sigma^2)$ es un ruido blanco
$-\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ es el polinomio autorregresivo
$-\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ es el polinomio de medias móviles

y los polinomios autorregresivo y media móvil no tienen factores en común.

Definición 2.6. Sea $\{Y_t\}$ un proceso $ARMA(p,q)$. Se dice que $\{Y_t\}$ es estacionario si existen las constantes $\{\psi_j\}$ tales que $\sum_{j=0}^{\infty}|\psi_j| < \infty$ y para cualquier t se cumple

$$\sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j} = \frac{\theta(z)}{\phi(z)}, |z| \leq 0$$

Definición 2.7. Sea d un entero no negativo.Se dice que $\{Y_t\}$ es un proceso $ARIMA(p,d,q)$ si la serie $X_t = (1-B)^d Y_t$ es un proceso $ARMA(p,q)$ estacionario. Esta definición implica que la serie $\{Y_t\}$ satisface la relación

$$\phi(B)(1-B)^d Y_t = c + \theta(B)\varepsilon_t$$

Donde $\phi(z)$ y $\theta(z)$ son los polinomios autorregresivo y media móvil de grado p y q respectivamente, tales que no tienen raíces en común y $\{\varepsilon_t\} \sim N(0,\sigma^2)$. Además, debe notarse que el proceso es estacionario si y solo si $d = 0$, en cuyo caso se trataría de un modelo $ARMA(p,q)$.

Definición 2.8. Si d y D son enteros no negativos, entonces se dice que $\{Y_t\}$ es un modelo $ARIMA(p,d,q)\ddot{O}(P,D,Q)_s$ estacional de periodo S si la serie $X_t = (1-B)^d(1-B^s)^D Y_t$ es un proceso $ARMA$ estacionario definido por

$$\phi(B)\Phi(B_s)X_t = c + \theta(B)\Theta(B^s)\varepsilon_t$$

donde

$-\{\varepsilon_t\} \sim N(0,\sigma^2)$ es un ruido blanco
$-\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ es el polinomio autorregresivo
$-\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ es el polinomio de medias móviles
$-\Phi(z) = 1 - \Phi_1 z - \ldots - \Phi_P z^P$ es el polinomio autorregresivo estacional
$-\Theta(z) = 1 + \Theta_1 z + \ldots + \Theta_Q z^Q$ es el polinomio media móvil estacional

2.2.2. Análisis de Componentes Principales

Para analizar las interacciones entre p variables correlacionadas que miden información común, es posible transformar el conjunto original de variables en uno nuevo. Este nuevo conjunto de variables se conoce como conjunto de componentes principales, el cual se caracteriza por ser un conjunto de variables no correlacionadas entre sí, sin repetición ni redundancia en la información.

En el proceso de construcción de las componentes principales, las nuevas variables se obtienen como combinaciones lineales de las variables originales. Estas combinaciones se realizan de manera que se preserve el orden de importancia en cuanto a la variabilidad total de la muestra. El objetivo es encontrar un conjunto de m variables (donde m es menor que p) que sean combinaciones lineales de las p variables originales y que estén incorrelacionadas entre sí. Además, se busca que estas nuevas variables capturen la mayor parte de la información o variabilidad presente en los datos.

Si las variables originales están incorrelacionadas desde el principio, no tiene sentido aplicar el análisis de componentes principales. Esta técnica se utiliza cuando hay correlación entre las variables, ya que busca encontrar nuevas variables que sean combinaciones lineales de las originales y que estén incorrelacionadas entre sí, sirviendo tanto para evitar el efecto duplicado o aumentado de determinados elementos como para reducir la dimensionalidad. Sin embargo, es importante destacar que el análisis de componentes principales no requiere la suposición de normalidad multivariante de los datos. Aunque si los datos cumplen con esta suposición, se puede obtener una interpretación más profunda de los componentes principales.

En el análisis de componentes principales, se parte de un conjunto de variables $(x_1, x_2, ..., x_p)$ que están asociadas a un grupo de objetos o individuos. El objetivo es calcular un nuevo conjunto de variables $(y_1, y_2, ..., y_m)$ que estén incorrelacionadas entre sí y cuyas varianzas disminuyan gradualmente.

Cada y_j (donde $j = 1, ..., m$) es una combinación lineal de las $x_1, x_2, ..., x_p$ originales, es decir:

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p = \mathbf{a}_j^\top \mathbf{x}$$

siendo $\mathbf{a}_0 j = (a_{1j}, a_{2j}, \ldots, a_{pj})$ un vector de constantes y $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$

Dado que deseamos maximizar la varianza, una estrategia simple podría ser incrementar los coeficientes a_{ij}. Para preservar la ortogonalidad de la transformación, se impone que

$$\mathbf{a}_j^\top \mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1$$

El primer componente se calcula eligiendo a_1 de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción de que $\mathbf{a}_1^\top \mathbf{a}_1 = 1$. El segundo componente principal se calcula obteniendo a_2 de modo que la variable obtenida, y_2 esté incorrelada con y_1.Del mismo modo se eligen $y_3, y_4, ..., y_m$, incorrelados entre sí, de manera que las variables aleatorias obtenidas vayan teniendo cada vez menor varianza.

2.2.3. Modelos de aprendizaje supervisado

En esta sección, presentaremos los tres modelos de aprendizaje supervisado que formarán parte de nuestro estudio. Estos modelos son ampliamente utilizados en el campo del aprendizaje automático y nos permitirán realizar predicciones precisas.

El aprendizaje supervisado es un enfoque del aprendizaje automático que se basa en el uso de modelos para aprender de un conjunto de datos. En este proceso, se busca establecer una relación entre las variables predictoras y la variable objetivo, con el objetivo de poder predecir o clasificar nuevas instancias.

En el aprendizaje supervisado, se parte de un conjunto de datos de entrenamiento en el que se conocen tanto las variables predictoras como la variable objetivo correspondiente para cada instancia. Estos datos de entrenamiento se utilizan para ajustar o entrenar

el modelo, de modo que este pueda aprender la función subyacente que relaciona las variables predictoras con la salida objetivo.

Durante el entrenamiento del modelo, se ajustan los parámetros o pesos del modelo para minimizar el error entre las predicciones del modelo y las salidas reales en el conjunto de entrenamiento. Esto se logra mediante algoritmos de optimización que buscan encontrar los valores óptimos de los parámetros del modelo.

Una vez entrenado, el modelo puede ser utilizado para hacer predicciones o clasificar nuevas instancias proporcionando las variables predictoras correspondientes. Esto implica aplicar la función aprendida durante el entrenamiento a las nuevas instancias para obtener las salidas predichas.

2.2.3.1. Algoritmo KNN (K-Nearest Neighbors)

El método de los k vecinos más cercanos (k-nearest neighbors o k-NN) es un algoritmo de aprendizaje supervisado. En lugar de hacer suposiciones sobre la distribución de los datos, este método se basa en el conjunto de entrenamiento y utiliza la información de los prototipos existentes para realizar las predicciones.

En nuestro caso, el algoritmo k-NN se utiliza en regresión para predecir el valor de una variable objetivo para una instancia de prueba. Se calcula un valor numérico basado en los k vecinos más cercanos en el conjunto de entrenamiento. La predicción se obtiene tomando el promedio de los valores de la variable objetivo de estos vecinos más cercanos. La cercanía se determina mediante una medida de distancia, como la distancia euclidiana, la distancia Manhattan o la distancia de Hamming, entre las características de las instancias.

Este método es no paramétrico, lo que significa que no asume ninguna forma particular para la distribución de los datos. En cambio, se basa en la información proporcionada por los prototipos existentes en el conjunto de entrenamiento. Al seleccionar un valor adecuado para k, se puede controlar la suavidad o la granularidad de las fronteras de decisión.

En el algoritmo de k-NN en regresión, los ejemplos de entrenamiento se representan como vectores en un espacio característico multidimensional. Cada ejemplo está descrito en términos de p atributos, y se considera una variable objetivo continua en lugar de clases. Los valores de los atributos del i-ésimo ejemplo (donde $1 \leq i \leq n$) se representan mediante un vector de dimensión p $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi}) \in X$.

El objetivo del algoritmo de k-NN en regresión es predecir el valor de la variable objetivo para una instancia de prueba. Para esto, se calcula la distancia entre los vectores de las instancias de entrenamiento y la instancia de prueba utilizando una medida de distancia como la distancia euclidiana.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}$$

Luego, se seleccionan los k ejemplos de entrenamiento más cercanos a la instancia de prueba. La predicción se obtiene tomando el promedio de los valores de la variable objetivo de estos k ejemplos de entrenamiento más cercanos. Esto proporciona una estimación del valor de la variable objetivo para la instancia de prueba en función de los valores de las instancias de entrenamiento más similares.

2.2.3.2. Árboles de decisión

En esta sección, se abordará el uso de los árboles de decisión, que pueden emplearse como modelos predictivos por sí solos o, para ganar más robustez, como elementos constituyentes básicos de los bosques aleatorios, como vemos en la siguiente sección y hemos usado en profundidad en este trabajo. Los árboles de decisión son modelos de clasificación y regresión que se basan en la estructura de un árbol, donde cada nodo representa una característica o atributo y cada borde representa una decisión o regla basada en el valor de ese atributo.

Los árboles de decisión se construyen mediante la partición recursiva del conjunto de datos de entrenamiento, seleccionando en cada paso el atributo que mejor divide los datos en función de alguna medida de impureza, como la ganancia de información o el índice de Gini.

2.2.3.2.1. Random Forest

El algoritmo de bosques aleatorios es un método de aprendizaje supervisado utilizado para problemas de clasificación y regresión. Se basa en la construcción de múltiples árboles de decisión para generar un modelo predictivo robusto.

En el caso de clasificación, el algoritmo Random Forest construye un conjunto de árboles de decisión independientes entre sí. Cada árbol se entrena utilizando una muestra aleatoria con reemplazo del conjunto de datos de entrenamiento original. Además, en cada nodo del árbol, se selecciona un subconjunto de características para realizar la división. Esto se conoce como "bagging" y ayuda a reducir la correlación entre los árboles y mejorar la generalización.

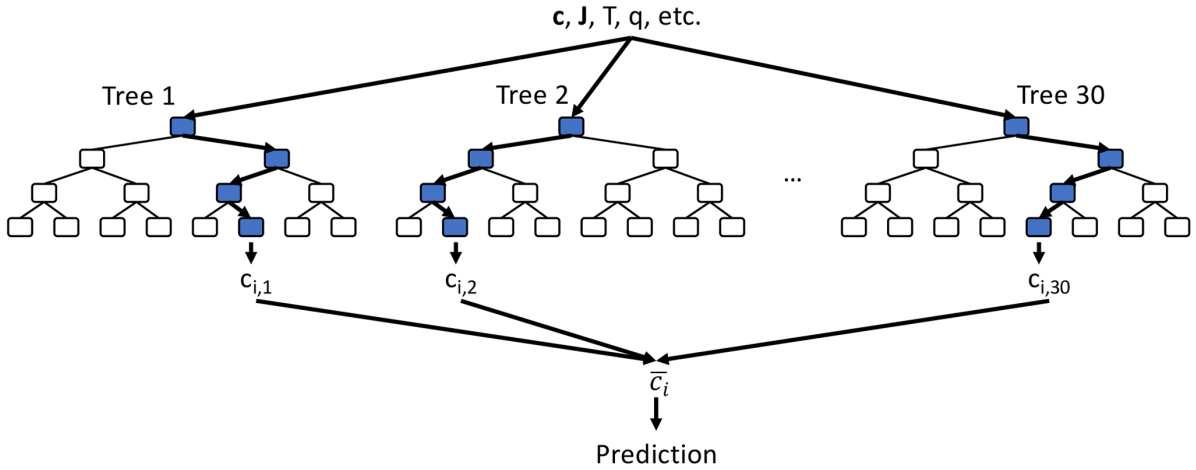


Figura 2.1: Random Forest (fuente: Application of random forest regression to the calculation of gas-phase chemistry within the geos-chem chemistry model)

Durante la fase de clasificación, cada árbol en el conjunto de Random Forest emite una predicción de clase y se utiliza un esquema de votación (mayoría de votos) para determinar la clase final asignada a la instancia de prueba.

En el caso de la regresión, el algoritmo Random Forest también construye un conjunto de árboles de decisión. Sin embargo, en lugar de utilizar votación, se utiliza el promedio de las predicciones de los árboles para estimar el valor objetivo de la instancia de prueba.

2.2.3.3. Red Neuronal

Una red neuronal es un modelo de aprendizaje automático inspirado en la estructura y funcionamiento del cerebro humano. Consiste en un conjunto interconectado de unidades de procesamiento llamadas neuronas artificiales, que trabajan en conjunto para resolver problemas complejos de clasificación, regresión o reconocimiento de patrones.

Cada neurona artificial en una red neuronal está asociada con un conjunto de pesos sinápticos, que determinan la importancia relativa de las entradas recibidas. Las neuronas reciben entradas de otras neuronas o directamente del entorno, es decir, la entrada, realizan cálculos en función de los pesos sinápticos y aplican una función de activación para producir una salida. Las funciones de activación más utilizadas son las siguientes:

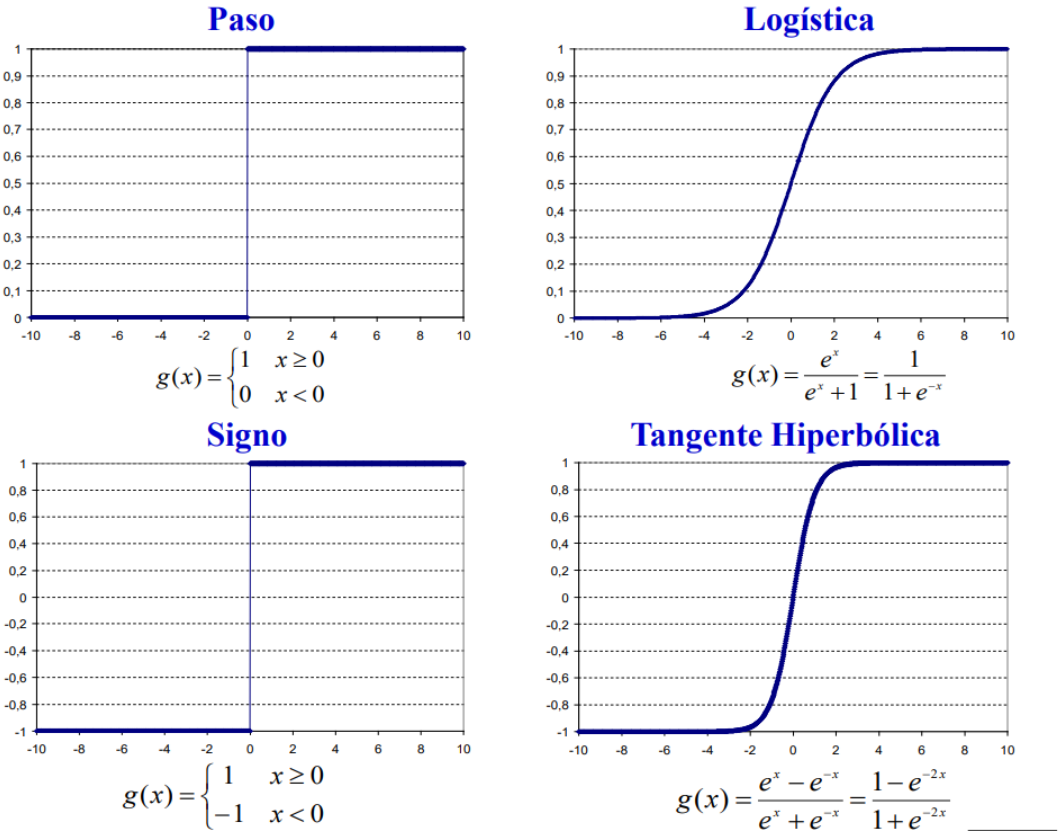


Figura 2.2: Funciones de activación (fuente: elaboración propia)

Las redes neuronales se organizan en capas, donde las neuronas de una capa se conectan con las neuronas de la capa siguiente. La capa de entrada recibe los datos de entrada y la capa de salida produce las respuestas o predicciones finales. Entre estas capas de entrada y salida, pueden existir una o varias capas ocultas, que proporcionan la capacidad de aprendizaje y generalización de la red.

Una Red Neuronal Artificial multicapa hacia delante (típicamente referida en inglés como multilayer perceptron o feedforward) es una Red Neuronal Artificial cuyos elementos

de procesamiento están organizados en capas o niveles sucesivos, y de forma que, una vez ordenadas las capas de izquierda a derecha, solo existen conexiones entre nodos de niveles sucesivos, en el sentido izquierda-derecha.

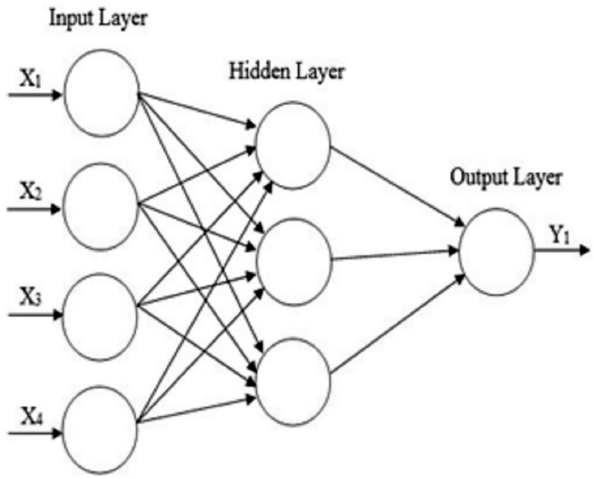


Figura 2.3: Red Neuronal (fuente: Artificial neural network with taguchi method for robust classification model to improve classification accuracy of breast cancer)

Dada una Red Neuronal Artificial, se llama algoritmo, método o regla de aprendizaje a cualquier algoritmo que permita obtener una asignación de valores para cada uno de los coeficientes sinápticos (en definitiva, encontrar la combinación de pesos que minimizan el error que cometemos sobre cada salida conocida a partir de sus entradas conocidas). El objetivo final, naturalmente, será desentrañar la función no lineal que hay detrás de las combinaciones de variables de entrada para obtener la salida, de modo que podamos predecir a futuro la salida esperada para una entrada desconocida, simplemente propagando los productos de las entradas por los pesos y la aplicación de las sucesivas funciones de activación.

2.3. Software empleado

En esta sección, se presentará una descripción detallada del software utilizado para el desarrollo del trabajo. El lenguaje de programación R y el entorno de desarrollo RStudio han sido las principales herramientas utilizadas para llevar a cabo las distintas tareas de análisis de datos, modelado estadístico y generación de resultados.

A lo largo de esta sección, se irán describiendo los diferentes paquetes y funciones más relevantes utilizados en el trabajo. Se destacarán aquellas herramientas que han sido clave en las distintas etapas del estudio, como la exploración y preparación de los datos, la implementación de los modelos y algoritmos, y la generación de gráficos y visualizaciones para la interpretación de los resultados.

Para realizar el análisis de la serie temporal utilizamos las siguientes librerías:

- TSA: esta librería se enfoca en el análisis de series de tiempo en R. Se utiliza la función BoxCox.ar() para realizar una transformación de Box-Cox para estabilizar la varianza de la serie temporal.

- tseries: esta librería se utiliza para realizar pruebas de estacionariedad en series de tiempo. Se utiliza la función adf.test() para realizar una prueba de Dickey-Fuller aumentada y evaluar si la serie temporal es estacionaria.

- forecast: esta librería se especializa en pronósticos y modelos de series de tiempo. Se utiliza la función auto.arima() para ajustar automáticamente un modelo ARIMA a la serie transformada y la función predict() para realizar predicciones futuras basadas en el modelo ajustado.

Para aplicar las técnicas de análisis de componentes principales, algoritmo KNN y random forest utilizaremos la librería Tidymodels. La librería Tidymodels es una poderosa herramienta para la creación de modelos predictivos en R. Proporciona una serie de funciones y utilidades que facilitan el proceso de construcción, entrenamiento y evaluación de modelos. A continuación, se presentan algunas de la principales funciones y conceptos ofrecidos por Tidymodels:

- Recetas (Recipes): las recetas son objetos que especifican cómo preprocesar los datos antes de alimentarlos al modelo. Podemos utilizar funciones como recipe() para definir una receta, y luego agregar pasos de preprocesamiento utilizando funciones como step_center(), step_scale(), step_log(), entre otras. Las recetas permiten estandarizar, transformar y seleccionar características de manera flexible y estructurada.

- Flujos de trabajo (Workflows): los flujos de trabajo son objetos que combinan una receta con un modelo para crear un flujo de trabajo completo. Podemos utilizar funciones como workflow() para crear un flujo de trabajo y luego agregar la receta y el modelo utilizando las funciones add_recipe() y add_model(), respectivamente. Los flujos de trabajo permiten encadenar de manera ordenada las etapas de preprocesamiento y modelado.

- Métodos de modelado: tidymodels proporciona una amplia gama de algoritmos de modelado que se pueden utilizar para construir modelos predictivos. Algunos ejemplos incluyen:

  - Modelos lineales: utiliza la función linear_reg() para crear modelos de regresión lineal.
  - Random Forest: utiliza la función rand_forest() para crear modelos basados en Random Forest.
  - Algoritmo KNN: utiliza la función nearest_neighbor() para crear modelos basados en el algortimo KNN.

- Validación cruzada (Cross-validation): tidymodels ofrece funciones para realizar la validación cruzada, una técnica que divide los datos en múltiples conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo de manera más robusta. Podemos utilizar funciones como vfold_cv() para crear particiones de validación cruzada y fit_resamples() para ajustar el modelo y recopilar las métricas de evaluación.

- Evaluación de modelos: la librería Tidymodels proporciona funciones para evaluar el rendimiento de los modelos utilizando diversas métricas. Podemos utilizar funciones como collect_metrics() para recopilar y visualizar las métricas de evaluación, como el error cuadrático medio (RMSE), la precisión, la sensibilidad, etc y funciones como collect_predictions() para recopilar las predicciones realizadas por el modelo.

Para crear los modelos de redes neuronales utilizaremos las librerías Neuralnet y Keras. La librería Neuralnet proporciona varias funciones para crear y entrenar modelos de redes neuronales en R. A continuación, se describen las principales funciones que ofrece esta librería:

- neuralnet(): se utiliza para crear un modelo de redes neuronales. Toma como argumentos la fórmula que especifica la relación entre la variable objetivo y las variables predictoras, los datos de entrenamiento, la estructura de capas ocultas de la red neuronal, el algoritmo de entrenamiento y otros parámetros opcionales. El resultado es un objeto de modelo que puede ser utilizado para hacer predicciones y evaluar el rendimiento.
- compute(): se utiliza para realizar predicciones utilizando un modelo de redes neuronales previamente entrenado. Toma como argumentos el modelo y los datos de prueba. Devuelve un objeto que contiene los valores predichos para la variable objetivo.
- algorithm: permite especificar el algoritmo de entrenamiento a utilizar. Algunos de los algoritmos disponibles son "rprop+" (propagación hacia atrás con actualización de peso resiliente), "backpropagation" (propagación hacia atrás clásica) y "slp" (perceptrón simple).
- hidden: este parámetro de la función neuralnet() permite especificar la estructura de capas ocultas de la red neuronal. Se proporciona un vector que indica el número de neuronas en cada capa oculta. Por ejemplo, hidden = c(16, 8) indica una red con dos capas ocultas, la primera con 16 neuronas y la segunda con 8 neuronas.
- startweights: permite especificar los pesos iniciales de la red neuronal. Se puede proporcionar una matriz de pesos inicializados manualmente para ajustar el proceso de entrenamiento.

La librería Keras es una librería de aprendizaje profundo de alto nivel. Proporciona una interfaz sencilla y eficiente para la construcción y entrenamiento de modelos de redes neuronales. Algunas de las principales funciones ofrecidas por Keras son las siguientes:

- keras_model_sequential(): crea un modelo de red neuronal secuencial. Un modelo secuencial es aquel en el que las capas se apilan secuencialmente una encima de la otra.
- layer_normalization(): añade una capa de normalización a un modelo de red neuronal. La normalización se utiliza para asegurar que las entradas de la capa se encuentren en la misma escala y facilitar el entrenamiento del modelo.

- layer_dropout(): añade una capa de dropout a un modelo de red neuronal. El dropout es una técnica de regularización que ayuda a prevenir el sobreajuste al eliminar aleatoriamente conexiones entre neuronas durante el entrenamiento.
- layer_dense(): añade una capa densa (totalmente conectada) a un modelo de red neuronal. En una capa densa, cada neurona se conecta con todas las neuronas de la capa anterior.
- compile(): compila el modelo de red neuronal especificando la función de pérdida y el optimizador a utilizar durante el entrenamiento. La función de pérdida mide qué tan bien se están haciendo las predicciones del modelo, mientras que el optimizador se encarga de ajustar los pesos de las conexiones para minimizar la pérdida.
- optimizer_adam(): configura el optimizador Adam, que es un algoritmo de optimización popular utilizado en el entrenamiento de redes neuronales. Se puede ajustar la tasa de aprendizaje a través del parámetro learning_rate.
- fit(): entrena el modelo utilizando los datos de entrenamiento. Se especifican las variables predictoras (x), la variable objetivo (y), el número de épocas de entrenamiento y la proporción de datos de validación.
- evaluate(): evalúa el modelo utilizando los datos de prueba. Calcula la pérdida del modelo en los datos de prueba para evaluar su rendimiento.
- predict(): realiza predicciones utilizando el modelo entrenado. Se proporcionan las variables predictoras y devuelve las predicciones del modelo.

Estas son algunas de las funciones más relevantes de la librería Keras que utilizaremos en nuestro código. Keras ofrece muchas más funciones y opciones para la construcción y entrenamiento de modelos de redes neuronales, lo que la convierte en una herramienta poderosa para el desarrollo de aplicaciones de aprendizaje profundo, incluyendo muchos tipos de capas distintas como las de regularización, o conjuntos de capas de neuronas preentrenadas para poder hacer transfer learning.

Capítulo 3

Análisis descriptivo de las variables empleadas en el estudio

En esta sección se llevará a acabo un análisis descriptivo de los datos relacionados con el cultivo del olivar y la producción de aceituna y aceite en la cooperativa de Arjonilla. Para este análisis, se cuenta con un conjunto de datos recopilados a lo largo del tiempo, que abarcan desde el año 2001 hasta la actualidad.

Entre los datos recopilados, se disponen de registros de precipitaciones y temperaturas máximas y mínimas, los cuales proporcionan información relevante sobre las condiciones climáticas en la región de Arjonilla durante el período de estudio. Estos datos climáticos son fundamentales para comprender cómo los factores ambientales influyen en la producción de aceituna y aceite de oliva, ya que la cantidad y distribución de las precipitaciones, así como las temperaturas extremas, pueden tener un impacto significativo en el rendimiento y la calidad de la cosecha.

Además de los datos climáticos, se cuenta con información histórica desde 2001 sobre la producción diaria de aceituna en kilogramos y su rendimiento, es decir, la cantidad de aceite obtenida a partir de la cantidad de aceituna procesada. Estos datos permitirán analizar las tendencias y variaciones en la producción de la zona a lo largo del tiempo, identificando posibles factores que hayan influido en los cambios observados.

A partir de la campaña 2017/2018, se cuenta con datos adicionales, específicamente sobre la humedad contenida en la aceituna y el rendimiento graso sobre materia seca. Estos datos son de gran relevancia, ya que la humedad de la aceituna en el momento de recolección y el rendimiento graso son indicadores clave de la calidad y cantidad de aceite de oliva producido. El contenido de humedad influye en la eficiencia del proceso de extracción de aceite, mientras que el rendimiento graso sobre materia seca indica la cantidad de aceite que se puede obtener a partir de la aceituna.

Durante esta sección, nos adentraremos en un análisis exhaustivo de las variables bajo estudio. En primer lugar, llevaremos a cabo una evaluación de la normalidad de las variables. Además, realizaremos un estudio detallado de las relaciones entre las variables, utilizando técnicas como la correlación para medir la fuerza y la dirección de las asociaciones. También emplearemos gráficos y visualizaciones para visualizar estas relaciones de manera más intuitiva y comprensible.

Por último, exploraremos la relación entre la recolección temprana y la campaña oficial,

analizando cómo se relacionan estas dos variables importantes en el contexto de nuestro estudio. Examincaremos los posibles efectos de la recolección temprana en los resultados de la campaña oficial, evaluando si existen diferencias significativas en términos de rendimiento.

3.1. Descripción de los datos

En esta sección se proporcionará una descripción detallada de la búsqueda e importación de los datos utilizados en el estudio, así como su naturaleza. Para obtener los datos relacionados con la producción en la cooperativa San Roque, se llevó a cabo una reunión con el presidente actual de la cooperativa, en la cual se explicó el objetivo del estudio y se solicitó acceso a los datos necesarios. Se obtuvo un historial completo desde el año 2001 hasta la fecha actual, el cual contiene registros diarios de cada entrega realizada en la cooperativa. Estos registros incluyen información como el número de ticket, la fecha de entrega, los kilogramos entregados, el rendimiento y los kilogramos de aceite producidos.

Fecha	ticket	Kg	Rdto. Laborato	Kg. aceite real
05-dic-02	1	398	20,01	78
05-dic-02	2	1632	24,11	385,37
05-dic-02	3	1810	21,58	382,55
05-dic-02	4	619	23,15	140,35
05-dic-02	5	415	21,76	88,44
05-dic-02	6	653	18,67	119,4
05-dic-02	7	1576	18,23	281,39
05-dic-02	8	1000	18,23	178,54
05-dic-02	9	1594	19,16	299,12
05-dic-02	10	759	21,21	157,67

Figura 3.1: Datos de los registros históricos (fuente: elaboración propia)

Además, se recibieron una serie de archivos de Excel correspondientes a cada campaña desde la temporada 2017/18 hasta la temporada 2021/2022. Estos archivos contienen datos diarios de las entradas en la cooperativa, pero esta vez también se incluyen variables adicionales de gran importancia para el estudio, como la humedad del fruto y el rendimiento graso sobre materia seca. Estas variables proporcionan información clave sobre la calidad y las características de las aceitunas, lo que permite un análisis más completo y preciso.

DIA	MES	AÑO	TICKET	*RENDIMIENTO*	HUMEDAD	RGSS%
12	12	2017	1	***21,05***	51,03	42,99
12	12	2017	2	***20,48***	48,43	39,72
12	12	2017	3	***20,36***	51,74	42,2
12	12	2017	4	***20,8***	51,29	42,71
12	12	2017	5	***22,06***	50,88	44,92
12	12	2017	6	***20,38***	48,01	39,21
12	12	2017	7	***21,89***	48,3	42,34
12	12	2017	8	***22,9***	51,13	46,86
12	12	2017	9	***18,17***	55,31	40,68
12	12	2017	10	***18,04***	51,65	37,31
12	12	2017	11	***22,51***	50,2	45,2

Figura 3.2: Datos de los registros por campaña (fuente: elaboración propia)

Por otra parte, para complementar los datos obtenidos de la cooperativa San Roque, se accedió a los datos climáticos de Arjonilla a través de AEMET (Agencia Estatal de Meteorología). Estos datos incluyen información sobre las precipitaciones, así como las temperaturas máximas y mínimas registradas desde el año 2001. La incorporación de estos datos climáticos es fundamental, ya que permiten analizar la influencia de las condiciones ambientales en la producción de aceitunas.

FECHA	Precipitación: l/m^2
01/01/2001	15.2
02/01/2001	0.4
03/01/2001	5.3
04/01/2001	2.9
05/01/2001	3.5
06/01/2001	2.5
07/01/2001	0.1
08/01/2001	0.0

Figura 3.3: Datos de las precipitaciones (fuente: elaboración propia)

FECHA	T. Máxima	T.Mínima
01/01/2001	13.4	3.5
02/01/2001	16.3	7.8
03/01/2001	15.5	7.9
04/01/2001	17.3	11.3
05/01/2001	16.8	11.9
06/01/2001	16.2	8.0
07/01/2001	12.9	2.8
08/01/2001	11.7	0.1
09/01/2001	10.0	-1.1

Figura 3.4: Datos de las temperaturas (fuente: elaboración propia)

En resumen, para llevar a cabo este estudio se realizó un proceso exhaustivo de búsqueda e importación de datos. Gracias a la colaboración de la cooperativa San Roque y

la disponibilidad de su historial de producción, así como la inclusión de variables relevantes como la humedad del fruto y el rendimiento graso sobre materia seca, se cuenta con una sólida base de datos para realizar un análisis completo. La incorporación de los datos climáticos también enriquece el estudio al considerar la influencia de las condiciones ambientales en la producción de aceitunas.

3.2. Análisis de la normalidad de las variables

Durante este apartado, llevaremos a cabo un estudio exhaustivo de la normalidad de las variables que tienen influencia en nuestro estudio. El objetivo principal de este análisis es determinar si las variables siguen una distribución normal o no. Para conseguirlo, utilizaremos una serie de técnicas y herramientas estadísticas.

En primer lugar, procederemos a realizar un resumen descriptivo de cada variable. En el mismo, se incluirán diferentes medidas que nos permitirán tener una visión general de la distribución de los datos. Estas medidas incluirán el valor mínimo y máximo observado, la media aritmética, la mediana y los cuartiles. Estas estadísticas descriptivas nos proporcionarán una idea de la tendencia central, la dispersión y la forma de la distribución de cada variable.

Además del análisis descriptivo, utilizaremos herramientas gráficas para visualizar la distribución de las variables. Para ello, emplearemos un conjunto de gráficos, incluyendo diagramas de caja y bigotes (boxplots), histogramas y gráficos Q-Q.

El boxplot nos permitirá identificar valores atípicos y evaluar la simetría de la distribución. El histograma, por su parte, nos proporcionará una representación visual de la forma y la frecuencia de los valores en cada variable. Finalmente, el gráfico Q-Q nos ayudará a comparar la distribución de los datos con una distribución normal teórica, mediante la comparación de los cuantiles empíricos y los cuantiles teóricos.

Como último paso de nuestro análisis de normalidad, aplicaremos el test de normalidad de Shapiro-Wilk. Como se ha comentado anteriormente, la hipótesis nula establece que la variable analizada sigue una distribución normal en la población. Este test es una prueba estadística ampliamente utilizada para evaluar la normalidad de una variable. Se basa en la comparación de los valores observados con los valores esperados bajo una distribución normal. El resultado del test nos proporcionará un p-valor, que nos indicará si los datos siguen una distribución normal o no. En nuestro caso, trabajaremos con un nivel de significación del 5 %.

3.2.1. Rendimiento

Nos centraremos en una de las variables más relevantes de nuestro estudio: el rendimiento. El rendimiento es una medida fundamental que nos permite evaluar la eficacia y la productividad en el contexto de nuestra investigación. En nuestro caso, esta medida medirá el rendimiento medio diario en cada una de las campañas estudiadas.

En nuestro análisis, el rendimiento asume un papel central, ya que será el objetivo principal de nuestros modelos predictivos como se detallará en el capítulo 4.

El resumen descriptivo de la variable rendimiento muestra los siguientes valores:

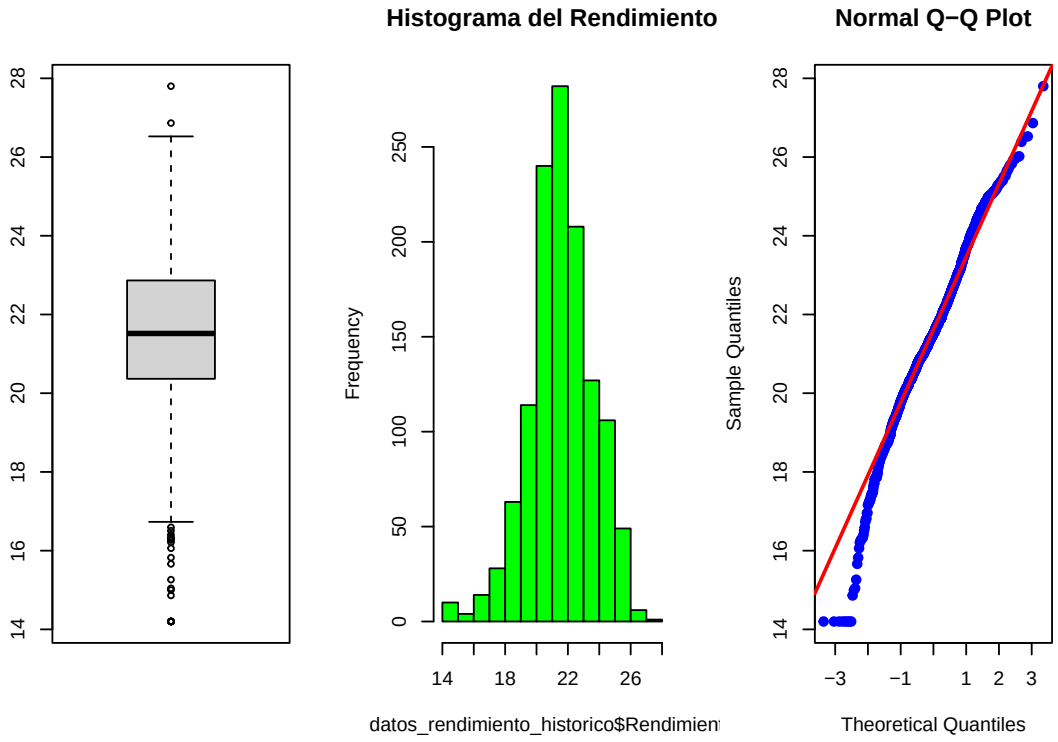
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.20   20.37   21.52   21.56   22.86   27.80
```

Estos valores proporcionan información sobre la tendencia central, la dispersión y la forma de la distribución de la variable rendimiento. La media aritmética indica que el valor promedio de rendimiento se encuentra alrededor de 21.56. La mediana de 21.52, que es muy similar a la media, sugiere que la distribución de la variable es aproximadamente simétrica.

Los cuantiles nos brindan información sobre la dispersión de los datos. El primer cuantil (20.37) y el tercer cuartil (22.86) indican que el 25 % de los valores de la variable rendimiento se encuentran por debajo de 20.37 y el 25 % de los valores se encuentra por encima de 22.36, respectivamente.

El valor mínimo de 14.20 y el valor máximo de 27.80 nos dan una idea de la amplitud de los datos observados, lo cual sugiere que hay una variabilidad considerable en la variable rendimiento.

En conjunto, estos resultados descriptivos proporcionan una visión general de la distribución de la variable rendimiento, indicando que la mayoría de los valores se encuentran alrededor de la media y la mediana, con una dispersión relativamente amplia. Sin embargo, estos resultados no son suficientes para determinar si la variable sigue una distribución normal. Para ello realizaremos un análisis gráfico y aplicaremos el test de normalidad de Shapiro-Wilk.




```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  datos_rendimiento_historico$Rendimiento
## W = 0.98553, p-value = 7.81e-10
```

En primer lugar, al observar el boxplot, se puede apreciar que la distribución de la variable rendimiento es simétrica. La caja central del boxplot, que representa el rango intercuartil (el 50 % central de los datos), muestra una simetría en su posición, indicando que la mediana se encuentra en el centro de la distribución. Además, los bigotes del boxplot no muestran una asimetría pronunciada, lo que refuerza la suposición de simetría en la distribución.

Sin embargo, también se identifican algunos valores atípicos (outliers) en la parte inferior del boxplot, correspondientes a rendimientos inferiores. Estos valores atípicos pueden deberse a la influencia de la campaña de recogida de aceite temprano, donde se espera que el rendimiento sea inferior en comparación con la campaña oficial de recogida. Es importante tener en cuenta estos valores atípicos al interpretar los resultados.

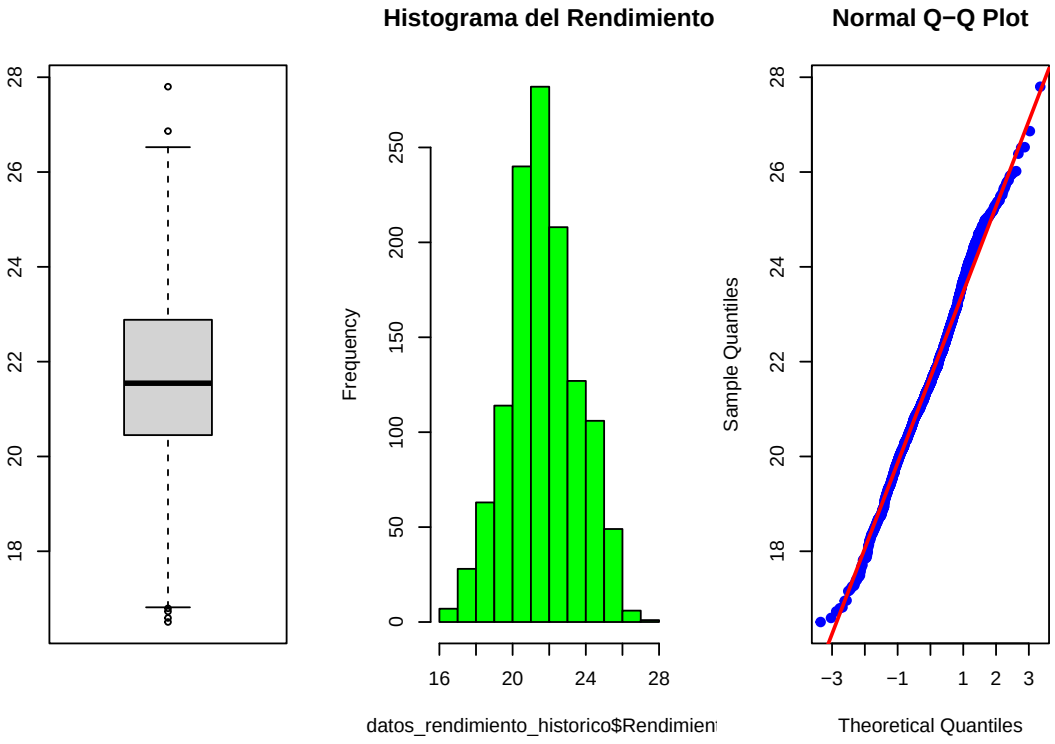
En cuanto al histograma, se puede observar que la forma de la distribución se asemeja a la distribución normal. La distribución de los valores se concentra alrededor de la media y la mediana, y la forma del histograma muestra una simetría característica de la distribución normal. Sin embargo, es importante tener en cuenta que el histograma proporciona una representación visual, y se necesitará un análisis adicional para determinar si la distribución sigue una forma normal de manera precisa.

El gráfico Q-Q (cuantil-cuantil) también brinda información valiosa sobre la normalidad de la distribución de rendimiento. Al examinar los puntos de datos en el gráfico, se observa que la mayoría de ellos se ajustan a la línea recta de manera general, lo que indica una distribución normal. Sin embargo, se aprecia que algunos puntos en las colas de la distribución no siguen exactamente la línea recta esperada. Esto podría sugerir cierta desviación de la normalidad en los extremos de la distribución, pero en su mayoría, los puntos parecen estar razonablemente distribuidos de acuerdo con una distribución normal.

Además de los análisis gráficos, aplicamos el test de normalidad de Shapiro-Wilk a la variable rendimiento. Los resultados del test indican un valor de p igual a 7.81e-10, lo que significa que el p-valor obtenido es extremadamente bajo.

Dado que el nivel de significación utilizado es 0.05, y el valor de p obtenido (7.81e-10) es mucho menor que este nivel de significación, se rechaza la hipótesis nula de normalidad. Por lo tanto, se concluye que la variable rendimiento no sigue una distribución normal.

Este resultado refuerza la observación anterior de que existen algunos indicios de desviación de la normalidad en los análisis gráficos. Aunque la forma general de la distribución parece ser similar a una distribución normal, los valores atípicos y las desviaciones en los extremos, junto con el resultado del test de Shapiro-Wilk, indican que la variable no se distribuye de manera completamente normal.




```
##
##  Shapiro-Wilk normality test
##
## data:  datos_rendimiento_historico$Rendimiento
## W = 0.99538, p-value = 0.0008356
```

Tras realizar un análisis adicional del rendimiento, excluyendo los valores atípicos asociados con la recogida temprana, hemos obtenido que los resultados del test revelan un p-valor igual a 0.0008356.

Por lo tanto, concluimos que, incluso después de excluir los rendimientos bajos asociados con la recogida temprana, la variable del rendimiento no sigue una distribución normal.

Este resultado refuerza aún más la conclusión anterior de que la variable rendimiento no se distribuye de manera normal en general. Aunque hemos eliminado los rendimientos más bajos para minimizar su influencia en el análisis, aún se observa una falta de normalidad en la distribución de los datos.

En conclusión, es de suma importancia complementar el análisis gráfico con pruebas estadísticas que permitan cuantificar y respaldar las observaciones realizadas. Si bien el análisis gráfico proporciona una primera impresión visual de los datos, los contrastes estadísticos nos permiten poner a prueba nuestras percepciones iniciales y evaluar la significancia de las diferencias observadas.

Sin embargo, también es crucial destacar la relevancia del análisis gráfico en sí mismo. Aunque las pruebas estadísticas son herramientas poderosas, no son infalibles y pueden presentar limitaciones en determinadas situaciones. El análisis gráfico proporciona una

representación visual directa de los datos, lo que facilita la identificación de patrones, tendencias y relaciones que podrían pasar desapercibidos en los resultados de las pruebas estadísticas.

Además, es importante tener en cuenta la dificultad de estimar una variable si no sigue una distribución normal. En muchos casos, las suposiciones de normalidad subyacentes en los métodos estadísticos pueden no ser cumplidas por los datos reales. Esto puede tener un impacto significativo en la interpretación de los resultados y en la precisión de las estimaciones. Por lo tanto, es fundamental tener en cuenta las características y la distribución de los datos al seleccionar y aplicar métodos estadísticos adecuados.

3.2.2. Humedad

La humedad es una variable de gran importancia en el estudio, ya que desempeña un papel fundamental en el desarrollo y la calidad de las aceitunas. La cantidad de humedad presente en el fruto puede influir en diversos aspectos, como el peso de la aceituna, la calidad del aceite producido y la resistencia a enfermedades. Por lo tanto, es crucial comprender y analizar cómo la humedad afecta a la producción de aceitunas y cómo puede ser utilizada como una variable predictiva en los modelos desarrollados en el estudio.

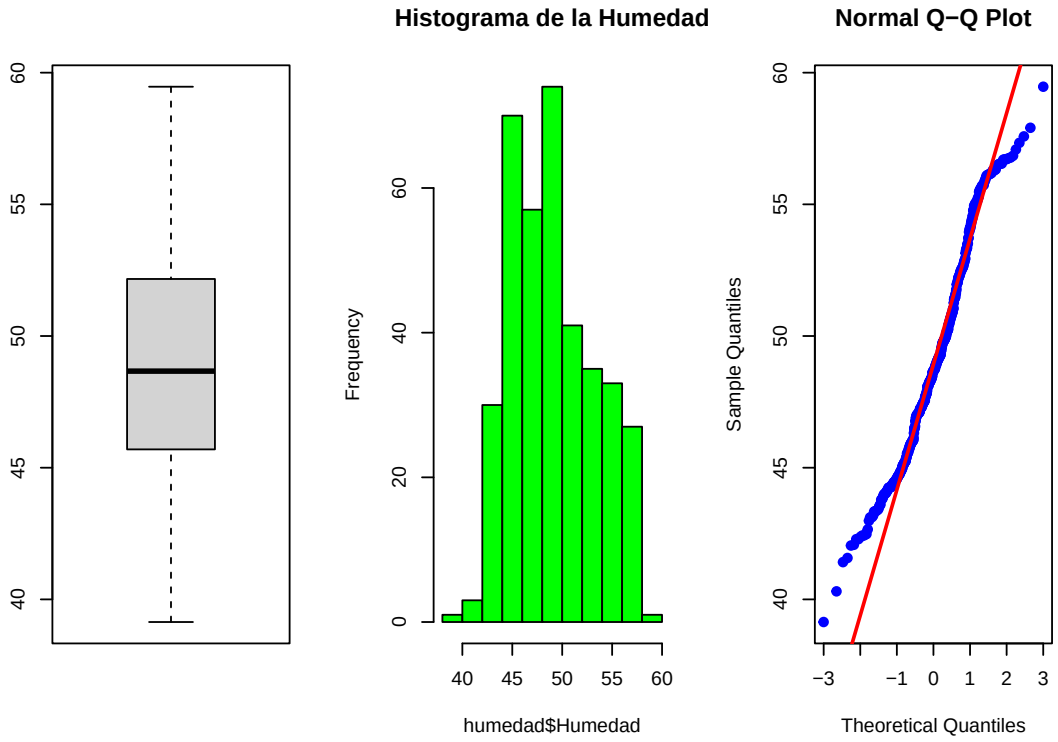
El resumen descriptivo de la variable humedad muestra los siguientes valores:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.14   45.71   48.67   49.05   52.15   59.47
```

En primer lugar, el valor mínimo de 39.14 y el valor máximo de 59.47 nos indican que la variable humedad abarca un rango amplio de valores. Esto sugiere que existe una variabilidad significativa en los niveles de humedad observados.

La media de 49.05 nos proporciona el valor promedio de la variable. Este valor nos indica que, en general, los niveles de humedad se sitúan alrededor de 49.05. Por otro lado, la mediana de 48.67, que es muy similar a la media, nos sugiere que la distribución de la variable presenta una tendencia hacia la simetría.

Al analizar los cuartiles, observamos que el primer cuartil (45.71) y el tercer cuartil (52.15) nos ofrecen información sobre la distribución de los datos en relación a la mediana. Estos cuartiles nos indican que el 25 % de los valores de humedad se encuentra por debajo de 45.71, mientras que el 25 % se encuentra por encima de 52.15.




```
##
##  Shapiro-Wilk normality test
##
## data:  humedad$Humedad
## W = 0.97171, p-value = 1.216e-06
```

Al examinar el boxplot de la variable, podemos observar que muestra una distribución simétrica. Esto se refleja en la posición de la caja central, indicando que la mediana se encuentra en el centro de la distribución. Además, los bigotes del boxplot no muestran una asimetría pronunciada.

En cuanto al histograma, al analizar su forma, notamos que no presenta una gran similitud con una distribución normal. Podemos observar que muestra diferentes patrones de distribución, en este caso, asimetrías.

Al observar el gráfico Q-Q (cuantil-cuantil) para evaluar la normalidad de la variable, notamos que los puntos de datos no siguen exactamente una línea recta a lo largo de todo el gráfico. En particular, encontramos valores que se alejan de la línea en las colas de la distribución. Esto indica que los datos en las colas están menos alineados con la distribución normal esperada. Sin embargo, los puntos centrales se aproximan a una distribución normal, lo que sugiere cierta aproximación a la normalidad en la distribución central de la variable.

Al realizar el test de Shapiro-Wilk para evaluar la normalidad de la variable, hemos obtenido un p-valor igual a 1.216e-06.

Dado que el nivel de significación utilizado es 0.05, y el p-valor obtenido (1.216e-06) es significativamente menor que este nivel de significación, rechazamos la hipótesis nula

de normalidad. Por lo tanto, podemos concluir que la variable no sigue una distribución normal.

3.2.3. Rendimiento Graso Sobre materia Seca (RGSS)

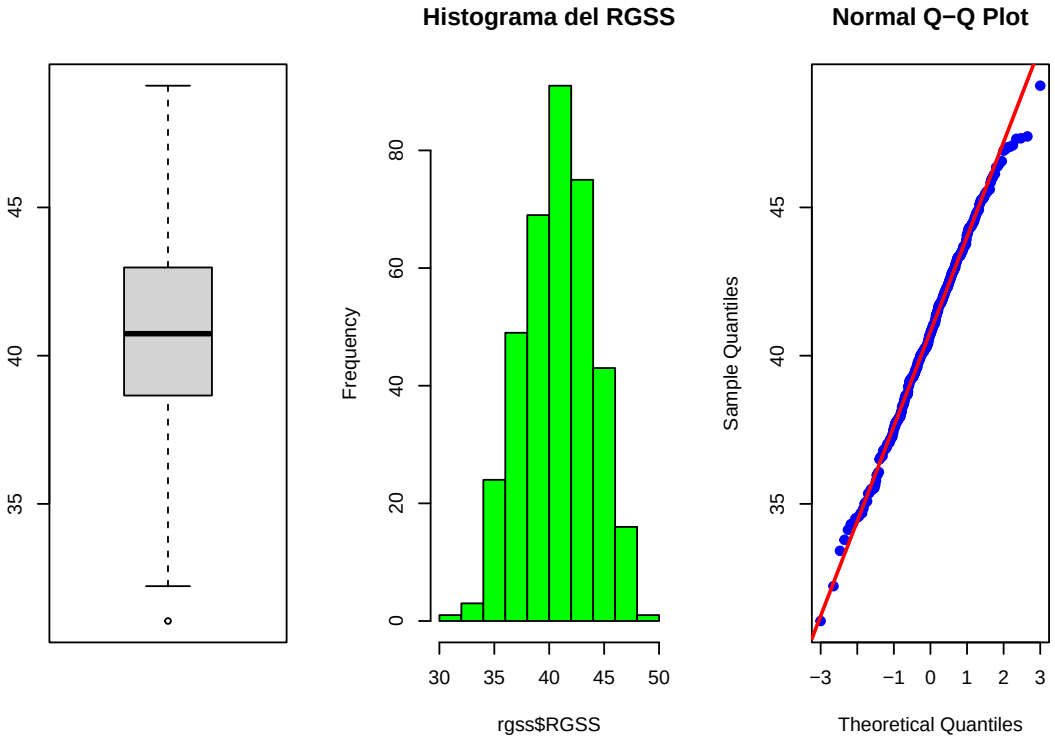
El rendimiento graso sobre materia seca es otra variable de gran relevancia en el estudio. Esta medida nos proporciona información sobre la cantidad de aceite de oliva que se puede obtener a partir de una determinada cantidad de materia seca de las aceitunas. Es un indicador clave de la calidad y el valor económico del aceite producido.

El resumen descriptivo de la variable rendimiento graso sobre materia seca proporciona información detallada y técnica sobre su distribución. Analicemos los valores presentados:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.05   38.66   40.74   40.76   42.97   49.11
```

El valor promedio del rendimiento graso sobre materia seca se encuentra alrededor de 40.76. La mediana de 40.74, que es cercana a la media, sugiere una distribución aproximadamente simétrica de los datos. El primer cuartil (38.66) y el tercer cuartil (42.97) indican que el 25 % de los valores de rendimiento graso sobre materia seca se encuentra por debajo de 38.66 y el 25 % se encuentra por encima de 42.97, respectivamente.

El valor mínimo de 31.05 y el valor máximo de 49.11 nos indican el rango de variación de los datos observados. Esta información muestra la amplitud de la distribución de los rendimientos grasos sobre materia seca.



```
##
##  Shapiro-Wilk normality test
##
## data:  rgss$RGSS
## W = 0.99612, p-value = 0.4984
```

Al analizar los gráficos y el resultado del test de Shapiro-Wilk para la variable rendimiento graso sobre materia seca, podemos obtener las observaciones que se describen a continuación.

El boxplot muestra una simetría en la distribución de los datos. La caja central, que representa el rango intercuartil, indica que la mediana se encuentra en el centro de la distribución. Los bigotes del boxplot también indican una simetría en la distribución de los datos, ya que no muestran una asimetría pronunciada hacia un lado u otro.

El histograma, al observar su forma, muestra una distribución que se asemeja a una distribución normal. La forma del histograma se asemeja a una campana característica de una distribución normal, lo que sugiere que los valores de rendimiento graso sobre materia seca se distribuyen de manera similar a una distribución normal.

En cuanto al gráfico Q-Q, la mayoría de los valores se encuentran cerca de la línea recta, lo que indica una buena aproximación a una distribución normal. Además, el resultado del test de Shapiro-Wilk nos proporciona un p-valor de 0.4984. Dado que este valor de p es mayor que el nivel de significación utilizado de 0.05, no tenemos suficiente evidencia para rechazar la hipótesis nula de normalidad. Esto sugiere que la variable rendimiento graso sobre materia seca se aproxima a una distribución normal.

3.2.4. Kilogramos de Aceite

Los kilogramos de aceite son una variable fundamental en el estudio, ya que representan la cantidad de aceite de oliva producido en cada campaña. Esta medida es de gran importancia tanto desde el punto de vista económico. Los kilogramos de aceite reflejan la eficiencia del proceso de extracción y la calidad de las aceitunas utilizadas.

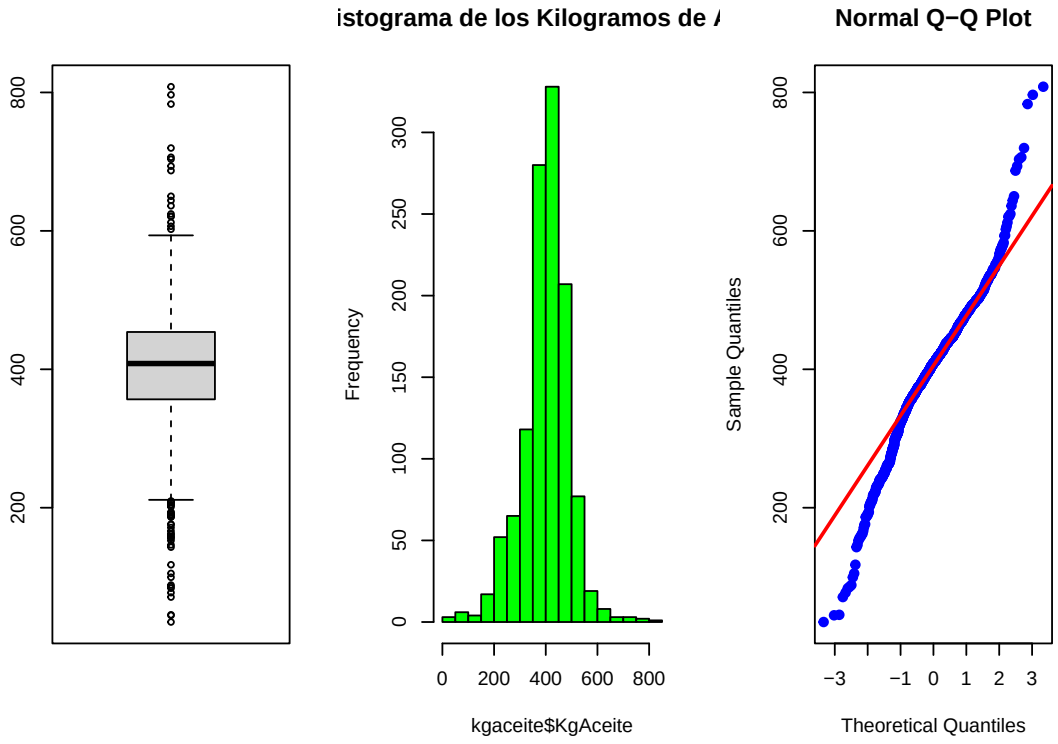
El resumen descriptivo de la variable "kilogramos de aceite" proporciona información detallada y técnica sobre su distribución. Vamos a analizar los valores presentados:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.86  356.59  408.32  400.53  453.92  808.28
```

La media de 400.53 indica que el valor promedio de los kilogramos de aceite es aproximadamente 400.53. La mediana de 408.32, que está cerca de la media, sugiere que la distribución de los datos está relativamente equilibrada. En otras palabras, hay una cantidad similar de valores por encima y por debajo de este punto medio.

Al observar los cuartiles, notamos que el primer cuartil (356.59) y el tercer cuartil (453.92) nos brindan información sobre la dispersión de los datos.

El rango, que es la diferencia entre el valor mínimo (34.86) y el valor máximo (808.28), nos muestra la amplitud total de la distribución de los kilogramos de aceite. Esto indica que hay una amplia variación en los rendimientos de aceite, desde valores muy bajos hasta valores muy altos.




```
##
##  Shapiro-Wilk normality test
##
## data:  kgaceite$KgAceite
## W = 0.96653, p-value = 5.847e-16
```

Al analizar el boxplot de la variable y los gráficos de histograma y Q-Q, podemos hacer las siguientes observaciones:

El boxplot muestra la presencia de numerosos valores atípicos tanto por encima como por debajo de los límites del rango intercuartil. Estos valores atípicos se deben a las irregularidades que ocurren en cada campaña de recolección. Factores como los días de lluvia, que impiden la recolección de frutos, y los registros al final de la campaña, donde la cantidad de frutos y, por ende, la cantidad de aceite recolectado es menor, contribuyen a la presencia de estos valores atípicos. Estas variaciones en las condiciones de recolección impactan en los resultados obtenidos.

En cuanto al histograma, se observa que la distribución de los kilogramos de aceite es más apuntada (tienen una mayor concentración de valores alrededor de la media) en comparación con una distribución normal estándar. Esto se refleja en una curtosis positiva, lo que indica que los datos tienen colas más pesadas y una mayor concentración en el centro de la distribución. Aunque no sigue una distribución perfectamente simétrica, la forma general del histograma sugiere que los kilogramos de aceite tienden a agruparse en un rango relativamente estrecho alrededor de la media. Sin embargo, es importante tener en cuenta que la presencia de valores atípicos puede afectar la forma y apariencia de la distribución.

El gráfico Q-Q muestra que hay muchos valores que se desvían de la línea recta, especialmente en las colas de la distribución. Esto indica que hay desviaciones significativas de la normalidad en los extremos de la distribución de los kilogramos de aceite.

El p-valor obtenido del test de Shapiro-Wilk es de 5.847e-16. Dado que este valor es extremadamente pequeño, rechazamos la hipótesis nula de normalidad. Esto proporciona evidencia sólida de que la distribución de los kilogramos de aceite no sigue una distribución normal.

3.3. Estudio comparativo de las variables en estudio

En esta sección, se abordará el estudio y la comparación del rendimiento, la humedad, el rendimiento graso sobre materia seca y los kilogramos de aceite producidos por campaña en el contexto del cultivo del olivar en la cooperativa San Roque.

El análisis de estas variables resulta fundamental para comprender y evaluar la eficiencia y la calidad de la producción de aceite de oliva en cada campaña. El rendimiento, expresado como la cantidad de aceite obtenido a partir de la cantidad de aceituna procesada, es un indicador clave para medir la productividad y la rentabilidad del cultivo del olivar. Además, la humedad contenida en la aceituna y el rendimiento graso sobre materia seca son factores determinantes para la calidad y el perfil organoléptico del aceite producido.

El estudio y comparación de estas variables a lo largo de diferentes campañas permitirán identificar patrones y tendencias en la producción de aceite de oliva, así como analizar posibles factores que puedan haber influido en los resultados observados. Además, se podrán realizar comparaciones entre campañas para evaluar el impacto de diferentes condiciones climáticas, prácticas agrícolas u otros factores que puedan influir en los resultados obtenidos.

El análisis de los kilos de aceite producidos durante la campaña proporcionará una visión general de la evolución de la producción a lo largo del tiempo, permitiendo así identificar años de mayor o menor producción y evaluar el impacto de los diferentes factores en los resultados obtenidos.

Para llevar a cabo este estudio y comparación, se utilizarán los datos recopilados desde el inicio de las campañas disponibles, analizando la información de rendimiento, humedad, rendimiento graso sobre materia seca y kilos de aceite producidos en cada una de ellas. Estos datos representan todas las entradas diarias de aceituna procesadas en la cooperativa a lo largo de las campañas analizadas con información precisa y detallada sobre la cantidad de aceituna, así como los parámetros asociados.

Para analizar las variables mencionadas, se utilizará la media diaria de las entradas en la cooperativa. La utilización de esta medida promedio permitirá obtener una visión general y representativa de la producción de aceite de oliva en cada campaña, al considerar la contribución de todas las entradas diarias de aceituna durante ese período. La media diaria será calculada a partir de los datos, considerando el peso de la aceituna recibida, los valores de humedad, el rendimiento graso sobre materia seca y el rendimiento obtenidos en cada entrada.

Es importante destacar que los datos utilizados en este estudio se centrarán exclusivamente en la campaña oficial de recogida de aceituna, dejando de lado los datos correspondientes a la campaña de recogida de aceituna destinada a la obtención de aceite temprano.

La decisión de obviar los datos de la campaña de recogida de aceituna destinada a la obtención de aceite temprano se basa en dos razones fundamentales. En primer lugar, se ha observado que esta campaña cuenta con un número muy reducido de entradas en comparación con la campaña normal. Esto implica que la disponibilidad de datos para el análisis sería limitada, lo que podría afectar a la robustez y representatividad de los resultados obtenidos.

En segundo lugar, se ha observado que los valores de los parámetros estudiados en la campaña de aceite temprano presentan ligeras diferencias en comparación con la campaña oficial. Al centrarnos únicamente en la campaña normal, se garantizará una comparación más precisa y consistente de los resultados, evitando posibles distorsiones ocasionadas por las particularidades de la campaña de aceite temprano.

Además de los datos recopilados sobre rendimiento, humedad, rendimiento graso sobre materia seca y kilos de aceite producidos en cada campaña, este estudio también utilizará información complementaria relacionada con las condiciones climáticas. Se incluirán los datos de temperatura máxima, temperatura mínima y precipitaciones correspondientes al año en estudio.

Es importante destacar que algunos de los datos climáticos necesarios para el análisis presentaban valores faltantes. Para abordar esta limitación, se realizó una imputación de datos utilizando la media mensual de los registros disponibles desde el año 2001 hasta el año 2022. De esta manera, se pudo estimar de manera precisa y fiable los valores ausentes, permitiendo una evaluación más completa y precisa de las condiciones ambientales a lo largo de la campaña.

La inclusión de estos datos climáticos en el análisis permitirá examinar la posible influencia de las condiciones meteorológicas en el rendimiento de la aceituna. Esto brindará una perspectiva más completa y enriquecedora para comprender los factores que afectan el proceso productivo y los resultados obtenidos en la cooperativa.

3.3.1. Análisis de correlaciones y contexto climático

En esta sección abordaremos un análisis exahustivo de las correlaciones entre las variables clave de nuestro estudio, que son: rendimiento, rendimiento graso sobre materia seca, humedad del fruto y kilogramos de aceituna recogidos. Nuestro objetivo es investigar las relaciones estadísticas significativas entre estas variables en el contexto de las campañas agrícolas comprendidas entre 2017/2018 y 2021/2022.

Para llevar a cabo este estudio, los datos están organizados por número de ticket, lo que significa que cada entrada en la cooperativa está asociada a un ticket que contiene los datos de las variables mencionadas anteriormente. Para abordar este estudio, calcularemos la media de cada variable por campaña.

Para enriquecer más nuestro estudio, consideraremos el contexto climático de cada campaña agrícola. Recopilaremos datos relacionados con las precipitaciones, temperatura

máxima y temperatura mínima registradas durante los meses correspondientes al año de la campaña. Para ello, utilizaremos la media en cada mes de estas tres variables.

Mediante el análisis y comparación de las correlaciones obtenidas entre las variables estudiadas y los datos climáticos asociados, podremos identificar patrones significativos y establecer relaciones causales potenciales. Estos resultados nos ayudarán a comprender mejor los mecanismos subyacentes que influyen en las variables, así como a tomar decisiones fundamentadas en ámbito agrícola.

3.3.1.1. Análisis de la influencia de la climatología en el rendimiento

En este apartado, se llevará a cabo un análisis detallado de la influencia de la climatología en el rendimiento de la producción de aceite de oliva. Se examinarán las relaciones y patrones existentes entre estas variables climáticas y el rendimiento de la producción, con el objetivo de comprender cómo los factores climáticos pueden influir en los resultados obtenidos.

3.3.1.1.1. Influencia de las temperaturas máximas

En esta sección se abordará el estudio de la influencia de las temperaturas máximas en el rendimiento de la aceituna. Al examinar el gráfico de correlaciones 3.1, se observa claramente que las temperaturas máximas en el mes de julio tienen una influencia significativa en el rendimiento de la aceituna. Es importante destacar que esta influencia es predominantemente negativa, lo cual era esperado debido al conocido efecto adverso de las altas temperaturas en la producción de aceite.

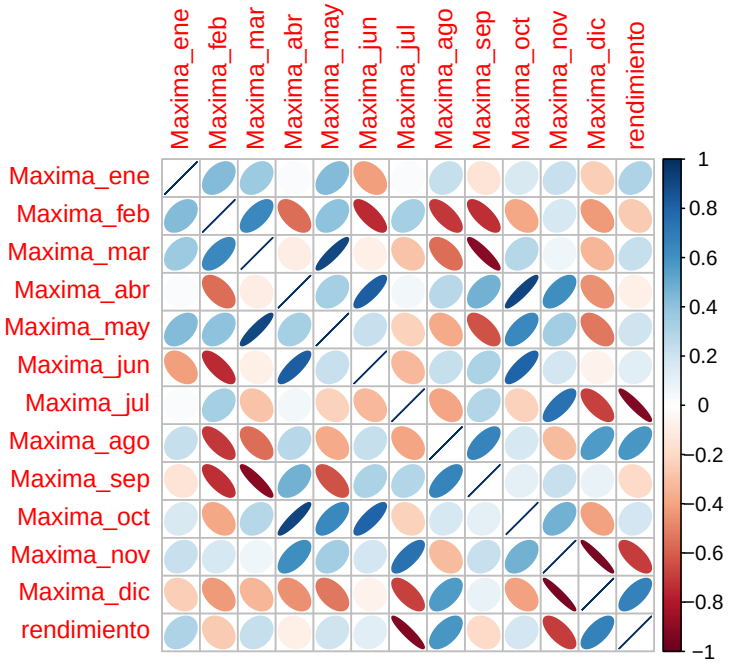


Figura 3.5: Influencia de las máximas en el rendimiento (fuente: elaboración propia)

El estrés térmico puede afectar negativamente en la acumulación de aceite en los frutos. Las altas temperaturas pueden interferir con la fotosíntesis y la capacidad de la planta

para sintetizar y almacenar aceite en los tejidos de la aceituna. Como resultado, se produce una disminución en la cantidad total de aceite obtenido.

Es importante destacar que estos hallazgos son consistentes con la literatura científica existente y respaldan la comprensión de los efectos negativos de las altas temperaturas en la producción de aceite de oliva. Estos resultados respaldan la importancia de implementar estrategias de manejo agrícola adecuadas para mitigar los efectos negativos de las altas temperaturas y preservar la calidad y cantidad de la producción de aceite de oliva.

3.3.1.1.2. Influencia de las temperaturas mínimas

En esta sección nos enfocaremos en estudiar la influencia de las temperaturas mínimas en el rendimiento de la aceituna.

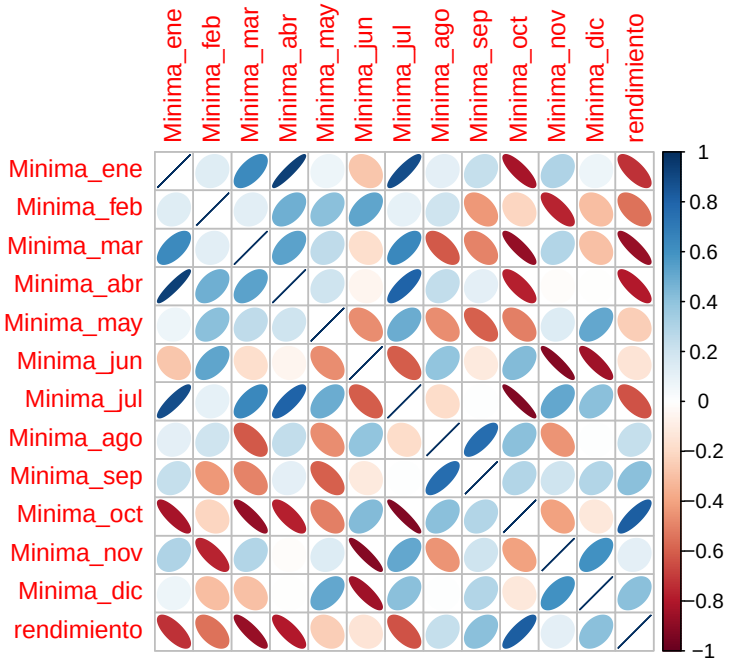


Figura 3.6: Influencia de las mínimas en el rendimiento (fuente: elaboración propia)

Al examinar el gráfico de correlaciones, se destacan algunos hallazgos significativos. En primer lugar, se observa una influencia negativa de las temperaturas mínimas durante el periodo de aparición de las yemas y floración, que abarca los meses de marzo y abril. Estos resultados indican que temperaturas mínimas más bajas en esta etapa pueden tener un efecto positivo en la polinización y la formación de los frutos.

Además, se identifica una influencia positiva de las temperaturas mínimas en el mes de julio. Este hallazgo sugiere que temperaturas mínimas más bajas en este mes pueden tener un impacto favorable en el rendimiento de la aceituna. Es posible que temperaturas mínimas más frescas en julio favorezcan el desarrollo de los frutos y la acumulación de aceite.

Por otro lado, el análisis revela una influencia positiva de las temperaturas mínimas en el mes de octubre. Este hallazgo sugiere que temperaturas mínimas más altas en este momento pueden tener un efecto beneficioso en el rendimiento de la aceituna. Con

temperaturas mínimas adecuadas en octubre, es probable que se promueva la maduración de los frutos y una mayor acumulación de aceite.

Estos hallazgos resaltan la importancia de considerar las temperaturas mínimas en distintas etapas del ciclo de cultivo de la aceituna. Es necesario mantener un equilibrio adecuado, evitando temperaturas mínimas extremas que puedan afectar negativamente la producción y la calidad del aceite de oliva.

3.3.1.1.3. Influencia de las precipitaciones

En esta sección nos centraremos en investigar la influencia de las precipitaciones en el rendimiento de la aceituna.

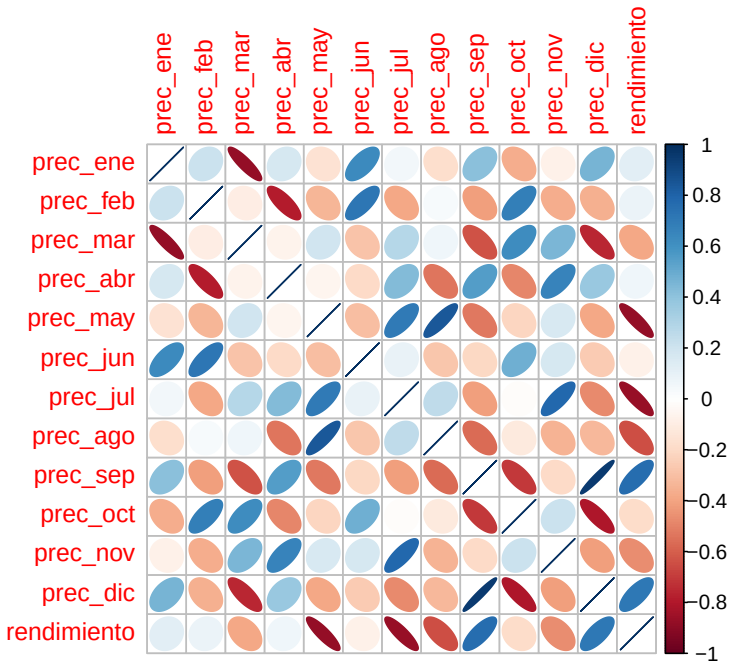


Figura 3.7: Influencia de las precipitaciones en el rendimiento (fuente: elaboración propia)

Al examinar el gráfico de correlaciones 3.3, se destacan hallazgos significativos en relación a las precipitaciones. Un punto destacable es la influencia muy positiva de las precipitaciones en el mes de septiembre. Este hallazgo sugiere que las precipitaciones en septiembre tienen un impacto significativo en el rendimiento de la aceituna. El suministro de agua en este mes crucial puede estimular el crecimiento de los frutos y la acumulación de aceite, lo que resulta en un mayor rendimiento y una mejor calidad del aceite de oliva.

Estos hallazgos resaltan la importancia de las precipitaciones adecuadas en diferentes momentos del ciclo de vida de la aceituna. La disponibilidad de agua en los momentos críticos del crecimiento y maduración de los frutos puede tener un impacto significativo en el rendimiento y la calidad del aceite de oliva producido.

3.3.1.2. Análisis de la influencia de la climatología en el rendimiento graso sobre materia seca

En este apartado, se llevará a cabo un análisis detallado de la influencia de la climatología en el rendimiento graso sobre materia seca de las aceitunas. Se examinará cómo variables climáticas como las temperaturas máximas y mínimas, así como las precipitaciones, pueden afectar el contenido de grasa en las aceitunas y, por ende, en el aceite producido.

3.3.1.2.1. Influencia de las temperaturas máximas

En esta sección nos enfocaremos en analizar la influencia de las temperaturas máximas en el rendimiento graso sobre materia seca de la aceituna.

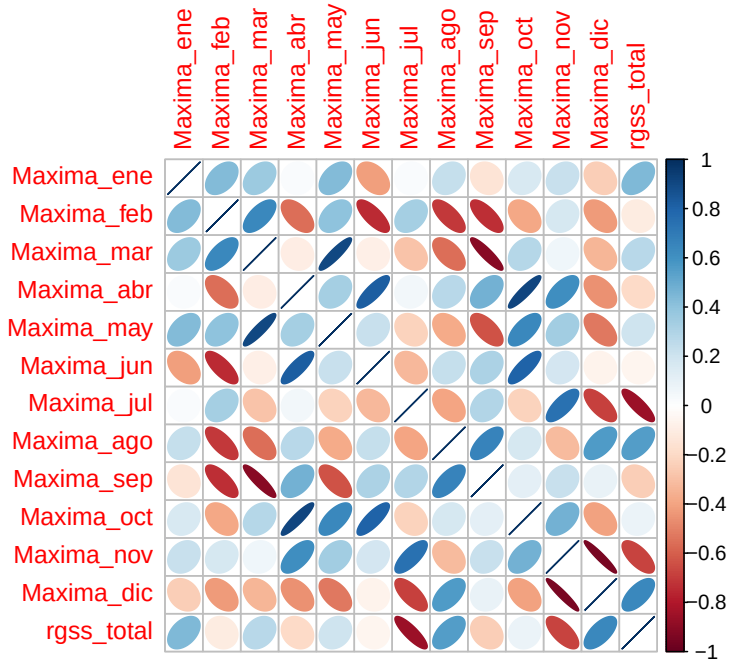


Figura 3.8: Influencia de las máximas en el RGSS (fuente: elaboración propia)

Al examinar detenidamente el gráfico de correlaciones 3.4, se destaca un hallazgo significativo. Es evidente que las temperaturas máximas en el mes de julio tienen un impacto negativo muy significativo en el rendimiento graso sobre materia seca.

El efeto negativo de las temperaturas máximas en el rendimiento graso sobre materia seca puede atribuirse a varios factores. En primer lugar, las altas temperaturas pueden acelerar el proceso de maduración de la aceituna, lo que conduce a una menor acumulación de aceite y, por tanto, a un menor rendimiento graso. Además, las altas temperaturas pueden afectar negativamente en la composición del aceite, reduciendo su contenido de ácidos grasos saludables y antioxidantes.

Además del mes de julio, se observa que las temperaturas máximas en el mes de noviembre también tienen un efecto negativo en el rendimiento graso sobre materia seca. Esto puede ser atribuido a que el mes de noviembre se encuentra cerca de la etapa de recolección de la aceituna, y las altas temperaturas en este momento pueden afectar la calidad de los frutos y la composición del aceite.

3.3.1.2.2. Influencia de las temperaturas mínimas

En esta sección de nuestro estudio, nos centraremos en investigar la influencia de las temperaturas mínimas en el rendimiento graso sobre materia seca de la aceituna.

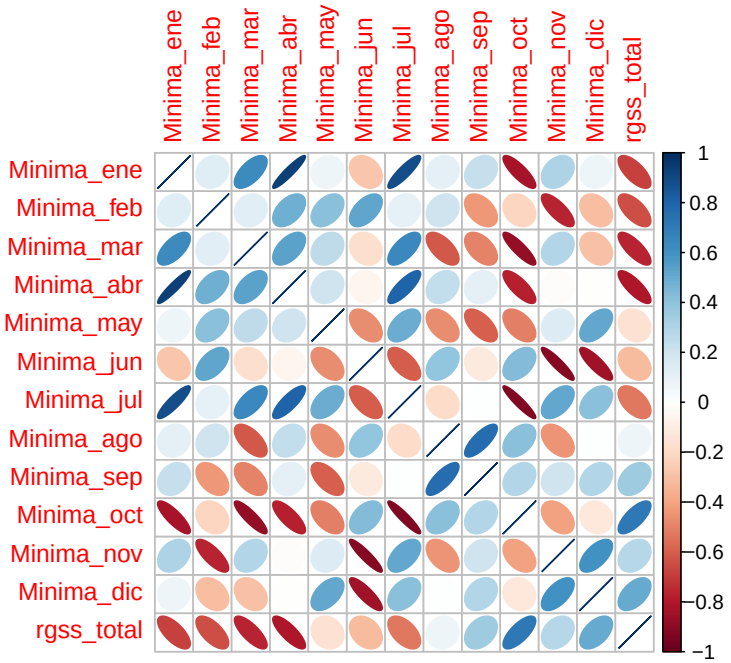


Figura 3.9: Influencia de las mínimas en el RGSS (fuente: elaboración propia)

Al examinar detalladamente el gráfico de correlaciones 3.5 se destacan resultados significativos en relación a las temperaturas mínimas. Es evidente que las temperaturas mínimas tienen un impacto negativo en el rendimiento graso sobre materia seca durante los meses de creación de yemas y floración, específicamente en marzo y abril. Estos hallazgos concuerdan con estudios previos que han demostrado que las bajas temperaturas en estas etapas críticas pueden afectar positivamente el desarrollo y la calidad de los frutos.

La influencia negativa de las temperaturas mínimas en los meses de creación de yemas y floración puede deberse a que las bajas temperaturas pueden promover un proceso de floración más adecuado y favorecer la formación adecuada de los frutos. Esto a su vez puede tener un impacto directo en el rendimiento graso sobre materia seca de la aceituna, contribuyendo a obtener un mayor contenido de grasa en el aceite producido.

Por otro lado, se observa un efecto positivo de las temperaturas mínimas en los meses cercanos a la recogida del fruto, como octubre y noviembre. Estos resultados indican que las bajas temperaturas en estas etapas finales del ciclo de crecimiento pueden tener un efecto beneficioso en la calidad del aceite de oliva producido. Las temperaturas mínimas adecuadas en estos meses pueden contribuir a una maduración gradual y equilibrada de los frutos, favoreciendo así un mayor rendimiento graso sobre materia seca.

3.3.1.2.3. Influencia de las precipitaciones

En este apartado de nuestro estudio, nos enfocaremos en investigar la influencia de las precipitaciones en el rendimiento graso sobre materia seca de la aceituna.

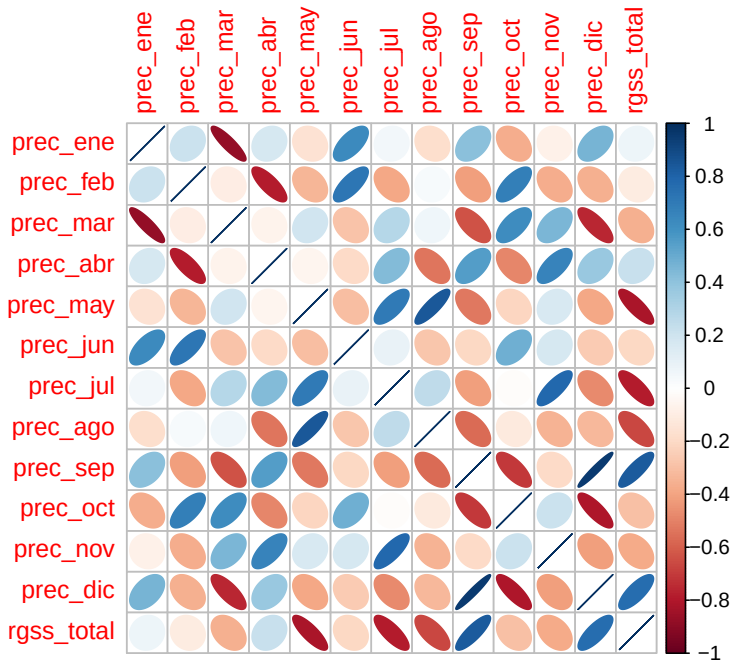


Figura 3.10: Influencia de las precipitaciones en el RGSS (fuente: elaboración propia)

Al examinar detalladamente el gráfico, se observa un patrón interesante en relación a las precipitaciones. Las precipitaciones tienen un impacto negativo en el rendimiento graso sobre materia seca durante el mes de mayo y durante los meses de verano.

Las precipitaciones en mayo pueden tener un efecto negativo en el rendimiento graso sobre materia seca debido a que un exceso de agua en ese momento puede provocar una menor acumulación de aceite en los frutos. Además, las altas temperaturas asociadas con los meses de verano pueden aumentar la evaporación y la pérdida de agua de los frutos, lo que también afecta negativamente al rendimiento graso sobre materia seca.

Por otro lado, es notable el impacto positivo de las precipitaciones en el mes de septiembre. Este hallazgo indica que las precipitaciones en septiembre pueden favorecer el rendimiento graso sobre materia seca de la aceituna. Esto puede estar relacionado con el hecho de que el mes de septiembre es crítico para la etapa de acumulación de aceite en los frutos, y un suministro adecuado de agua en ese momento puede contribuir a una mayor acumulación de aceite.

3.3.1.3. Análisis de la influencia de la climatología en la humedad del fruto

En este apartado, se analizará la influencia de la climatología en la humedad del fruto de la aceituna. La humedad del fruto es un factor crucial que puede afectar tanto la calidad como la cantidad de aceite producido. Se examinarán las variables climáticas, como las precipitaciones y las temperaturas, y se explorará cómo estas condiciones ambientales pueden influir en la humedad del fruto.

3.3.1.3.1. Influencia de las temperaturas máximas

En la siguiente sección de nuestro estudio, nos enfocaremos en investigar la influencia de las temperaturas máximas en la humedad del fruto de la aceituna.

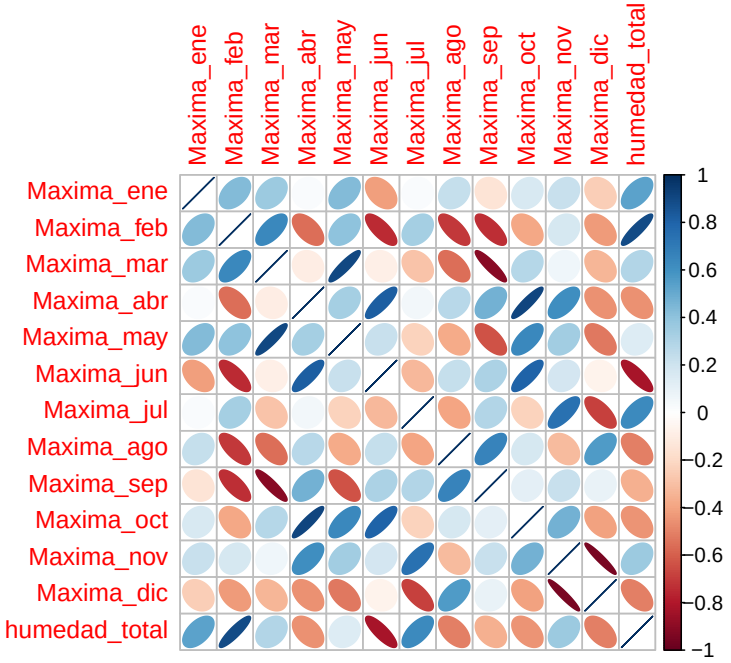


Figura 3.11: Influencia de las máximas en la humedad (fuente: elaboración propia)

Al examinar detalladamente el gráfico de correlaciones, se observa un patrón significativo en relación a las temperaturas máximas. Es evidente que los meses de verano, caracterizados por altas temperaturas, tienen un impacto negativo en la humedad del fruto de la aceituna.

Este resultado era de esperarse, ya que las altas temperaturas pueden acelerar la evaporación y la transpiración de la planta, lo que lleva a una mayor pérdida de agua en los frutos. Como consecuencia, la humedad del fruto disminuye durante los meses de verano. La influencia negativa de las altas temperaturas en la humedad del fruto de la aceituna puede tener implicaciones importantes en el desarrollo y la calidad de los frutos. Una baja humedad del fruto puede afectar la turgencia de las células, el metabolismo de la planta y la acumulación de nutrientes. Además, una baja humedad puede influir en la resistencia a enfermedades y en la calidad del aceite de oliva producido.

3.3.1.3.2. Influencia de las temperaturas mínimas

En este apartado de nuestro estudio, nos dedicaremos a analizar la influencia de las temperaturas mínimas en la humedad del fruto de la aceituna.

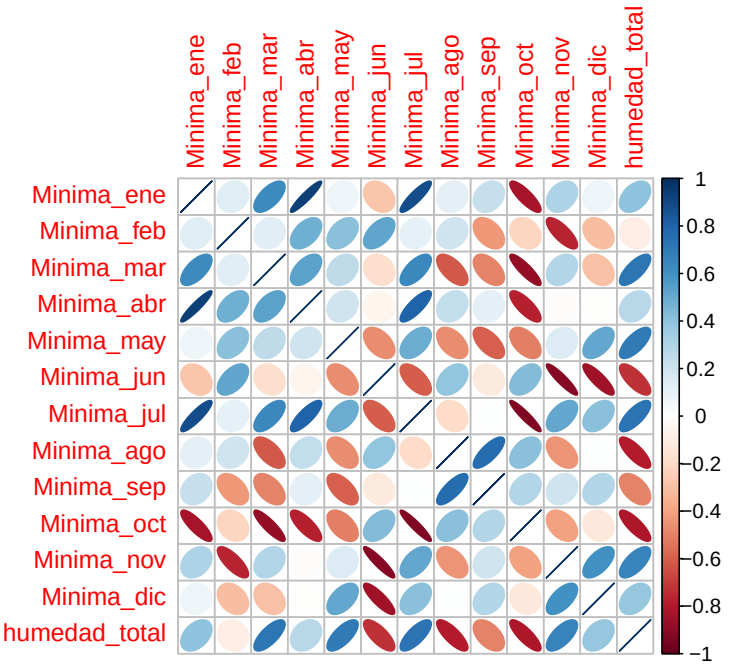


Figura 3.12: Influencia de las mínimas en la humedad (fuente: elaboración propia)

Al examinar detenidamente el gráfico de correlaciones, se puede observar un patrón significativo en relación a las temperaturas mínimas. Es notable que las temperaturas mínimas durante los meses de marzo, abril y mayo tienen una influencia positiva en la humedad del fruto.

Otro hallazgo interesante es la influencia positiva de las temperaturas mínimas durante el mes de julio en la humedad del fruto. Esto puede ser atribuido al hecho de que julio es un mes crítico en el desarrollo de la aceituna.

Sin embargo, es importante destacar que las temperaturas mínimas en el mes de octubre presentan una influencia negativa en la humedad del fruto. Esto puede deberse a que temperaturas mínimas más bajas en este momento pueden acelerar la maduración de la aceituna y, como resultado, se produce una disminución en la humedad del fruto.

3.3.1.3.3. Influencia de las precipitaciones

En esta sección, abordaremos la influencia de las precipitaciones en la humedad del fruto de la aceituna.

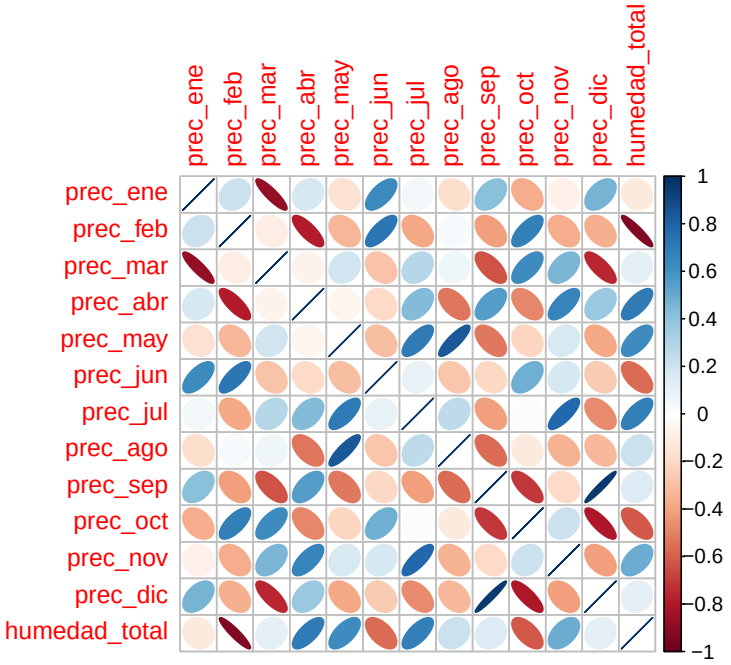


Figura 3.13: Influencia de las precipitaciones en la humedad (fuente: elaboración propia)

Al examinar detenidamente los datos, podemos apreciar que las precipitaciones tienen una influencia positiva en la humedad del fruto en la mayoría de los meses. Esto indica que un mayor volumen de precipitaciones favorece el mantenimiento de una adecuada humedad en los frutos de la aceituna.

Las precipitaciones durante la mayoría de los meses contribuyen al suministro de agua necesario para la planta y el desarrollo de los frutos. Además, es interesante destacar que las precipitaciones también influyen positivamente en la humedad del fruto durante el mes de julio. Esta correlación puede ser explicada por el hecho de que las precipitaciones en este mes contribuyen a mantener una hidratación adecuada de los frutos durante su fase de crecimiento y maduración.

En general, las precipitaciones actúan como una fuente fundamental de agua para la planta de olivo y sus frutos. Proporcionan la humedad necesaria para el adecuado funcionamiento de los procesos fisiológicos de la planta y, por ende, la humedad del fruto se ve favorecida.

3.3.1.4. Análisis de la influencia de la climatología en los kilogramos producidos

En esta sección, nos adentraremos en el análisis de cómo la climatología influye en la cantidad de kilogramos de aceituna producidos. Exploraremos variables climáticas importantes, como las precipitaciones y las temperaturas, y examinaremos cómo estas condiciones pueden impactar la producción de aceitunas en diferentes campañas. También investigaremos posibles relaciones y patrones entre la climatología y la cantidad de kilogramos producidos, lo que nos permitirá comprender mejor cómo los factores ambientales pueden afectar la productividad de los cultivos.

3.3.1.4.1. Influencia de las temperaturas máximas

En esta sección, nos enfocaremos en el estudio de la influencia de las temperaturas máximas en la cantidad de kilogramos de aceituna recogidos.

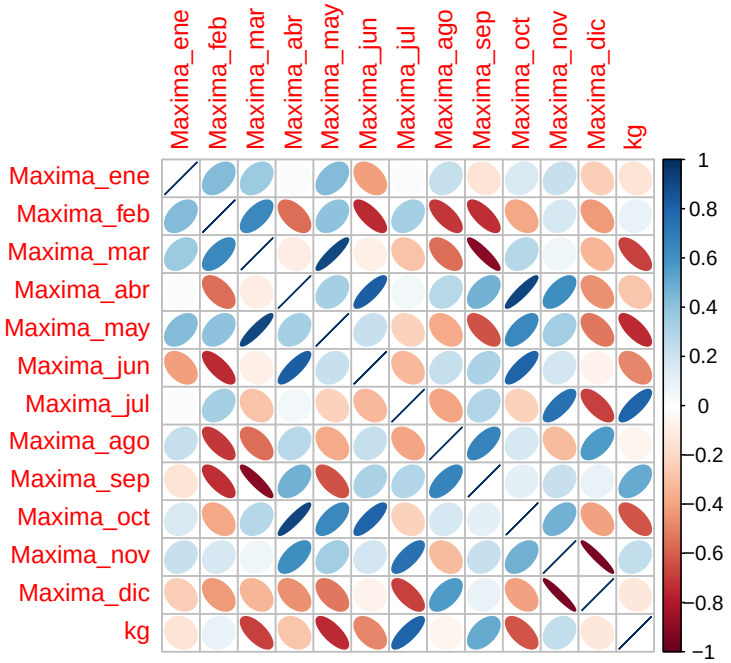


Figura 3.14: Influencia de las máximas en los kilogramos producidos (fuente: elaboración propia)

Al examinar detenidamente los datos, se observa claramente que las temperaturas máximas tienen una influencia muy negativa durante los meses de floración y crecimiento de la aceituna. Esto indica que temperaturas máximas más altas durante estas etapas tienen un impacto perjudicial en la cantidad de kilogramos de aceituna que se logran recoger.

Durante los meses de floración y crecimiento, las altas temperaturas máximas pueden provocar una serie de efectos adversos en los frutos de la aceituna. Estas altas temperaturas pueden causar estrés térmico en la planta, lo que a su vez puede afectar negativamente el desarrollo y la calidad de los frutos. Además, las altas temperaturas máximas pueden acelerar el proceso de maduración de los frutos, lo que resulta en una menor acumulación de aceituna y, por lo tanto, una disminución en la cantidad de kilogramos recogidos.

Es importante destacar que la floración y el crecimiento son etapas críticas en el ciclo de vida de la aceituna, ya que determinan en gran medida la formación de frutos y la posterior cosecha. Por lo tanto, es crucial mantener condiciones óptimas durante estos períodos, incluyendo temperaturas máximas moderadas, para asegurar un rendimiento satisfactorio en términos de kilogramos recogidos.

3.3.1.4.2. Influencia de las temperaturas mínimas

En esta sección, nos enfocamos en estudiar la influencia de las temperaturas mínimas en la cantidad de kilogramos de aceituna recogidos.

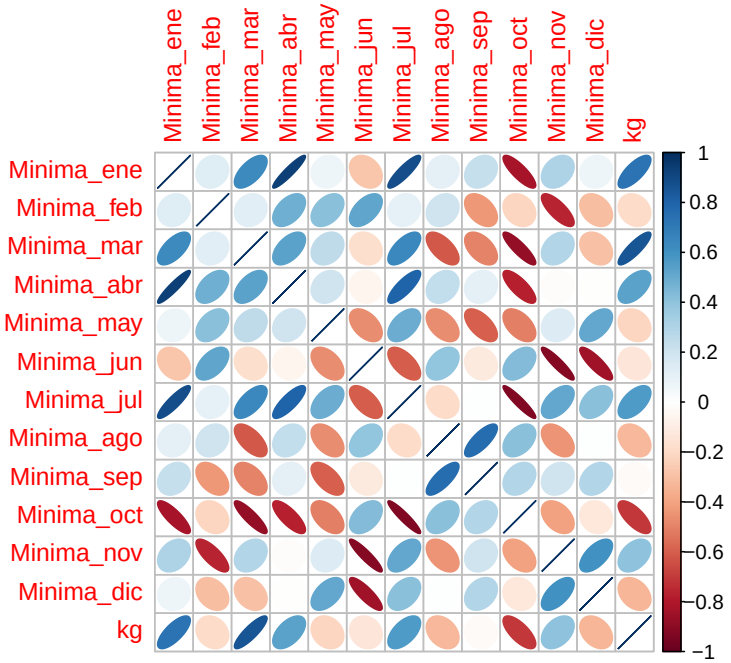


Figura 3.15: Influencia de las mínimas en los kilogramos producidos (fuente: elaboración propia)

Al examinar detenidamente los datos, podemos observar que las temperaturas mínimas durante los meses de floración y crecimiento del fruto tienen una influencia favorable en la consecución de mayores kilogramos de aceituna. Esto indica que temperaturas mínimas más altas durante estas etapas tienen un impacto positivo en la producción de kilogramos.

Durante la floración y el crecimiento de la aceituna, las temperaturas mínimas más altas proporcionan condiciones más favorables para el desarrollo adecuado de los frutos. Las altas temperaturas mínimas ayudan al proceso de maduración de los frutos, lo que a su vez promueve una mayor acumulación de aceituna y, por lo tanto, una mayor cantidad de kilogramos recogidos.

Por otro lado, es importante señalar que las temperaturas mínimas en meses como octubre pueden tener un efecto negativo en la cantidad de kilogramos recogidos. Las bajas temperaturas mínimas en este mes, que está cerca del período de recolección, pueden provocar la caída prematura del fruto o una menor acumulación de aceite en los mismos, lo que a su vez resulta en una disminución de los kilogramos recolectados.

3.3.1.4.3. Influencia de las precipitaciones

En esta sección, abordaremos la influencia de las precipitaciones en la cantidad de kilogramos de aceituna producidos.

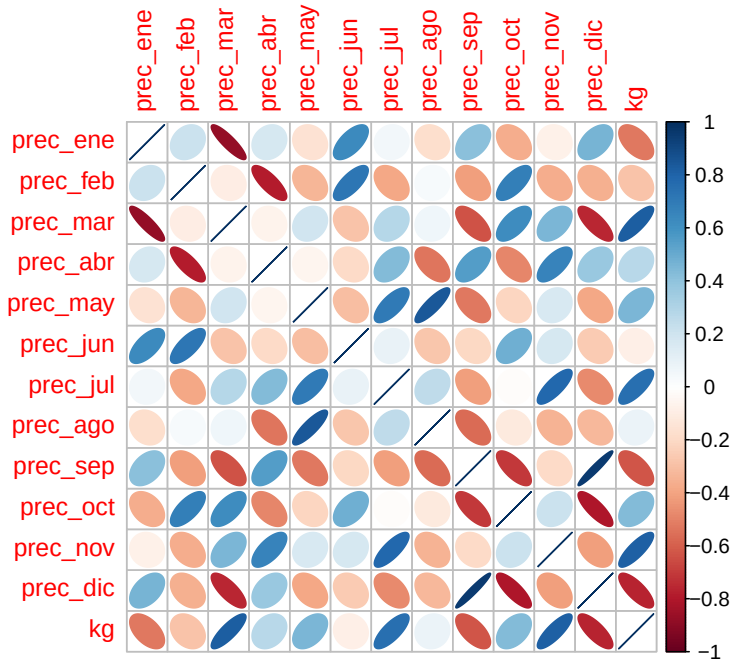


Figura 3.16: Influencia de las precipitaciones en los kilogramos producidos (fuente: elaboración propia)

Al analizar detenidamente los datos, podemos observar claramente que las precipitaciones tienen una influencia favorable en la producción de aceituna a lo largo de todos los meses. Esto indica que las precipitaciones, independientemente del momento del año, tienen un impacto positivo en la cantidad de kilogramos producidos.

Las precipitaciones desempeñan un papel crucial en el desarrollo de los árboles de olivo y en la formación de los frutos. Proporcionan el agua necesaria para el crecimiento adecuado de la planta y contribuyen a la acumulación de recursos necesarios para la producción de aceituna, como los nutrientes y la energía. Además, las precipitaciones ayudan a mantener niveles adecuados de humedad en el suelo, lo que favorece la absorción de agua y nutrientes por parte de las raíces de los olivos.

Es importante destacar que el suministro adecuado de agua a lo largo de todo el ciclo de vida de la aceituna es esencial para lograr una producción óptima de kilogramos. Las precipitaciones regulares y bien distribuidas proporcionan las condiciones ideales para el desarrollo saludable de los árboles y la formación de frutos de calidad.

Sin embargo, es importante mencionar que un exceso de precipitaciones o una distribución irregular pueden tener efectos negativos en la producción de aceituna. El exceso de agua puede llevar a problemas de drenaje y saturación del suelo, lo que puede afectar negativamente el desarrollo de las raíces y la disponibilidad de oxígeno para las plantas. Además, las lluvias intensas pueden provocar daños en la estructura de los frutos y favorecer la aparición de enfermedades.

3.3.1.5. Estudio de la influencia de las variables en estudio

En esta sección, nos enfocaremos en analizar la relación entre las variables de estudio.

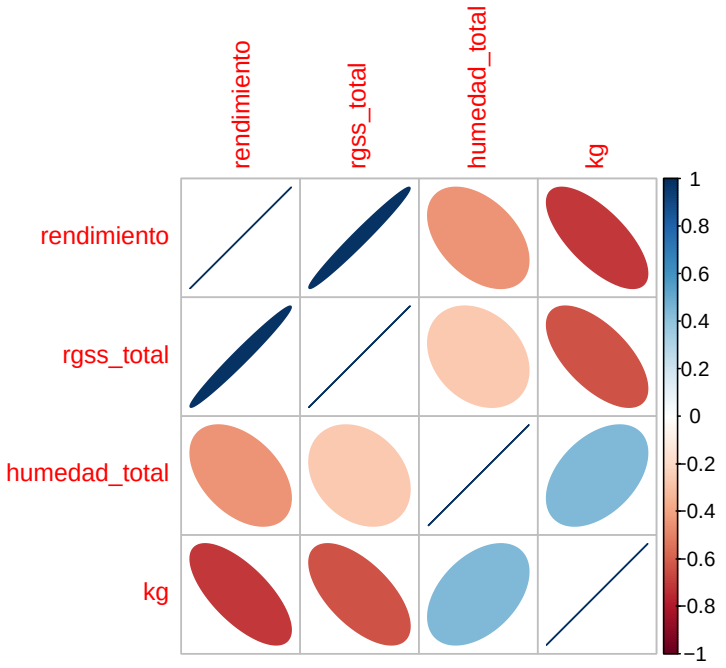


Figura 3.17: Influencia de las variables en estudio (fuente: elaboración propia)

Al examinar los datos y realizar el análisis de correlación, hemos encontrado una correlación positiva muy alta entre el rendimiento y el rendimiento graso. Esto significa que a medida que aumenta el rendimiento, también se observa un aumento en el rendimiento graso sobre materia seca. Esta asociación positiva indica que hay una tendencia de obtener un mayor rendimiento cuando se obtiene un mayor rendimiento graso de la aceituna.

Cuando se obtiene un mayor rendimiento graso sobre materia seca, significa que hay una mayor proporción de aceite en relación con la cantidad total de materia seca presente en la aceituna. Esto implica que, al procesar una determinada cantidad de aceitunas, se obtendrá una mayor cantidad de aceite. Por lo tanto, cuanto mayor sea el rendimiento graso, mayor será el rendimiento general de aceite obtenido.

Otro hecho significativo tiene lugar en relación a la correlación entre la humedad del fruto y la cantidad de kilos producidos de aceituna. Se ha encontrado una correlación positiva entre ambos factores, lo que implica que a medida que aumenta la humedad del fruto, también se incrementa la cantidad de kilos de aceituna producidos.

Esta relación puede explicarse por el aumento en el volumen del fruto debido a su mayor contenido de humedad. Cuanto más humedad tenga el fruto, mayor será su volumen y, por lo tanto, su peso. Esto conlleva una mayor cantidad de kilogramos de aceituna producidos en la cosecha.

La humedad del fruto está estrechamente relacionada con su capacidad para retener agua. Cuando el fruto retiene una mayor cantidad de agua, su volumen aumenta, lo que se traduce en un incremento en el peso de la aceituna recolectada. Por lo tanto, un fruto

con una mayor humedad tendrá un mayor peso y contribuirá a la producción de más kilos de aceituna.

Es importante destacar que el manejo adecuado de la humedad del fruto es fundamental para evitar problemas como la sobrehidratación o la deshidratación excesiva. Un equilibrio óptimo en la humedad del fruto, mediante técnicas de riego adecuadas y una correcta gestión de la cosecha, puede contribuir significativamente a maximizar la producción de kilos de aceituna.

Otro hecho destacable es la presencia de una correlación negativa entre el rendimiento y la cantidad de kilos producidos de aceituna. Esto significa que a medida que aumenta la cantidad de kilos de aceituna recolectados, el rendimiento tiende a disminuir.

Existen varias razones que pueden explicar esta correlación negativa. En primer lugar, cuando se obtiene una mayor cantidad de kilos de aceituna, es posible que la planta tenga que distribuir sus recursos de manera más diluida entre un mayor número de frutos. Esto puede llevar a una menor concentración de nutrientes y energía en cada fruto individual, lo que puede afectar negativamente su calidad y rendimiento.

Además, un aumento en la cantidad de kilos recolectados puede indicar una mayor densidad de plantas en un determinado espacio. Esto puede conducir a una mayor competencia por los recursos disponibles, como agua, nutrientes y luz solar. Como resultado, las plantas pueden tener un crecimiento más limitado y, en última instancia, una menor producción de aceitunas de calidad.

Otro factor que puede influir en la correlación negativa entre el rendimiento y la cantidad de kilos producidos es la variabilidad en la calidad de la cosecha. Es posible que, al recolectar una mayor cantidad de aceitunas, también se incluyan frutos de menor calidad, como aquellos que están dañados, inmaduros o de menor tamaño. Estos frutos de menor calidad pueden arrastrar el promedio hacia abajo, afectando así el rendimiento general.

3.3.2. Estudio gráfico de las variables en cada campaña

En esta sección, se presentarán dos gráficas para cada campaña agrícola, con el objetivo de analizar y visualizar las relaciones entre variables clave.

En la primera gráfica, se representarán las variables de humedad, rendimiento, rendimiento graso sobre materia seca y kilogramos de aceite producido en esa campaña. En el eje Y principal, se utilizará una escala porcentual para representar el rendimiento, la humedad y el rendimiento graso sobre materia seca. Estas variables estarán interrelacionadas y se trazarán como líneas para observar su evolución a lo largo de la campaña. En el segundo eje Y, se utilizará una escala de kilogramos para representar los kilogramos de aceite producidos.

En la segunda gráfica, se visualizarán los datos climáticos correspondientes a cada mes del año de la campaña agrícola. Se incluirán la precipitación, la temperatura máxima y la temperatura mínima. La precipitación se expresará en litros, mientras que las temperaturas se medirán en grados Celsius y se trazarán como líneas.

Estas gráficas proporcionarán una representación visual detallada y técnica de los datos recopilados. Permitirán observar las tendencias y relaciones entre las variables estudiadas, así como su posible influencia en la producción de aceite de oliva. Además, facilitarán la identificación de patrones estacionales y posibles correlaciones entre las condiciones climáticas y la producción de aceituna.

3.3.2.1. Campaña 2017/2018

Comenzaremos examinando, a través de una representación gráfica, la campaña 2017/2018, con el objetivo de analizar las tendencias observadas en relación al rendimiento, el rendimiento graso sobre materia seca, la humedad y los kilos de aceite producidos diariamente.

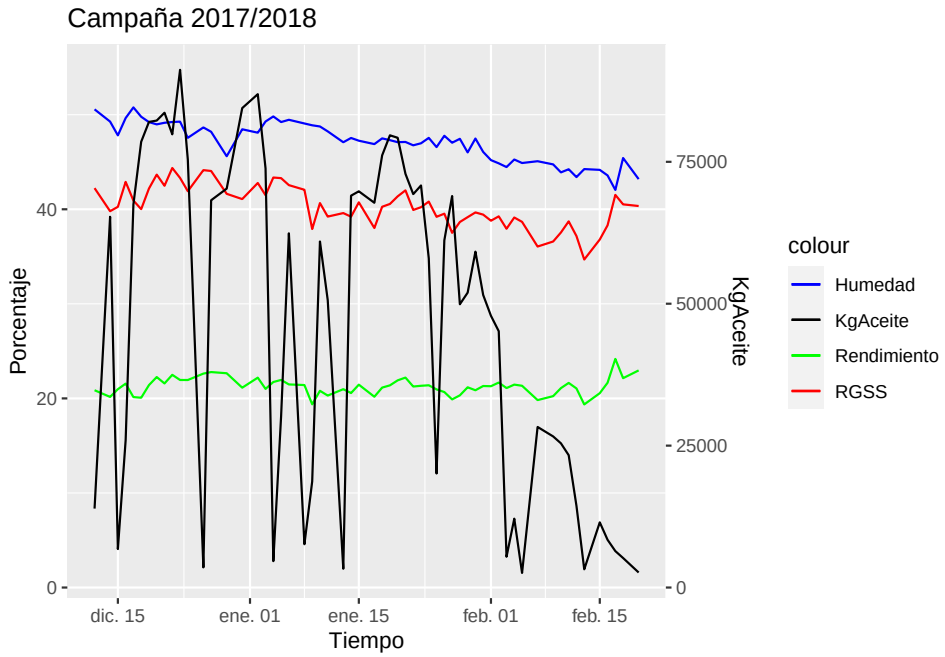


Figura 3.18: Campaña 2017/2018 (fuente: elaboración propia)

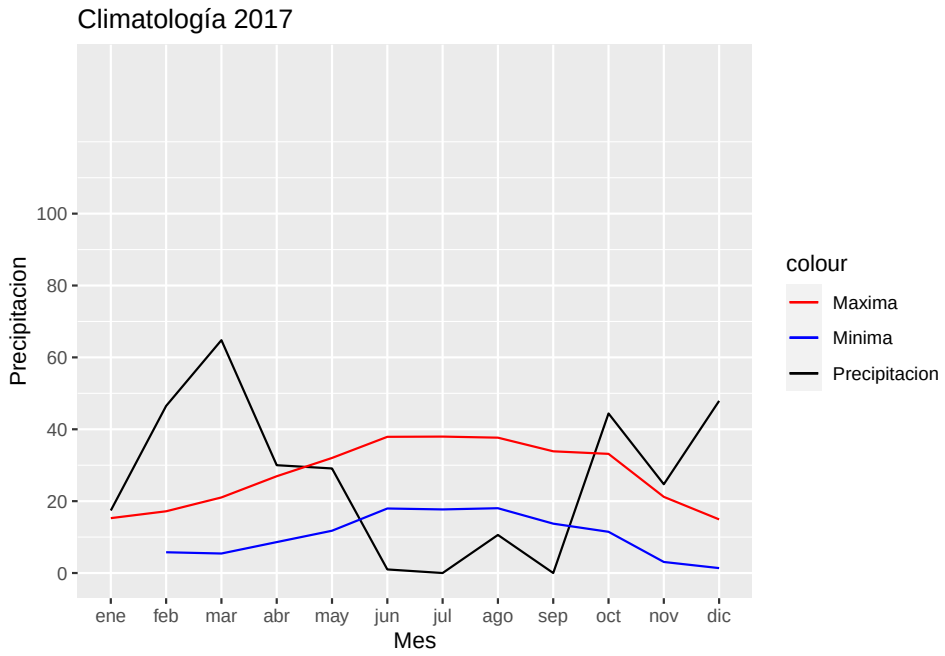


Figura 3.19: Condiciones climáticas en 2017 (fuente: elaboración propia)

En la gráfica 3.14, se puede apreciar que el rendimiento se mantiene consistentemente por encima del 20 % a lo largo de la campaña. No obstante, se observan variaciones suaves en forma de descensos y aumentos a lo largo del tiempo. Estos cambios pueden deberse a factores como las condiciones climáticas o la variabilidad en la calidad de la aceituna recolectada.

En cuanto al rendimiento graso sobre materia seca, se constata que inicialmente comienza por encima del 40 %. Sin embargo, a medida que avanza la campaña, se evidencia una disminución progresiva, de manera que hacia mediados de enero se sitúa por debajo del umbral del 40 %. Este descenso en el rendimiento graso sobre materia seca puede deberse a diversos factores, como el estado de madurez de la aceituna recolectada o las condiciones ambientales durante el proceso de producción.

Un patrón similar se observa en los niveles de humedad de la aceituna. Al inicio de la campaña, se registran niveles cercanos al 50 % de humedad en el fruto. Sin embargo, a medida que transcurren los días, se constata una disminución gradual de la humedad, alcanzando niveles cercanos al 40 % hacia el final de la campaña. Esta reducción de humedad puede estar influenciada por un factor fundamental: la madurez del fruto.

En relación a los kilos de aceite producidos diariamente, se observa que existen picos superiores a los 75.000 kilogramos hasta mediados de enero. No obstante, se evidencian descensos significativos debido a las precipitaciones, las cuales impiden la recolección del fruto. Además, hacia el final de la campaña, se puede apreciar un descenso dramático en la producción diaria de aceite. Esto es debido a que muchos socios han finalizado su campaña.

En cuanto a la climatología del año analizado, se destaca un patrón caracterizado por altas temperaturas y escasas precipitaciones. Desde el mes de mayo hasta octubre, se registraron temperaturas persistentemente elevadas, lo que indica un año especialmente caluroso. Por otro lado, en términos de precipitaciones, se observó una notable falta

de lluvia a lo largo del año. Especificamente, el mes de marzo se destacó como el más lluvioso en comparación con los demás meses, siendo el momento en el que se registraron las mayores cantidades de precipitación.

3.3.2.2. Campaña 2018/2019

En la campaña 2018/2019, se puede observar un patrón interesante en los datos recopilados. Inicialmente, los rendimientos se encuentran por debajo del 20 %, indicando un inicio de temporada con un desempeño relativamente bajo. Sin embargo, a partir de mediados de diciembre, estos rendimientos comienzan a situarse consistentemente por encima de dicho umbral, alcanzando sus valores más altos a mediados de enero.

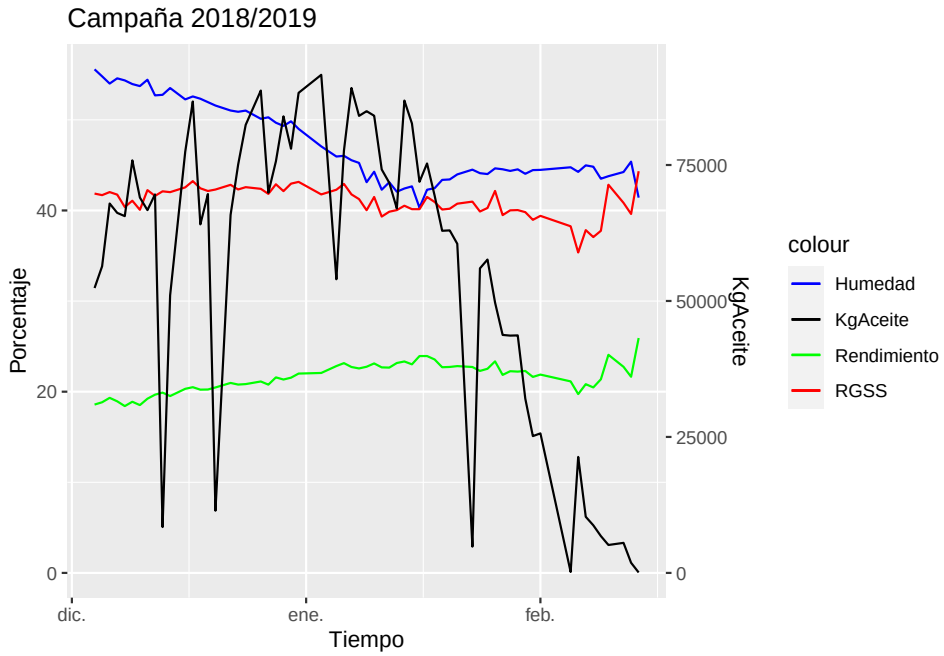


Figura 3.20: Campaña 2018/2019 (fuente: elaboración propia)

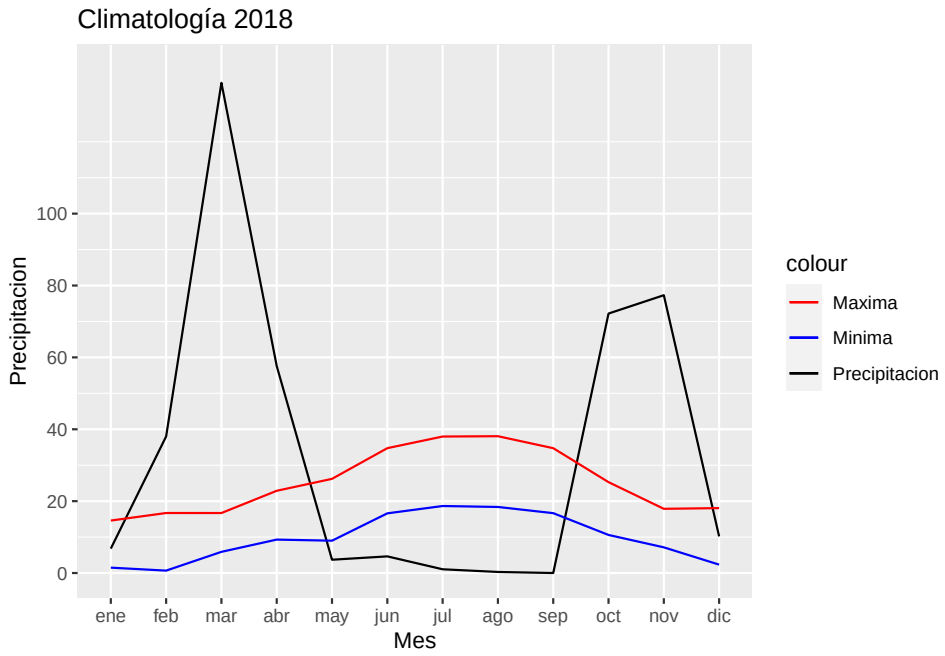


Figura 3.21: Condiciones climáticas en 2018 (fuente: elaboración propia)

En cuanto al rendimiento graso sobre materia seca, durante toda la campaña se mantiene por encima del 40 %, lo que indica un contenido de grasa significativo en el producto final. Sin embargo, a principios de febreo se observa una ligera disminución de este valor, descendiendo por debajo de dicho umbral.

En relación a la humedad en el fruto, se registra un inicio de campaña con niveles muy altos, superando el 50 %. No obstante, a medida que avanza la temporada, la humedad disminuye considerablemente, llegando a situarse un 10 % por debajo de los valores iniciales.

En términos de los kilogramos de aceite recogidos, se puede notar un incremento progresivo a lo largo de la campaña, superando los 75.000 kilogramos a medida que avanza el tiempo. Sim embargo, a mediados de enero, con el fin de la campaña, se registra una disminución en la producción, descendiendo por debajo de dicho umbral.

Por último, respecto a la climatología de esta camapaña en particular, se evidencia un verano caracterizado por condiciones secas y calurosas. Sin embargo, es importante destacar que durante los meses de marzo y abril se registraron precipitaciones significativas, así como en los meses de octubre, noviembre y diciembre. Por otro lado, es interesante observar que, excluyendo los meses de junio, julio, agosto y septiembre, el resto del año se caracterizó por temperaturas suaves.

3.3.2.3. Campaña 2019/2020

La campaña 2019/2020 destaca por presentar un rendimiento muy satisfactorio. Inicia con rendimientos ligeramente superiores al 20 %, los cuales van incrementando a lo largo de la campaña hasta alcanzar un valor cercano al 23 %. Aunque al final de la campaña se registra una ligera disminución, los rendimientos se mantienen en torno al 21 %, lo que indica un buen desempeño general.

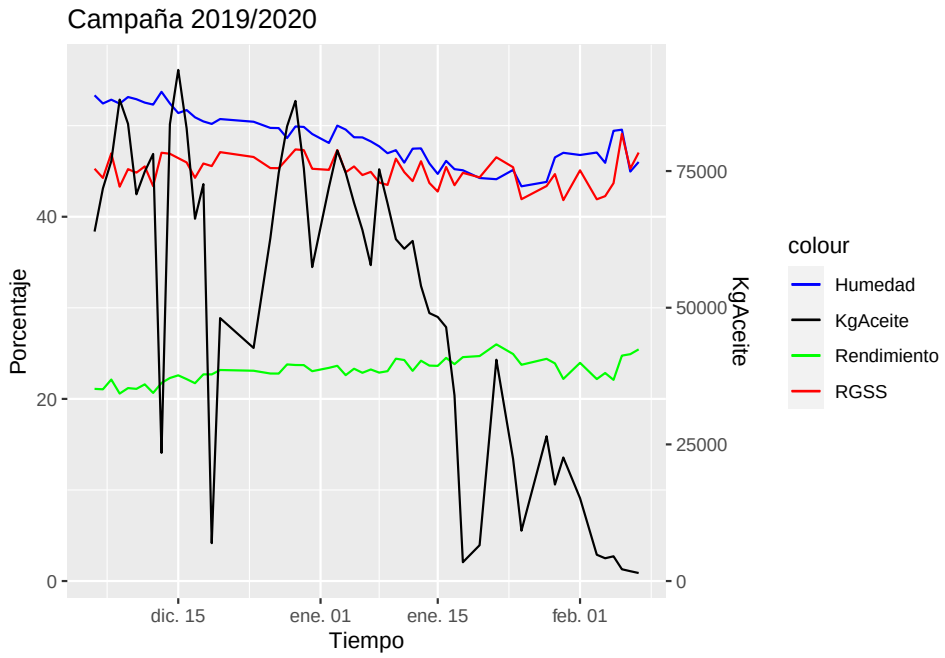


Figura 3.22: Campaña 2019/2020 (fuente: elaboración propia)

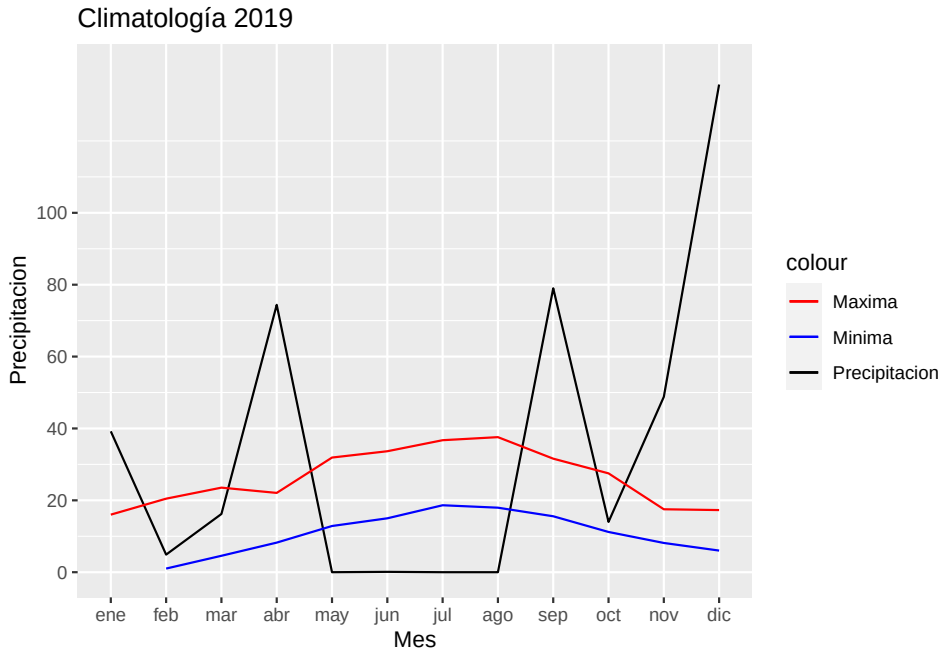


Figura 3.23: Condiciones climáticas en 2019 (fuente: elaboración propia)

En cuanto al rendimiento graso sobre materia seca, se mantiene prácticamente constante, con fluctuaciones poco significativas, durante toda la campaña, con valores que oscilan entre el 42 % y el 43 %. Esto demuestra una consistencia en la calidad del aceite producido a lo largo del tiempo.

En relación a la humedad en el fruto, se observa que al incio de la campaña es ligeramente superior al 50 %. Sin embargo, experimenta un descenso gradual hasta estabilizarse en torno al 42 % para experimentar una subida al final de la campaña.

Respecto a los kilogramos de aceite producidos, esta campaña se caracteriza por obtener números favorables hasta mediados de enero. A partir de ese momento, la producción comienza a disminuir, marcando el final de la campaña. Es importante destacar que, a pesar de la disminución, se mantienen valores aceptables en términos de producción de aceite.

En cuanto a la climatología, el año de estudio se caracteriza por ser excepcionalmente seco y caruloso. Durante un período extendido de mayo hasta agosto, no se registra ninguna precipitación significativa. Sin embargo, es importante destacar que los meses de abril y septiembre presentan una mayor cantidad de litros de agua registrados, siendo estos los meses más lluviosos del año.

3.3.2.4. Campaña 2020/2021

La campaña 2020/2021 se caracteriza por presentar rendimientos relativamente bajos. Al analizar la gráfica, podemos observar que los rendimientos iniciales de esta campaña comienzan aproximadamente 2 puntos por debajo del umbral del 20 %. A medida que avanza el mes de diciembre, se produce una disminución en los rendimientos, pero hacia finales de mes, se registra una ligera subida que los acerca nuevamente al 20 %. A lo largo de la campaña, los rendimientos se mantienen en torno a este valor, lo que indica que no se trata de una campaña destacada en cuanto a rendimientos.

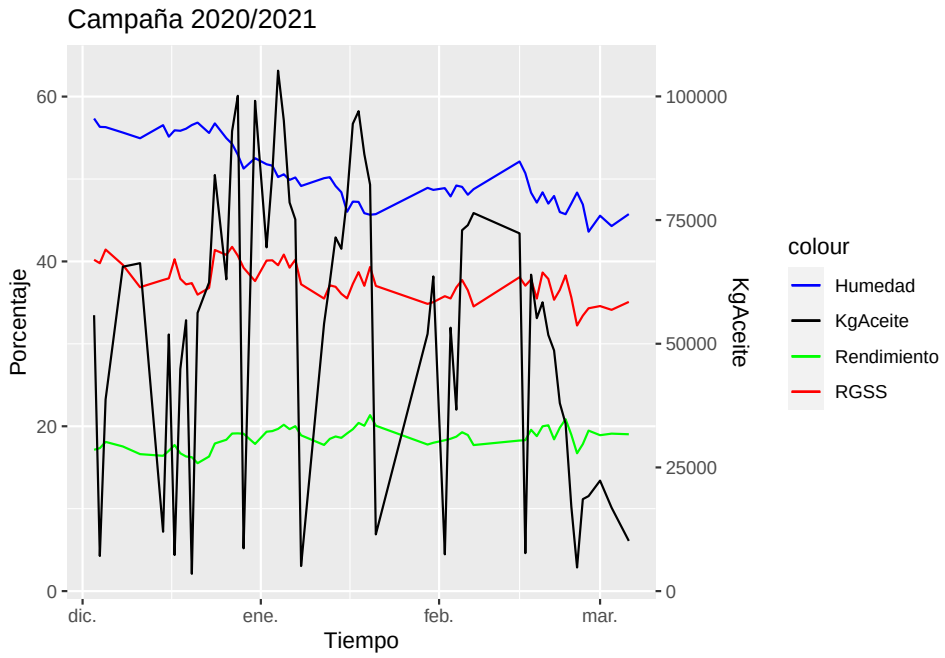


Figura 3.24: Campaña 2020/2021 (fuente: elaboración propia)

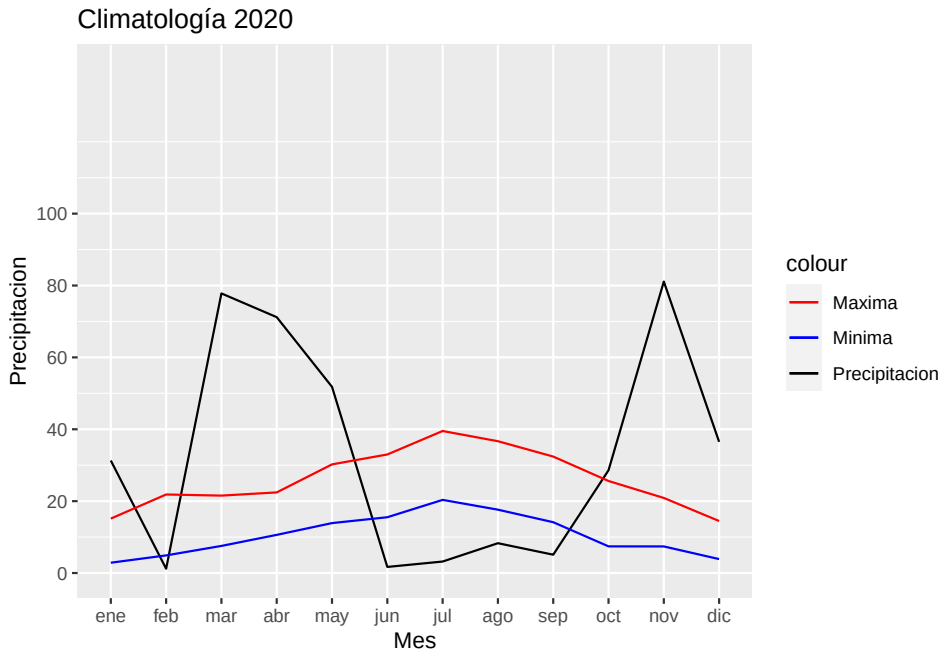


Figura 3.25: Condiciones climáticas en 2020 (fuente: elaboración propia)

En relación al rendimiento graso sobre materia seca, se observa que comienza en un 40 % y experimenta un ligera disminución, que luego es contrarrestada con un aumento hacia finales de diciembre. Sin embargo, posteriormente continúa descendiendo hasta ubicarse en alrededor del 37 % durante el resto de la campaña.

En cuanto a la humedad del furto, se inicia la campaña con un nivel muy alto, del 57 %, que se mantiende durante diciembre. A principios de enero, la humedad comienza a disminuir y se situa por debajo del 50 %. Durante la última semana de enero y principios de febrero, experimenta un ligero aumneto hasta alcanzar el 50 %, pero posteriormente vuelve a descender hasta el 47 % al final de la campaña.

En relación a la cantidad de aceite producido, durante el primer mes de la campaña se registran cifras por debajo de 40.000 kilogramos. Sin embargo, desde finales de diciembre hasta mediados de enero, se observa un aumento significativo, de manera que pasa a estar por debajo de los 60.000 kilogramos diarios. Durante el mes de febrero, los registros continúan siendo destacables, y es notable destacar que la campaña se prolonga hasta el mes de marzo.

Respecto a la climatología del año 2020, es notable que a pesar de ser un año con un verano caluroso, se registraron precipitaciones significativas. Es importante destacar que la primavera se caracterizó por ser muy lluviosa y además, el otoño también presentó buenos registros en términos de precipitaciones.

3.3.2.5. Campaña 2021/2022

La campaña 2021/2022 se considera una campaña exitosa en cuanto a rendimientos. Al analizar la gráfica, podemos observar que en ningún momento los rendimientos caen por debajo del umbral del 20 %, lo cual es muy positivo. De hecho, la campaña comienza con rendimientos superiores al 20 % y experimenta un crecimiento durante las tres primeras

semanas, alcanzando niveles cercanos al 25 %. A partir de ese punto, los rendimientos disminuyen, pero se mantienen ligeramente por encima del 20 % durante el resto de la campaña.

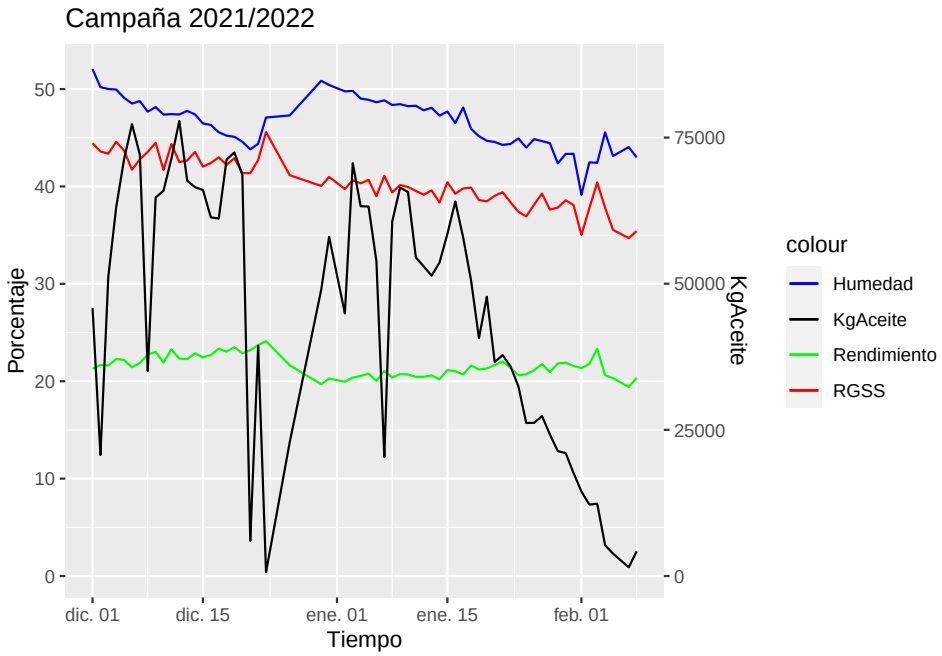


Figura 3.26: Campaña 2021/2022 (fuente: elaboración propia)

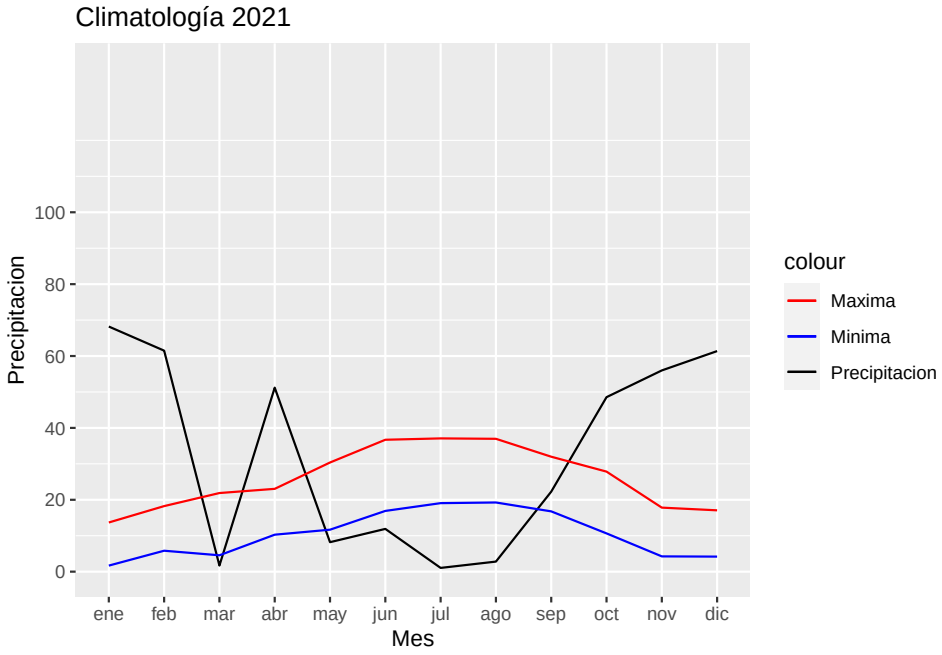


Figura 3.27: Condiciones climáticas en 2021 (fuente: elaboración propia)

Respecto a la humedad del fruto, al inicio se registra un nivel del 50 %. Durante las tres pimeras semanas, se observa una disminución de la humedad, alcanzando un mínimo del 45 %, seguido de un aumento hasta el 50 % durante la última semana del mes.

Posteriormente, se observa una disminución constante de la humedad durante el resto de la campaña, llegando a un nivel del 40 %. Sin embargo, es importante mencionar que los últimos días de la campaña se registra un aumneto hasta el 45 %.

En relación a la cantidad de aceite producido, se observa que durante el primer mes de la campaña, se registran cifras cercanas a los 75,000 kilogramos. Sin embargo, se observa un descenso significativo a finales de diciembre y principios de enero, debido a las precipitaciones. Posteriormente, se vuelve a registrar un incremento en las cifras hasta finales de enero, momento en el cual se produce un descenso debido al final de la campaña.

3.3.3. Conclusiones

A partir de este exhaustivo estudio, que incluye el análisis de correlaciones y la representación gráfica de las variables relacionadas tanto con la aceituna como con las condiciones climáticas, se derivan conclusiones significativas. Una de ellas se refiere a una interrogante frecuente entre los agricultores, que es cuándo es mejor iniciar la campaña de recolección de aceitunas, ya sea a principios o mediados de diciembre. Esta pregunta surge debido a que, como se puede observar, los rendimientos en las dos primeras semanas tienden a ser más bajos. Los agricultores suelen considerar esta medida como un indicador para determinar el momento adecuado de inicio de la campaña, lo que lleva a muchos agricultores a preferir retrasar dicho inicio.

Sin embargo, es crucial tener en cuenta un factor imperceptible pero relevante: la humedad del fruto. Mediante el análisis de correlación y los gráficos correspondientes a las distintas campañas, se ha constatado una correlación negativa entre ambas variables. Es decir, a medida que los rendimientos aumentan a lo largo de la campaña, la humedad del fruto tiende a disminuir. Por el contrario, al inicio de la campaña, cuando los rendimientos son más bajos, la humedad del fruto se encuentra en niveles más elevados.

Esta condición de mayor humedad del fruto en las etapas iniciales de la campaña resulta beneficioso por diversas razones. En primer lugar, una mayor humedad en las aceitunas contribuye a incrementar su peso, lo cual puede ser favorable desde el punto de vista comercial, ya que se obtiene una mayor cantidad de materia prima para la extracción de aceite de oliva.

Además, la humedad en el fruto puede tener un impacto positivo en la calidad del aceite producido. Una mayor humedad se asocia con una menor oxidación de los compuestos presentes en las aceitunas, lo que puede preservar y potenciar las características organolépticas y nutricionales del aceite de oliva resultante.

Por lo tanto, el hecho de que la aceituna tenga mayor humedad al comienzo de la campaña puede ser beneficioso tanto en términos de kilogramos recogidos como en la calidad del fruto. Luego, considerar este aspecto en la toma de decisiones relacionadas con el momento de inicio de la campaña resulta fundamental para maximizar tanto la calidad como la cantidad de aceite producido.

3.4. Promedio de la recolección temprana y de la campaña y correlación entre ambas

En esta sección, llevaremos a cabo un estudio de gran relevancia centrado en la comparación de los rendimientos entre lo que denominaremos la campaña de recolección de aceituna para aceite temprano y la campaña oficial. El objetivo principal de este estudio es analizar la posible influencia del rendimiento en la cosecha temprana como un indicador predictivo del rendimiento durante la campaña oficial.

La campaña de recolección de aceituna para aceite temprano se refiere al periodo en el que se realiza la recolección de aceitunas en etapas tempranas para obtener aceites con características organolépticas específicas. Por otro lado, la campaña oficial se refiere al periodo en el que se realiza la recolección de aceitunas en su punto óptimo de maduración para la obtención de aceite de oliva de calidad.

El objetivo de establecer esta comparación es determinar si los rendimientos obtenidos durante la cosecha temprana pueden proporcionar información relevante y predictiva sobre los rendimientos que se obtendrán durante la campaña oficial. De esta manera, se pretende evaluar la viabilidad de utilizar el rendimiento de la cosecha temprana como herramienta para anticipar y estimar el rendimiento global de la campaña.

Para llevar a cabo este estudio, emplearemos datos recopilados durante las campañas de recolección de aceituna correspondientes a los periodos 2017/2018 a 2021/2022. Durante estas campañas, se registraron datos tanto para la recolección temprana como para la campaña oficial, lo que nos proporcionará una amplia muestra de información.

Para cada campaña, calcularemos la media de los rendimientos obtenido tanto en la cosecha temprana como en la campaña oficial. Estos valores medios nos servirán como representación de los rendimientos en cada etapa, permitiéndonos comparar y analizar su relación.

```
##   media_temprano media_notemprano
## 1       16.23000         21.56738
## 2       14.42000         23.12245
## 3       14.21000         18.47580
## 4       16.64583         21.54198
## 5       13.35667         19.31586
```

Realizaremos un estudio de correlación entre los rendimientos de la cosecha temprana y los rendimientos de la campaña oficial en cada una de las campañas consideradas.

```
##                  media_temprano media_notemprano
## media_temprano        1.0000000        0.4826289
## media_notemprano      0.4826289        1.0000000
```

En el análisis realizado, se encuentra una correlación positiva significativa de 0.4826289 entre los rendimientos de la cosecha temprana y los rendimientos de la campaña oficial de aceituna. Esta correlación indica que existe una relación entre ambas variables, aunque no es una relación directa y fuerte.

La correlación positiva de 0.4826289 sugiere que hay una tendencia general de que los rendimientos obtenidos durante la cosecha temprana estén relacionados con los rendimientos de la campaña oficial. Sin embargo, es importante tener en cuenta que la correlación no implica causalidad directa y otros factores también pueden influir en los rendimientos finales.

Esta correlación moderada puede considerarse como un indicio o una pista que nos indica que los rendimientos tempranos podrían proporcionar cierta información predictiva sobre los rendimientos generales de la campaña oficial. Sin embargo, es necesario realizar análisis más detallados y considerar otros factores relevantes para obtener conclusiones más precisas.

Es importante destacar que la correlación de 0.4826289 no es una relación perfecta y que otros factores, como condiciones climáticas, prácticas agrícolas y eventos imprevistos, también pueden influir en los rendimientos de la campaña oficial. Por lo tanto, es recomendable utilizar esta correlación como una referencia inicial y complementarla con otros análisis y consideraciones antes de tomar decisiones basadas únicamente en el rendimiento temprano.

Capítulo 4

Predicción del rendimiento y kilogramos de aceituna recogidos.

En el próximo capítulo, nos centraremos en la predicción del rendimiento y los kilogramos de aceituna recogidos. Utilizaremos técnicas de modelado estadístico y de aprendizaje automático para desarrollar modelos que nos permitan estimar con precisión y fiabilidad estas variables de interés. Para ello, exploraremos diferentes enfoques y algoritmos, ajustaremos los modelos a nuestros datos y evaluaremos su desempeño. El objetivo es encontrar el modelo que mejor se ajuste a nuestros datos y nos brinde las predicciones más precisas. Analizaremos los resultados obtenidos y discutiremos sus implicaciones en relación con el problema de investigación. Además, consideraremos posibles mejoras y limitaciones de los modelos, así como áreas de estudio adicionales que podrían explorarse en el futuro.

4.1. Desarrollo de modelos predictivos para estimar el rendimiento

En esta sección, se procederá a aplicar las técnicas de predicción descritas en la sección 2.2 en el contexto de modelos predictivos basados en la predicción del rendimiento. Para llevar a cabo este análisis, se utilizará un conjunto de datos previamente mencionado, el cual cuenta con variables predictoras relacionadas con las precipitaciones y temperaturas mensuales, así como la variable objetivo que representa el rendimiento.

El primer paso consistirá en dividir este conjunto de datos en dos subconjuntos: el conjunto de entrenamiento y el conjunto de prueba. El conjunto de entrenamiento representará el 75 % del total de datos, mientras que el conjunto de prueba será el 25 % restante. Esta división permitirá evaluar la capacidad predictiva del modelo en datos no vistos previamente. Para ello nos valdremos de la función initial_split() del paquete rsample introducido en la sección 2.3.

A continuación, se aplicarán técnicas de ingeniería de características, prepocesando los datos para normalizar las variables predictoras. Estas técnicas incluirán el centrado y escalado de las variables, lo cual permitirá que todas ellas se encuentren en una escala comparable. El centrado consistirá en restar la media de cada variable, mientras que el escalado implicará dividir por la desviación estándar. Esta normalización será crucial para evitar problemas de sesgo ocasionados por diferencias en las magnitudes de las variables.

Una técnica que nos ofrece la librería tidymodels es la herramienta strata en initial_split(). Al utilizar strata en initial_split(), se especifica una variable que se utilizará para estratificar los datos. Esta variable debe ser la variable objetivo o cualquier otra variable categórica relevante que represente las clases o grupos de interés. La función entonces realiza la partición asegurando que cada conjunto (entrenamiento y prueba) contenga una representación adecuada de todas las clases presentes en la variable estratificadora (o al menos manteniendo la misma representación en el conjunto de entrenamiento y el de prueba).

Sin embargo, en nuestro caso particular, hemos encontrado una limitación para aplicar la partición estratificada utilizando strata en initial_split() debido al tamaño pequeño de nuestro conjunto de datos. La partición estratificada requiere una cantidad suficiente de datos en cada clase para realizar una distribución equitativa en los conjuntos de entrenamiento y prueba. Si el conjunto de datos es pequeño, puede haber dificultades para lograr una estratificación adecuada y se corre el riesgo de obtener subconjuntos desequilibrados, lo que puede afectar la validez de la evaluación del modelo.

La técnica de validación cruzada k-fold, con un valor de k igual a 10, se utilizará para evaluar el rendimiento de los modelos de manera más robusta y representativa. Sin embargo, es importante tener en cuenta que esta técnica sufrirá el mismo problema relacionado con la estratificación. A pesar de este desafío, el uso de la validación cruzada k-fold será beneficioso para comparar los modelos en promedio a través de 10 particiones distintas, lo que proporcionará medidas de rendimiento más confiables y robustas. Esto es especialmente relevante dado el tamaño relativamente pequeño del conjunto de datos, con alrededor de 20 observaciones. Además, esta técnica permitirá mitigar la influencia de la elección específica de la muestra de prueba, que puede tener un impacto significativo en los resultados de los modelos.

Es importante destacar esta limitación en nuestra metodología y reconocer que la partición estratificada habría sido deseable si tuviéramos un conjunto de datos más grande con clases bien representadas.

En el proceso de evaluación de los modelos predictivos, es fundamental analizar las métricas de rendimiento para comprender su eficacia y seleccionar el modelo más adecuado. En nuestro estudio, hemos observado que el coeficiente de determinación R cuadrado de los modelos es notablemente bajo. El coeficiente de determinación R cuadrado es una medida estadística que indica la proporción de la varianza en la variable dependiente que puede ser explicada por el modelo. Un valor bajo de R cuadrado sugiere que el modelo no puede explicar de manera adecuada la variabilidad en los datos observados.

Debido a esta baja eficacia en la capacidad de los modelos para explicar la varianza, hemos decidido utilizar el error cuadrático medio (RMSE) como índice principal para la comparación de los modelos. El RMSE es una medida de la diferencia entre los valores predichos por el modelo y los valores observados en los datos de prueba. Cuanto menor sea el valor del RMSE, mejor será la capacidad del modelo para predecir con precisión los resultados reales, y nos servirá como decíamos para poder comparar los modelos, aunque no dejemos de tener en mente que sería deseable que todos ellos tuvieran un R2 más aceptable.

El uso del RMSE como índice de evaluación es especialmente apropiado en nuestro caso, ya que nos permite cuantificar el error promedio de las predicciones en términos

de las unidades originales de la variable objetivo. Esto nos brinda una comprensión más intuitiva y significativa de la calidad de las predicciones en el contexto de nuestro estudio.

Al considerar el RMSE como nuestro índice principal, podemos comparar y seleccionar entre los diferentes modelos propuestos aquel que presente el menor error promedio en las predicciones. Esto nos permitirá identificar el modelo que mejor se ajuste a nuestros datos y que tenga la capacidad de predecir con mayor precisión el rendimiento en función de las variables predictoras utilizadas.

4.1.1. Serie Temporal

En esta sección abordaremos la predicción del rendimiento de los meses de diciembre, enero y febrero, utilizando un enfoque basado en series temporales. Nuestro objetivo es utilizar los datos de rendimiento disponibles desde 2001 para calcular la media de rendimiento de cada uno de los meses mencionados anteriormente.

Para llevar a cabo este análisis, comenzaremos recopilando los datos de rendimiento de los años comprendidos entre 2001 y el presente. A continuación, identificaremos los meses de diciembre, enero y febrero de cada año y calcularemos la media de rendimiento correspondiente a cada uno de estos meses.

Sin embargo, es importante destacar que en los años 2005 y 2006 no contamos con datos disponibles para el mes de febrero. Para evitar interrupciones en la serie temporal y asegurar una frecuencia consistente, imputaremos los valores faltantes del mes de febrero utilizando la media de los meses de febrero de los años restantes. De esta manera, todos los años tendrán un valor de 3 para el mes de febrero, garantizando la coherencia de la serie.

Una vez que hayamos calculado las medias de rendimiento para los meses de diciembre, enero y febrero, estaremos en condiciones de utilizar estos valores como base para predecir el rendimiento de estos meses en el futuro.

En primer lugar, realizaremos una representación gráfica de la serie temporal obtenida para los meses de diciembre, enero y febrero.

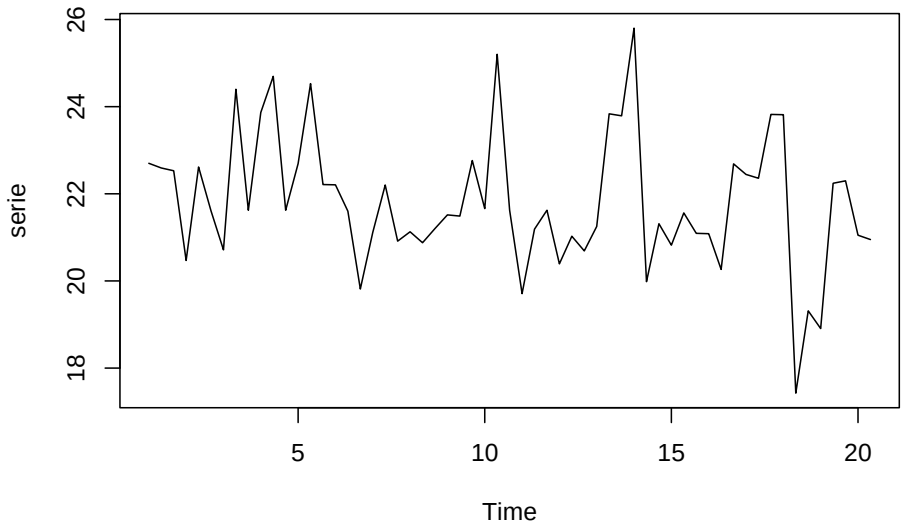


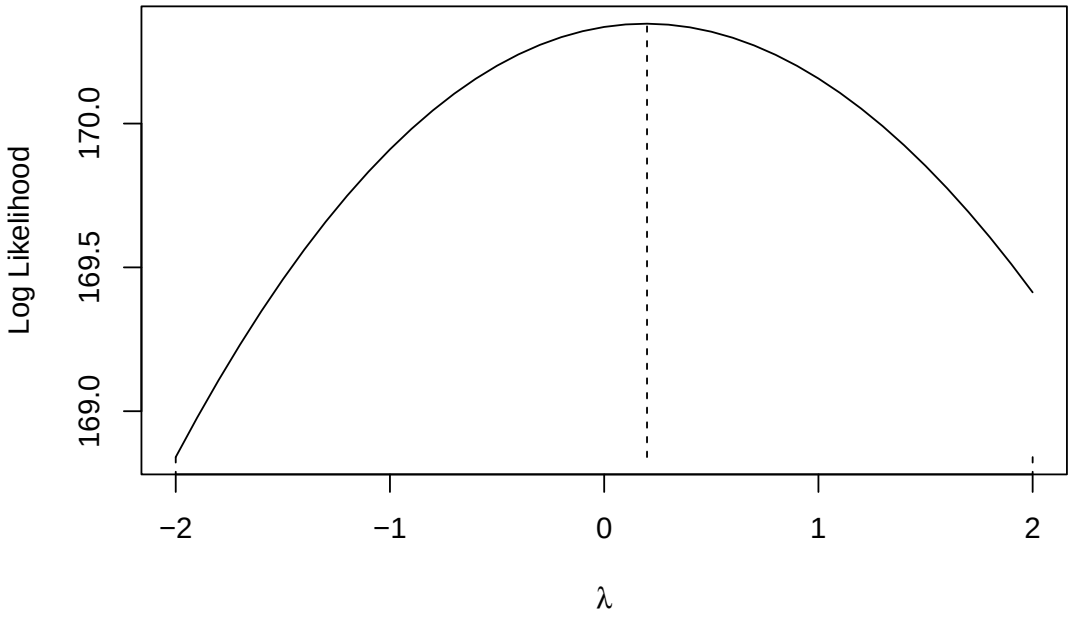
Figura 4.1: Serie de rendimiento (fuente: elaboración propia)

Al observar la gráfica, es evidente que la varianza de la serie no es estable a lo largo del tiempo. Esto implica que la serie no cumple con uno de los supuestos fundamentales de muchos modelos estadísticos y de series temporales, que es la homocedasticidad.

Para abordar esta problemática, utilizaremos una técnica conocida como transformación de Box-Cox (descrita con detalle en la sección 2.1). Esta transformación se emplea para estabilizar la varianza de la serie temporal y garantizar una mayor validez en la aplicación de modelos estadísticos y de series temporales.

La transformación de Box-Cox se basa en la idea de aplicar una función matemática a los valores de la serie que ayuda a lograr una distribución más cercana a la normalidad y, en consecuencia, una varianza más estable.

```
## Transformaciones para que la varianza sea estable con el tiempo
library(TSA)
bc=BoxCox.ar(y=serie)
```



La función utilizada en esta transformación depende de un parámetro lambda, que determina el tipo de transformación aplicada. En nuestro caso, utilizaremos el logaritmo neperiano. Una vez aplicada la transformación de Box-Cox y obtenida una serie temporal con varianza estable, podremos proceder con el análisis y modelado de la serie para realizar las predicciones correspondientes.

```
serie_transformada = log(serie)
```

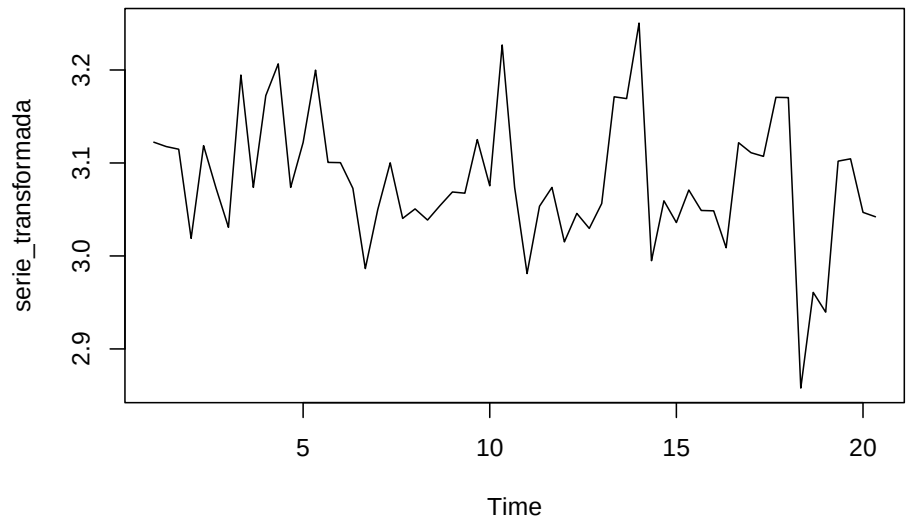


Figura 4.2: Serie de rendimiento transformada (fuente: elaboración propia)

Una vez hemos aplicado la transformación de Box-Cox y logrado estabilizar la varianza de la serie, podemos observar que la media también es estable. En caso de no serlo, deberíamos estabilizarla.

Después de confirmar que la serie está estabilizada tanto en términos de varianza como de media, procedemos a realizar el test de Dickey-Fuller. Este test es una prueba estadística utilizada para evaluar la estacionaridad de una serie temporal. La hipótesis nula del test es que la serie no es estacionaria, mientras que la hipótesis alternativa es que la serie es estacionaria.

Es importante destacar que la estacionaridad de la serie es un requisito fundamental para muchos modelos y técnicas de análisis de series temporales. Si la serie no es estacionaria, es posible que sea necesario aplicar técnicas adicionales, como diferenciación, para lograr la estacionaridad deseada.

```
### Contrastar estacionariedad
library(tseries)
adf.test(serie_transformada)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  serie_transformada
## Dickey-Fuller = -4.6364, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

El análisis del test de Dickey-Fuller nos dió un p-valor de 0.01. Este resultado indica que el p-valor es menor que el umbral de significación seleccionado,que es 0.05. Por lo tanto, podemos rechazar la hipótesis nula de no estacionaridad y concluir que la serie es estacionaria.

El p-valor de 0.01 es una medida de la evidencia estadística en contra de la hipótesis nula. Al ser un valor tan bajo, indica una alta confianza en la estacionaridad de la serie temporal. En otras palabras, hay una sólida base para afirmar que la serie presenta propiedades estacionarias. Procedemos a ajustar nuestra serie transformada a un modelo.

```
library(forecast)
ajuste=auto.arima(serie_transformada)
ajuste
```

```
## Series: serie_transformada
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1    mean
##       0.2259  3.0800
## s.e.  0.1262  0.0115
##
## sigma^2 = 0.00488:  log likelihood = 74.29
## AIC=-142.59   AICc=-142.15   BIC=-136.35
```

Después de realizar diversos análisis y pruebas, encontramos que el modelo que mejor se ajusta a nuestra serie es un modelo $ARIMA(1,0,0)$ con una media no nula.

Los modelos $ARIMA$ (Autoregressive Integrated Moving Average) son una clase de modelos ampliamente utilizados para el análisis y la predicción de series temporales. Estos modelos combinan componentes autoregresivos (AR), de medias móviles (MA) y una integración (I) de la serie para capturar tanto la dependencia temporal como las tendencias presentes en los datos. (Para un análisis más detallado, se recomienda consultar la sección 2.2).

En cuanto a la expresión del modelo $ARIMA(1,0,0)$, se representa de la siguiente manera:

$$\Delta y_t = \alpha + \beta_1 y_{t-1} + \varepsilon_t$$

Donde:

- Δy_t representa la primera diferencia de la serie, que es la serie original menos la serie desfasada en un período.
- α es la constante del modelo.
- β_1 es el coeficiente autoregresivo de orden 1, que mide la correlación entre la serie en el tiempo anterior y_{t-1} y el cambio en la serie actual Δy_t
- ε_t es el término de error

Una vez que hemos ajustado nuestra serie al modelo, procedemos a realizar las predicciones para la siguiente campaña. Sin embargo, es importante recordar que debemos deshacer los cambios realizados previamente, ya que a nuestra serie se le aplicó el logaritmo neperiano.

Para deshacer el logaritmo neperiano y obtener las predicciones en la escala original, utilizamos la función exponencial. Esta función nos permite revertir la transformación logarítmica y obtener los valores predichos en su forma original.

```
## Predición de resultados
pred1=predict(ajuste,n.ahead=3)
pred1$pred
```

```
## Time Series:
## Start = c(20, 3)
## End = c(21, 2)
## Frequency = 3
## [1] 3.071446 3.078068 3.079564
```

```
## Deshacemos las transformaciones
plot(serie, xlim = c(1,30))
lines(exp(pred1$pred), col="red")
```

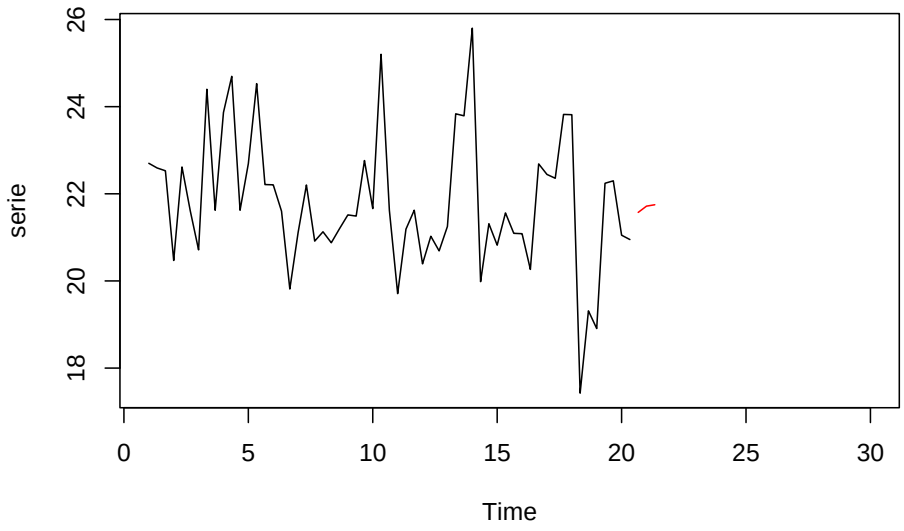


Figura 4.3: Serie de rendimiento con la predicción (fuente: elaboración propia)

```
exp(pred1$pred)
```

```
## Time Series:
## Start = c(20, 3)
## End = c(21, 2)
## Frequency = 3
## [1] 21.57308 21.71640 21.74891
```

Las predicciones obtenidas para la campaña 2022/23 son las siguientes:

- Predicción para diciembre 2022: 21.57308
- Predicción para enero 2023: 21.71640
- Predicción para febrero 2023: 21.74891

Estas predicciones representan las estimaciones para los meses específicos de la campaña 2022/23. Es importante tener en cuenta que los valores están en la escala original de la serie, ya que se ha deshecho la transformación logarítmica previa.

4.1.2. Regresión Lineal y Análisis de Componentes Principales

En esta sección, se abordará el proceso de modelado predictivo utilizando regresión lineal y la aplicación de la técnica de Análisis de Componentes Principales (PCA).

La regresión lineal se implementará utilizando el paquete parsnip de tidymodels, el cual proporciona una interfaz común para varios motores de modelos, permitiendo una fácil adaptación a diferentes tipos de modelos según el motor seleccionado.

En primer lugar, es importante resaltar la relevancia del preprocesamiento de los datos en el análisis predictivo. Los conjuntos de datos a menudo contienen una gran cantidad de variables, lo que puede generar problemas de dimensionalidad y complejidad en los modelos. Para abordar este desafío, se emplea la técnica de Análisis de Componentes Principales (PCA), que permite reducir la dimensionalidad del conjunto de datos mientras se conserva la mayor cantidad posible de información relevante.

En nuestro caso, contamos con un conjunto de datos que presenta una dimensionalidad inicial de 36 variables. Mediante el uso de PCA, aplicaremos una transformación en los datos para generar un nuevo conjunto de variables, conocidas como componentes principales.

El paquete recipes de tidymodels nos proporciona las herramientas necesarias para llevar a cabo el proceso de ingeniería de características. Utilizaremos las funciones incluidas en este paquete para estandarizar y normalizar las variables, así como para realizar la reducción de dimensionalidad mediante PCA. En este caso, configuraremos PCA para que mantenga un total de 5 componentes principales, lo que nos permitirá representar de manera más compacta la variabilidad presente en los datos originales.

A continuación, se presentará el código correspondiente al preprocesamiento de datos. Este incluirá pasos como centrar y escalar las variables, realizar la selección de variables (step_zv), y finalmente aplicar el PCA utilizando la función step_pca con un número determinado de componentes principales (num_comp = 5). Se destacará que el paso de PCA se realiza después de haber completado los demás pasos de preprocesamiento.

```
library(tidymodels)
datos_prediccion_rendimiento = data.frame(datos_prediccion,rendimiento)
datos_prediccion_rendimiento = datos_prediccion_rendimiento %>%
  mutate_all(as.numeric)
## Eliminamos las columnas referidas a los años
datos_prediccion_rendimiento = datos_prediccion_rendimiento[,-c(1,26)]
attach(datos_prediccion_rendimiento)

set.seed(123)
datos_prediccion_rendimiento_split <- datos_prediccion_rendimiento %>%
  initial_split(prop = 0.75)


## Creamos la receta
pca_rec <- training(datos_prediccion_rendimiento_split) %>%
  recipe(rendimiento ~ .,) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_pca(all_predictors(), num_comp = 5)
```

Una vez completado el proceso de preprocesamiento de datos, llegamos al paso de implementar el modelo de regresión lineal. Para llevar a cabo esto, integraremos la receta de preprocesamiento y el modelo en un flujo de trabajo coherente utilizando el paquete parsnip de tidymodels.

El flujo de trabajo, también conocido como workflow, nos permite organizar y estructurar todas las etapas del análisis predictivo de manera ordenada y reproducible. En este caso, combinaremos la receta de preprocesamiento que hemos creado previamente con el modelo de regresión lineal.

Para evaluar el rendimiento del modelo de manera robusta, utilizaremos la técnica de validación cruzada k-fold. La función vfold_cv se encargará de generar las particiones del conjunto de datos en k pliegues (folds), donde k tomará el valor 10. Cada partición se utilizará como conjunto de entrenamiento y prueba de manera rotativa, lo que nos permitirá evaluar el modelo en diferentes configuraciones de datos.

La función fit_resamples es la responsable de entrenar, predecir y evaluar el modelo en cada partición generada por vfold_cv. Esto se realiza de forma automática y eficiente, obteniendo métricas de evaluación para cada configuración. El resultado final es un objeto que contiene las métricas recopiladas de todas las particiones.

A continuación, se proporcionará el código correspondiente a esta sección, donde se implementa el flujo de trabajo completo. Esto incluirá la integración de la receta de preprocesamiento, el modelo de regresión lineal, la generación de particiones con vfold_cv y el cálculo de métricas con fit_resamples.

```
## Preparamos la receta
preps <- pca_rec %>%
  prep()
datos_procesados= preps %>%
  bake(new_data=datos_prediccion_rendimiento_split)


# Partiendo del modelo a aplicar
lm_spec <-
  linear_reg() %>%
  set_engine(engine = "lm") %>%
  set_mode(mode = "regression")

# y del flujo de trabajo establecido
all_wf <-
  workflow() %>%
  add_recipe(pca_rec) %>%
  add_model(lm_spec)

## Instrucción para realizar la validación cruzada
cv_folds <-
  vfold_cv(training(datos_prediccion_rendimiento_split), v = 10)


 all_wf1 = all_wf %>%
  fit_resamples(resamples = cv_folds,
                metrics = metric_set(rmse)) %>%
  collect_metrics()
```

```
all_wf1
```

```
## # A tibble: 1 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard    1.50    10   0.200 Preprocessor1_Model1
```

Finalmente, nos adentraremos en el análisis de las métricas de evaluación para comprender el desempeño del modelo de regresión lineal. Utilizaremos la función collect_metrics para recopilar y calcular diversas métricas de evaluación, como el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Estas métricas nos proporcionarán información valiosa sobre la capacidad predictiva del modelo y su ajuste a los datos.

El código correspondiente a esta sección incluirá la implementación de la función collect_metrics, que tomará como entrada el objeto generado por fit_resamples. A través de este proceso, obtendremos las métricas promediadas y consolidadas, lo que nos permitirá realizar comparaciones más robustas entre diferentes configuraciones del modelo.

```
rmse_pca = all_wf1$mean

last_fit_wf <-  all_wf %>%
  last_fit(split = datos_prediccion_rendimiento_split)

last_fit_wf %>% collect_predictions()
```

```
## # A tibble: 6 x 5
##   id               .pred  .row rendimiento .config
##   <chr>            <dbl> <int>       <dbl> <chr>
## 1 train/test split  21.3     1        22.5 Preprocessor1_Model1
## 2 train/test split  20.8     7        21.5 Preprocessor1_Model1
## 3 train/test split  20.9    12        20.8 Preprocessor1_Model1
## 4 train/test split  22.6    16        21.3 Preprocessor1_Model1
## 5 train/test split  21.4    17        21.5 Preprocessor1_Model1
## 6 train/test split  22.2    20        21.5 Preprocessor1_Model1
```

```
last_fit_wf %>% collect_metrics()
```

```
## # A tibble: 2 x 4
##   .metric .estimator  .estimate .config
##   <chr>   <chr>           <dbl> <chr>
## 1 rmse    standard   0.822      Preprocessor1_Model1
## 2 rsq     standard   0.00000542 Preprocessor1_Model1
```

Al analizar los resultados de nuestro modelo predictivo, observamos que el índice de error cuadrático medio (RMSE) promedio obtenido es de aproximadamente 1.5. El RMSE es una medida de la diferencia entre los valores predichos por el modelo y los valores reales

del rendimiento. En nuestro caso, hemos logrado obtener un RMSE del modelo de 0.82, lo que indica que nuestro modelo tiene un buen desempeño en la tarea de predicción.

Al examinar las predicciones generadas por el modelo, nos centramos en la columna .pred(), que representa las predicciones realizadas por nuestro modelo, y la columna "rendimiento", que contiene los valores reales del rendimiento. Al comparar ambas columnas, podemos observar que las predicciones generadas por nuestro modelo son bastante cercanas y se ajustan adecuadamente a los valores reales de rendimiento. Esto indica que nuestro modelo ha logrado capturar y comprender las relaciones subyacentes entre las variables predictoras y el rendimiento, lo que resulta en predicciones aceptables y confiables.

4.1.3. Random Forest

En esta sección de nuestro trabajo, vamos a utilizar el modelo de Random Forest para realizar predicciones del rendimiento. En este caso, hemos seleccionado "ranger" como motor de cálculo para nuestro modelo de Random Forest. Ranger es una implementación eficiente y escalable del algoritmo de Random Forest que ofrece un rendimiento óptimo para conjuntos de datos grandes y complejos. Además, hemos establecido el modo de nuestro modelo como "regresión", ya que estamos realizando una tarea de predicción de un valor numérico continuo, que es el rendimiento en nuestro caso. A continuación se presenta el modelo generado y se discuten los resultados.

```
modelo_randomforest = rand_forest() %>%
  set_engine("ranger") %>%
  set_mode("regression")

wflow_rf = workflow() %>%
              add_model(modelo_randomforest) %>%
              add_recipe(pca_rec_rf)

wflow_rf1 = wflow_rf %>%
  fit_resamples(resamples = cv_folds,
                metrics = metric_set(rmse)) %>%
  collect_metrics()

wflow_rf1
```

```
## # A tibble: 1 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard    1.56    10   0.290 Preprocessor1_Model1
```

```
rmse_rf =  wflow_rf1$mean

predicciones_rf = wflow_rf %>%
              last_fit(split = datos_prediccion_rendimiento_split) %>%
```

```
              collect_predictions()
predicciones_rf
```

```
## # A tibble: 6 x 5
##   id                .pred  .row rendimiento .config
##   <chr>             <dbl> <int>       <dbl> <chr>
## 1 train/test split  21.6     1        22.5 Preprocessor1_Model1
## 2 train/test split  21.8     7        21.5 Preprocessor1_Model1
## 3 train/test split  21.6    12        20.8 Preprocessor1_Model1
## 4 train/test split  22.1    16        21.3 Preprocessor1_Model1
## 5 train/test split  21.8    17        21.5 Preprocessor1_Model1
## 6 train/test split  21.9    20        21.5 Preprocessor1_Model1
```

```
last_fit_wf_rf =wflow_rf %>%
  last_fit(split = datos_prediccion_rendimiento_split) %>%
  collect_metrics()
last_fit_wf_rf
```

```
## # A tibble: 2 x 4
##   .metric .estimator .estimate .config
##   <chr>   <chr>          <dbl> <chr>
## 1 rmse    standard     0.649   Preprocessor1_Model1
## 2 rsq     standard     0.00710 Preprocessor1_Model1
```

El RMSE promedio obtenido para nuestras predicciones fue de 1.57, lo cual indica que, en promedio, nuestras predicciones difieren del rendimiento real en 1.57 unidades. Sin embargo, al observar el RMSE del modelo en sí, encontramos que este valor es de 0.649. Un RMSE más bajo en el modelo indica una mejor capacidad para ajustarse a los datos y realizar predicciones precisas.

Al analizar las predicciones generadas por el modelo, nos percatamos de que las predicciones se encuentran en línea con los valores reales de rendimiento. En general, podemos concluir que las predicciones generadas por nuestro modelo de Random Forest son bastante buenas. El hecho de que el RMSE del modelo sea significativamente menor que el RMSE promedio sugiere que el modelo ha logrado capturar y modelar de manera efectiva los patrones y la variabilidad presentes en los datos de rendimiento. Esto nos brinda confianza en la capacidad del modelo para realizar predicciones precisas y útiles en el contexto de nuestro estudio.

Es importante destacar que tanto este como la mayoría de los modelos presentan un coeficiente de determinación (R^2) relativamente bajo, aún así, continuamos con la comparación y mejora de los modelos utilizando otras métricas de evaluación, en este caso, el RMSE.

4.1.4. Algoritmo KNN

En esta sección de nuestro trabajo, exploraremos la utilización del algoritmo de vecinos más cercanos (KNN) para la predicción del rendimiento. Para ello, implementaremos tres modelos distintos utilizando diferentes valores de “K” (número de vecinos).

El primer modelo se construirá utilizando un único vecino cercano. Esto significa que, para realizar una predicción, el modelo considerará únicamente la información del vecino más cercano al punto de interés en el espacio de características. Este enfoque puede ser útil cuando se busca una predicción basada en patrones muy específicos y cercanos en los datos.

El segundo modelo se basará en tres vecinos cercanos. Al considerar la información de múltiples vecinos, el modelo podrá capturar patrones más amplios y tener en cuenta una mayor variedad de casos similares al momento de realizar las predicciones. Esta estrategia puede proporcionar un equilibrio entre precisión y generalización.

Por último, construiremos un tercer modelo utilizando cinco vecinos cercanos. Al aumentar el número de vecinos, el modelo se vuelve más robusto frente a valores atípicos o ruido en los datos. Esto puede conducir a predicciones más estables y confiables, especialmente en conjuntos de datos con mayor variabilidad o complejidad.

Cada uno de estos modelos de KNN nos permitirá evaluar el rendimiento del algoritmo en diferentes configuraciones, considerando distintos niveles de vecindad. Esto nos brindará una visión más completa de cómo el número de vecinos influye en la precisión y la capacidad predictiva del modelo.

4.1.4.1. Algoritmo KNN (K=1)

Abordaremos la implementación del método k-Nearest Neighbors (k-NN) con 1 vecino.

```
modelo_KNN1 = nearest_neighbor(neighbors = 1) %>%
  set_mode("regression") %>%
  set_engine("kknn")

wflow_knn1 = workflow() %>%
              add_model(modelo_KNN1) %>%
              add_recipe(pca_rec_knn1)

wflow_knn11 = wflow_knn1 %>%
  fit_resamples(resamples = cv_folds,
                metrics = metric_set(rmse)) %>%
  collect_metrics()

wflow_knn11
```

```
## # A tibble: 1 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard    1.88    10   0.356 Preprocessor1_Model1
```

En la ejecución del algoritmo KNN con un solo vecino, se realizó una evaluación exhaustiva utilizando métricas de rendimiento para medir la calidad del modelo. Los resultados presentados aquí corresponden a la validación cruzada.

El valor promedio de RMSE obtenido fue de 1.88, lo que indica que, en promedio, existe una diferencia de 1.88 unidades entre las predicciones generadas por el modelo y los valores reales de rendimiento en cada partición del conjunto de datos. Estos resultados de validación cruzada brindan una evaluación más completa y confiable del rendimiento del modelo, ya que se considera la variabilidad inherente de los datos y se promedian los resultados de múltiples particiones.

```
## # A tibble: 6 x 5
##   id               .pred  .row rendimiento .config
##   <chr>            <dbl> <int>       <dbl> <chr>
## 1 train/test split  23.1     1        22.5 Preprocessor1_Model1
## 2 train/test split  23.1     7        21.5 Preprocessor1_Model1
## 3 train/test split  21.2    12        20.8 Preprocessor1_Model1
## 4 train/test split  21.0    16        21.3 Preprocessor1_Model1
## 5 train/test split  19.4    17        21.5 Preprocessor1_Model1
## 6 train/test split  21.0    20        21.5 Preprocessor1_Model1

## # A tibble: 2 x 4
##   .metric .estimator .estimate .config
##   <chr>   <chr>          <dbl> <chr>
## 1 rmse    standard       1.15  Preprocessor1_Model1
## 2 rsq     standard       0.247 Preprocessor1_Model1
```

El resultado del RMSE que se presenta se refiere a la evaluación realizada sobre el conjunto de prueba. Es importante tener en cuenta que el resultado puede variar considerablemente dependiendo de la muestra específica utilizada en esta evaluación. Esto se debe a que el conjunto de prueba es solo una muestra de los datos disponibles, y diferentes muestras podrían dar lugar a resultados ligeramente diferentes. Se obtuvo un valor de 1.15, lo que significa que el modelo en sí tiene una capacidad para predecir el rendimiento con un error cuadrático medio de 1.15 unidades. Este valor indica que el modelo es capaz de realizar predicciones relativamente precisas en comparación con los valores reales de rendimiento.

Al examinar las predicciones generadas por el modelo, se observa que son aceptables. Esto implica que el modelo es capaz de capturar patrones relevantes y establecer relaciones significativas entre las variables predictoras y el rendimiento.

4.1.4.2. Algoritmo KNN (K=3)

Al aplicar el algoritmo KNN con tres vecinos, se llevó a cabo un análisis detallado de evaluación para examinar la efectividad del modelo en la predicción del rendimiento. Los resultados obtenidos a través de validación cruzada revelaron un valor promedio de RMSE de 1.72. Esto indica que, en promedio, existe una diferencia de 1.72 unidades entre las predicciones generadas por el modelo y los valores reales de rendimiento.

```
## # A tibble: 1 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard    1.72    10   0.279 Preprocessor1_Model1
```

```
## # A tibble: 6 x 5
##   id               .pred  .row rendimiento .config
##   <chr>            <dbl> <int>       <dbl> <chr>
## 1 train/test split  22.6     1        22.5 Preprocessor1_Model1
## 2 train/test split  23.1     7        21.5 Preprocessor1_Model1
## 3 train/test split  21.5    12        20.8 Preprocessor1_Model1
## 4 train/test split  21.9    16        21.3 Preprocessor1_Model1
## 5 train/test split  20.2    17        21.5 Preprocessor1_Model1
## 6 train/test split  21.2    20        21.5 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator .estimate .config
##   <chr>   <chr>          <dbl> <chr>
## 1 rmse    standard       0.911 Preprocessor1_Model1
## 2 rsq     standard       0.124 Preprocessor1_Model1
```

Además del RMSE promedio a través de validación cruzada, se calculó el RMSE del modelo referido a la evaluación realizada sobre el conjunto de prueba, el cual arrojó un valor de 0.911. Este resultado sugiere que el modelo, al considerar tres vecinos cercanos en la clasificación, tiene una capacidad para predecir el rendimiento con un error cuadrático medio de 0.911 unidades. Este valor relativamente bajo refuerza la precisión y la capacidad predictiva del modelo en comparación con los valores reales de rendimiento.

Al examinar las predicciones generadas por el modelo, se observó que son altamente aceptables en términos de su concordancia con los valores reales de rendimiento. Esto implica que el modelo, al considerar múltiples vecinos cercanos en la clasificación, es capaz de identificar y capturar patrones significativos y relaciones relevantes entre las variables predictoras y el rendimiento. Estos resultados respaldan la utilidad y el potencial del algoritmo KNN para abordar el problema de predicción del rendimiento en el contexto de estudio.

4.1.4.3. Algoritmo KNN (K=5)

Durante la aplicación del algoritmo KNN con una vecindad de cinco elementos más cercanos, se llevó a cabo un análisis minucioso para evaluar la capacidad predictiva del modelo en relación al rendimiento. Los resultados, a través de validación cruzada, obtenidos revelaron un valor promedio de RMSE de 1.63, lo cual indica que existe una discrepancia promedio de 1.63 unidades entre las predicciones generadas por el modelo y los valores reales de rendimiento.

```
## # A tibble: 1 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 rmse    standard    1.63    10   0.270 Preprocessor1_Model1
```

```
## # A tibble: 6 x 5
##   id               .pred  .row rendimiento .config
```

```
##   <chr>            <dbl> <int>       <dbl> <chr>
## 1 train/test split  22.2     1        22.5 Preprocessor1_Model1
## 2 train/test split  22.7     7        21.5 Preprocessor1_Model1
## 3 train/test split  21.4    12        20.8 Preprocessor1_Model1
## 4 train/test split  22.3    16        21.3 Preprocessor1_Model1
## 5 train/test split  20.6    17        21.5 Preprocessor1_Model1
## 6 train/test split  21.4    20        21.5 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator .estimate .config
##   <chr>   <chr>          <dbl> <chr>
## 1 rmse    standard      0.790  Preprocessor1_Model1
## 2 rsq     standard      0.0735 Preprocessor1_Model1
```

Además, al calcular el RMSE referido a la evaluación realizada sobre el conjunto de prueba, se obtuvo un valor de 0.79, lo que sugiere que el modelo tiene una capacidad aceptable para predecir el rendimiento con un error cuadrático medio de 0.79 unidades al considerar cinco vecinos cercanos en la clasificación. Este resultado resalta la precisión y efectividad del modelo en términos de su capacidad para aproximar los valores reales de rendimiento.

4.1.4.4. Comparación de modelos

En esta sección, realizaremos una comparación exhaustiva de los resultados obtenidos por los modelos KNN estudiados. Para ello, nos centraremos en la evaluación del error cuadrático medio (RMSE) utilizando dos enfoques: validación cruzada y una muestra específica seleccionada.

Tabla 4.1: Tabla comparativa del RMSE para los modelos KNN basados en el rendimiento

RMSE_CV	RMSE_Muestra	Modelos
1.879839	1.1461596	KNN_1
1.719169	0.9111748	KNN_3
1.626450	0.7901662	KNN_5

Al analizar detalladamente la tabla de resultados, se observa una tendencia clara: a medida que incrementamos el número de vecinos en el algoritmo KNN, se produce una mejora tanto en el RMSE obtenido mediante validación cruzada como en el RMSE obtenido a través de la muestra específica seleccionada.

Este patrón sugiere que al considerar un mayor número de vecinos en el proceso de regresión, el modelo tiene en cuenta una mayor cantidad de puntos de datos cercanos para realizar predicciones. Esta mayor cantidad de información contribuye a una mejor estimación y, por lo tanto, se obtiene un menor valor de RMSE.

En particular, al comparar los resultados obtenidos con diferentes números de vecinos, se observa que el modelo con 5 vecinos muestra el menor valor de RMSE tanto en la

validación cruzada como en la muestra elegida. Esto indica que este modelo tiene un desempeño superior en términos de precisión y capacidad para predecir el rendimiento de manera más acertada.

4.1.5. Red Neuronal

En esta sección, nos centraremos en la aplicación de modelos de redes neuronales para predecir el rendimiento en la producción oleícola. En el contexto de este estudio, se han llevado a cabo pruebas de modelado de redes neuronales sin realizar asunciones a priori sobre el número de capas ocultas o el número de neuronas en cada capa. Se ha realizado una exploración exhaustiva de diferentes arquitecturas de redes neuronales, con el objetivo de encontrar la configuración óptima que maximice el desempeño del modelo.

En primer lugar, diseñaremos una red neuronal con una única capa oculta compuesta por 20 nodos. Cada nodo de esta capa se conectará con los nodos de entrada, que representan las variables predictoras relevantes para nuestra predicción del rendimiento. El algoritmo que se utiliza es "rprop+", que es una variante del algoritmo de retropropagación con actualización de pesos adaptativa y los pesos iniciales de la red neuronal al darle la opción "NULL" significa que la red se inicializará de forma aleatoria.

```
library(neuralnet)

datos_prediccion_rendimiento = data.frame(datos_prediccion,rendimiento)
datos_prediccion_rendimiento = datos_prediccion_rendimiento %>%
  mutate_all(as.numeric)
datos_prediccion_rendimiento = datos_prediccion_rendimiento[,-c(1,26)]
attach(datos_prediccion_rendimiento)
set.seed(123)
datos_prediccion_rendimiento_split <- datos_prediccion_rendimiento %>%
  initial_split(prop = 0.75)


train_nrm <- training(datos_prediccion_rendimiento_split)
test_nrm <- testing(datos_prediccion_rendimiento_split)

pca_rec_nn <- train_nrm %>%
  recipe(rendimiento ~ .) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_predictors())

train_nrm =  pca_rec_nn %>%
  prep() %>%
  bake(new_data = NULL)

test_nrm =  pca_rec_nn %>%
  prep() %>%
  bake(new_data = test_nrm)
```

```
# FORMULA
# ------------------------------------------------------
nms  <- names(train_nrm)
frml <- as.formula(paste("rendimiento ~",
                    paste(nms[!nms %in% "rendimiento"], collapse = " + ")))


# MODELO
# ------------------------------------------------------
modelo.nn1 <- neuralnet(frml,
                        data          = train_nrm,
                        hidden        = c(20),
                        threshold     = 0.01,
                        algorithm     = "rprop+",
                        startweights = NULL
)


# PREDICCION
# ------------------------------------------------------
pr.nn1   <- compute(modelo.nn1,within(test_nrm,rm(rendimiento)))

pred1 = pr.nn1$net.result

comparacion = as.data.frame(cbind(pred1,test_nrm$rendimiento))
colnames(comparacion) = c("Predicción","Datos Reales")
comparacion
```

```
##   Predicción Datos Reales
## 1   13.59820     22.53933
## 2   23.15317     21.50002
## 3   10.46461     20.78898
## 4   20.72721     21.33736
## 5   16.71400     21.49914
## 6   15.85027     21.46424
```

```
## Error cuadratico medio

rmse_rn1 = sqrt((1/6)*sum((pr.nn1$net.result - test_nrm$rendimiento)^2))
rmse_rn1
```

```
## [1] 6.377772
```

En segundo lugar, se empleará una red neuronal con dos capas ocultas en el modelo de predicción. La arquitectura de la red constará de una primera capa oculta con 16 nodos y una segunda capa oculta con 8 nodos. La elección de estas capas y su tamaño se basa en

consideraciones teóricas y empíricas, buscando encontrar un equilibrio entre la capacidad de representación del modelo y la complejidad del mismo.

Al emplear esta configuración, se realizará el ajuste del modelo a los datos de entrenamiento y se buscará encontrar los pesos óptimos que minimicen el error de predicción. La red neuronal aprenderá a partir de los patrones presentes en los datos y generará predicciones basadas en la estructura y conexiones de la red.

```
##   Predicción Datos Reales
## 1   20.76822     22.53933
## 2   23.05044     21.50002
## 3   20.45712     20.78898
## 4   24.33934     21.33736
## 5   20.40780     21.49914
## 6   23.29352     21.46424
```

```
## [1] 1.78885
```

Al comparar los resultados obtenidos, podemos observar que la red neuronal de dos capas ocultas, con 16 y 8 nodos respectivamente, ha logrado obtener un RMSE de 1.7888, mientras que la red neuronal de una sola capa oculta ha obtenido un RMSE de 6.377.

Esto indica que la red neuronal de dos capas ha logrado reducir significativamente el error en las predicciones en comparación con la red neuronal de una capa. El RMSE más bajo obtenido por la red neuronal de dos capas sugiere que este modelo es más preciso y se ajusta mejor a los datos de entrenamiento.

La introducción de capas adicionales en la red neuronal permite que el modelo capture relaciones más complejas y no lineales entre las variables predictoras y la variable objetivo. Esto puede conducir a una mayor capacidad de generalización y a una mejora en la precisión de las predicciones.

Es importante destacar que en el modelado de redes neuronales, la elección de la arquitectura y los hiperparámetros es un proceso altamente exploratorio y puede involucrar muchas más combinaciones de parámetros de las que se han implementado en este estudio. Los resultados obtenidos con las arquitecturas de una y dos capas ocultas son solo dos ejemplos de posibles configuraciones.

Existen numerosos factores a considerar al diseñar una red neuronal, como el número de capas ocultas, la cantidad de nodos en cada capa, la función de activación utilizada, la tasa de aprendizaje, el número de épocas de entrenamiento y la técnica de regularización aplicada, entre otros. Cada combinación de parámetros puede tener un impacto significativo en el rendimiento y la precisión del modelo.

4.1.6. Comparacion de los modelos predictivos empleados

En esta sección, evaluamos y comparamos el rendimiento de diferentes modelos predictivos utilizados para predecir el valor del rendimiento. Para realizar una comparación más confiable y robusta, nos basamos en el RMSE promedio obtenido a través de validación cruzada. Esta medida nos brinda una perspectiva más sólida debido al tamaño de nuestro

conjunto de datos y a la sensibilidad que puede presentar dependiendo de la partición de los datos utilizada. A continuación, presentamos la tabla con los valores de RMSE correspondientes a cada modelo para su análisis y comparación.

Tabla 4.2: Tabla comparativa del RMSE de los modelos basados en el rendimiento

RMSE	Modelos
1.500199	LR con PCA
1.558210	RF
1.879839	KNN_1
1.719169	KNN_3
1.626450	KNN_5
1.788850	RN

Basándonos en los valores de RMSE promedio de los modelos evaluados, el modelo de análisis de componentes principales muestra el menor valor de RMSE promedio, lo cual indica que tiene un mejor rendimiento en la predicción del valor de rendimiento en comparación con los otros modelos. Esto significa que el modelo tiene una discrepancia promedio de 1.5002 unidades entre las predicciones generadas y los valores reales de rendimiento.

Por lo tanto, con base en el análisis realizado, se recomienda seleccionar el modelo de análisis de componentes principales como el modelo preferido para predecir el rendimiento debido a su menor RMSE promedio. Sin embargo, es importante considerar que el modelo basado en random forest o KNN con 5 vecinos tienen un RMSE muy cercano y que podrían ser otra alternativa para predecir el rendimiento.

4.2. Desarrollo de modelos predictivos para estimar la producción en kilos de aceituna

En esta sección, nos adentraremos en la aplicación de técnicas de predicción para estimar los kilogramos de aceituna producidos por campaña. A diferencia de la sección anterior, donde nos enfocamos en predecir el rendimiento, ahora dirigiremos nuestro análisis hacia la variable objetivo de los kilogramos totales de aceituna cosechados.

Para llevar a cabo este proceso, utilizaremos las mismas técnicas de preprocesamiento de datos que hemos aplicado previamente. Sin embargo, es importante destacar que, en esta ocasión, llevaremos a cabo una etapa adicional para eliminar los datos atípicos (outliers) que podrían influir negativamente en la precisión y confiabilidad de nuestras predicciones.

La eliminación de los outliers nos permitirá obtener un conjunto de datos más limpio y representativo, evitando así posibles distorsiones en nuestros modelos predictivos. Esto es esencial, ya que los outliers pueden generar una variabilidad excesiva en los datos y afectar significativamente la calidad de las predicciones.

4.2.1. Serie Temporal

En esta sección, nos centraremos en predecir los kilogramos de aceituna producidos utilizando un enfoque basado en series temporales, similar al utilizado para predecir el rendimiento.

Para llevar a cabo esta tarea, contamos con un conjunto de datos que consiste en los kilogramos de aceituna producidos en los meses de diciembre, enero y febrero de cada campaña desde el año 2001. Sin embargo, al igual que ocurría con la variable de rendimiento, los meses de febrero de los años 2005 y 2006 no disponen de información registrada.

Para abordar esta falta de información, optaremos por realizar una imputación utilizando la media de los meses de febrero que sí contienen datos. Esto nos permitirá mantener una frecuencia constante en la serie, con tres puntos de datos por campaña para los meses mencionados. Esta elección se basó en la comprobación de que no existe una clara tendencia marcada al alza ni a la baja en los datos. En situaciones donde se observa una tendencia clara, puede ser adecuado considerar la imputación utilizando modelos de regresión u otras estrategias.

En primer lugar, realizaremos una representación gráfica de la serie temporal obtenida para los meses de diciembre, enero y febrero.

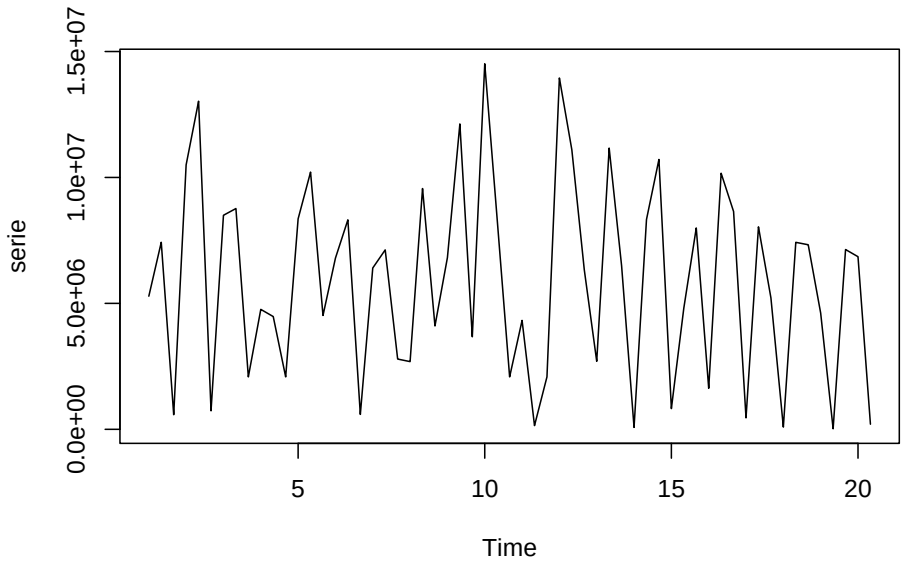


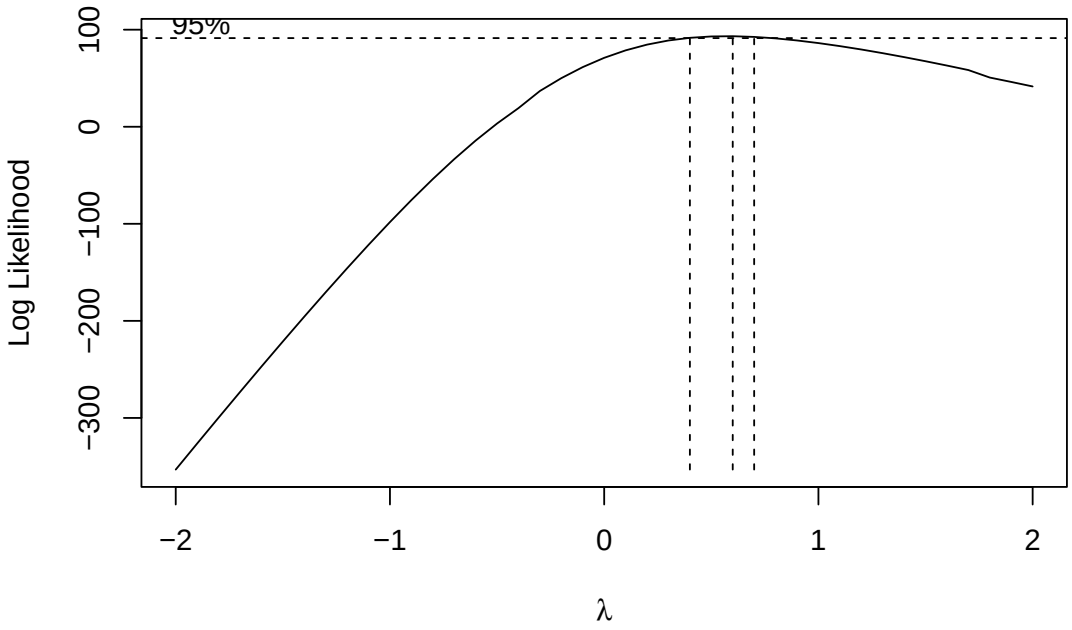
Figura 4.4: Serie de kilogramos producidos (fuente: elaboración propia)

Al analizar la gráfica de la serie de kilogramos de aceituna producidos, podemos notar que no muestra estabilidad en cuanto a la varianza a lo largo del tiempo. La presencia de variaciones significativas en la varianza puede dificultar el análisis y la predicción precisa de la serie.

Para abordar este problema y lograr la estabilización de la serie, seguiremos el mismo enfoque que utilizamos en la sección anterior para predecir el rendimiento. Aplicaremos

las transformaciones Box-Cox, una técnica comúnmente utilizada en el análisis de series temporales para estabilizar la varianza.

```
## Transformaciones para que la varianza sea estable con el tiempo
library(TSA)
bc=BoxCox.ar(y=serie)
```



Al aplicar el logaritmo neperiano como parte de la transformación de Box-Cox, buscamos reducir las diferencias en la escala y estabilizar la varianza de la serie. Esta transformación logarítmica es útil cuando se observan cambios proporcionales en los datos y se espera que la variación relativa sea más constante en el tiempo.

```
serie_transformada = log(serie)
```

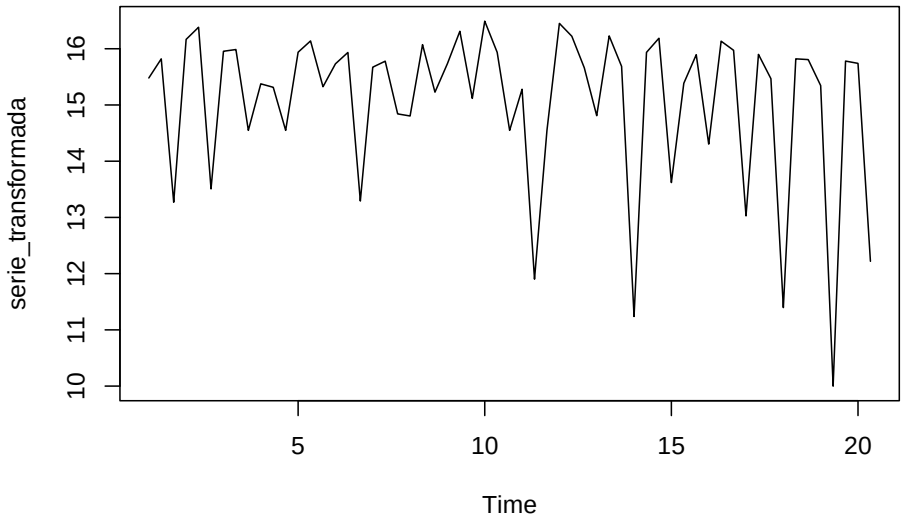


Figura 4.5: Serie de kilogramos producidos transformada (fuente: elaboración propia)

Después de aplicar la transformación de Box-Cox a la serie de kilogramos de aceituna producidos, podemos observar que la serie se vuelve estable tanto en términos de varianza como de media. Esta estabilización es un paso crucial para garantizar la confiabilidad de los análisis y las predicciones.

Si examinamos la serie transformada, notaremos que la variación en la escala se ha reducido y la varianza se ha vuelto más constante a lo largo del tiempo. Esto indica que hemos logrado estabilizar la serie en términos de varianza.

Además, si comparamos las medias en diferentes períodos de tiempo, veremos que la serie también muestra estabilidad en cuanto a la media. Esto significa que la media de la serie no varía de manera significativa a lo largo del tiempo después de la transformación.

Una vez que hemos obtenido una serie transformada estable en términos de varianza y media, procedemos a realizar el test de Dickey-Fuller para comprobar si la serie es estacionaria.

```
### Contrastar estacionariedad
library(tseries)
adf.test(serie_transformada)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  serie_transformada
## Dickey-Fuller = -3.8941, Lag order = 3, p-value = 0.0204
## alternative hypothesis: stationary
```

Podemos concluir que hay evidencia estadística suficiente para rechazar la hipótesis nula de no estacionaridad de la serie. Esto sugiere que la serie es estacionaria, lo que significa que no presenta tendencias significativas o patrones no estacionarios en el tiempo.

Una vez confirmada la estacionaridad de la serie transformada de kilogramos de aceituna producidos, podemos proceder a ajustar un modelo a la misma. El objetivo de ajustar un modelo es capturar los patrones y estructuras presentes en la serie para poder realizar predicciones futuras con base en estos.

```
library(forecast)
ajuste=auto.arima(serie_transformada)
ajuste
```

```
## Series: serie_transformada
## ARIMA(0,0,0)(0,0,1)[3] with non-zero mean
##
## Coefficients:
##         sma1     mean
##       0.3850  15.0653
## s.e.  0.1101   0.2309
##
## sigma^2 = 1.746:  log likelihood = -99.38
## AIC=204.75   AICc=205.19   BIC=210.98
```

Tras ajustar el modelo, obtenemos un modelo $ARIMA(0,0,0)(0,0,1)[3]$. Este utiliza una componente de media móvil estacional de orden 1. Para entenderlo mejor, desglosemos su estructura y sus implicaciones:

$ARIMA(0,0,0)$ implica que no hay componente autoregresivo (AR) ni componente de media móvil (MA) en el modelo. El orden (0,0,0) indica que no se utiliza ninguna observación pasada o ningún error pasado para predecir el valor actual de la serie. En otras palabras, el valor de la serie en un momento dado no depende de los valores anteriores ni de los errores pasados.

El componente (0,0,1) se refiere al componente de media móvil estacional (SMA). El orden (0,0,1) indica que se utiliza un único término de media móvil estacional en el modelo. Este término tiene en cuenta la dependencia lineal de la serie con los errores residuales en períodos de tiempo anteriores, pero solo a una escala estacional con una periodicidad de 3 (indicado por [3]).

En términos prácticos, esto significa que el modelo utiliza solo la información de la observación actual y el error estacional pasado para predecir el valor actual de la serie. La presencia de un término de media móvil estacional indica que el modelo captura algún patrón cíclico o estacional en los datos.

La expresión del modelo $ARIMA(0,0,0)(0,0,1)[3]$ es la siguiente:

$$y_t = 15.0653 + 0.3850 * \varepsilon_{t-3}$$

donde:

- y_t representa el valor de la serie en el tiempo t, es decir, el número de kilogramos de aceituna producidos en ese período específico.
- ε_{t-3} es el error estacional en el tiempo t_3. Representa la diferencia entre el valor real de la serie en el tiempo t_3 y el valor predicho por el modelo en ese momento. El modelo utiliza este error estacional pasado para ajustar la predicción de la serie en el tiempo t.

Una vez hemos ajustado nuestro modelo a la serie, procedemos a realizar las predicciones para la próxima campaña. Sin embargo, es importante tener en cuenta que debemos deshacer los cambios previos aplicados, ya que la serie fue transformada mediante el logaritmo neperiano. Para ello, utilizamos la función exponencial.

```
## Predición de resultados
pred1=predict(ajuste,n.ahead=3)
pred1$pred
```

```
## Time Series:
## Start = c(20, 3)
## End = c(21, 2)
## Frequency = 3
## [1] 15.23769 15.11297 14.75445
```

```
## Deshacemos las transformaciones
plot(serie, xlim = c(1,30))
lines(exp(pred1$pred), col="red")
```

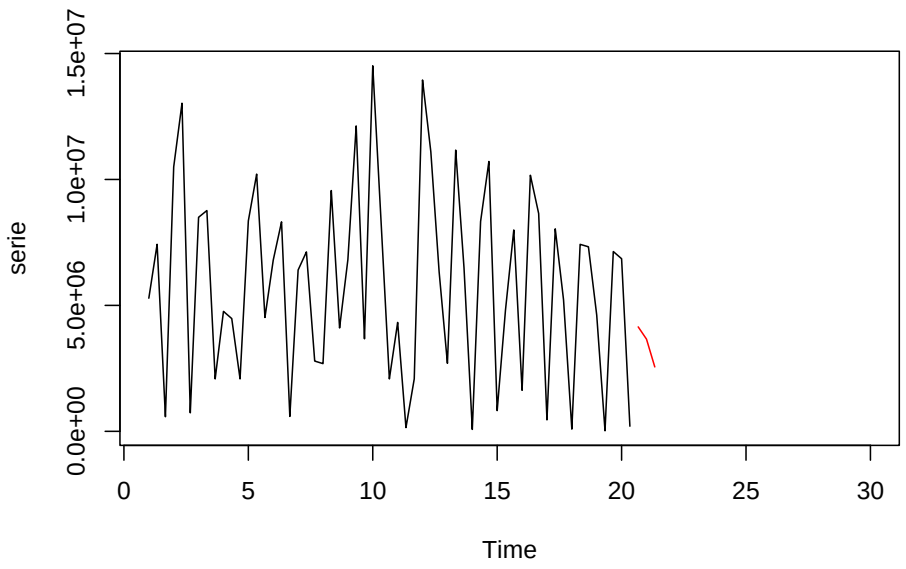


Figura 4.6: Serie de kilogramos producidos con la predicción (fuente: elaboración propia)

```
exp(pred1$pred)
```

```
## Time Series:
## Start = c(20, 3)
## End = c(21, 2)
## Frequency = 3
## [1] 4146145 3660001 2557268
```

Las predicciones para la próxima campaña son las siguientes:

- 4.146.145 kilogramos de aceituna producidos para el mes de diciembre.
- 3.660.001 kilogramos de aceituna producidos para el mes de enero.
- 2.557.268 kilogramos de aceituna producidos para el mes de febrero.

Estos valores representan las estimaciones de la cantidad de kilogramos de aceituna que se producirán en cada uno de los meses mencionados, basadas en el modelo ajustado a la serie temporal transformada y después de revertir la transformación logarítmica aplicada previamente.

4.2.2. Regresión Lineal y Análisis de Componentes Principales

En esta sección, siguiendo la metodología aplicada previamente para predecir el rendimiento, se realizará un modelo de regresión lineal múltiple utilizando todas las variables disponibles en nuestro conjunto de datos. Este enfoque nos permitirá explorar las posibles relaciones y la influencia conjunta de las diferentes variables en la variable objetivo y emplearemos el análisis de componentes principales (PCA) como técnica para obtener una representación más compacta y significativa de nuestros datos.

En este caso, al igual que hicimos para predecir el rendimiento, aplicaremos el análisis de componentes principales utilizando 5 componentes principales. Estos componentes serán seleccionados en función de su capacidad para explicar la variabilidad de los datos relacionada con los kilogramos totales de aceituna producidos. Al utilizar únicamente las componentes principales más relevantes, reduciremos la complejidad del modelo y evitaremos la inclusión de información redundante o poco relevante.

```
## # A tibble: 1 x 6
##   .metric .estimator     mean     n std_err .config
##   <chr>   <chr>         <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   3669865.    10 897765. Preprocessor1_Model1
```

Al evaluar el modelo predictivo de regresión lineal, apoyado en el análisis de componentes principales, para estimar los kilogramos totales de aceituna producidos por campaña, hemos obtenido resultados que merecen ser analizados en detalle. El RMSE promedio, a través de validación cruzada, se sitúa en 3.669.865 kilogramos. Esto indica que, en promedio, nuestras predicciones tienen una discrepancia de aproximadamente 3.669.865 kilogramos con respecto a los valores reales de producción.

```
## # A tibble: 4 x 5
##   id                    .pred  .row       kg .config
##   <chr>                 <dbl> <int>    <dbl> <chr>
## 1 train/test split 21338560.     8 22615423 Preprocessor1_Model1
## 2 train/test split 19542723.     9 22851016 Preprocessor1_Model1
## 3 train/test split 17607310.    11 19868255 Preprocessor1_Model1
## 4 train/test split 18201574.    13 19268546 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator   .estimate .config
##   <chr>   <chr>            <dbl> <chr>
## 1 rmse    standard   2169417.    Preprocessor1_Model1
## 2 rsq     standard         0.694 Preprocessor1_Model1
```

Por otro lado, al considerar el RMSE específico del modelo referido a la evaluación realizada sobre el conjunto de prueba, observamos que este valor se reduce a 2.169.417 kilogramos. Esto sugiere que el modelo es capaz de realizar predicciones más precisas y acertadas en comparación con el promedio de las predicciones. Un RMSE más bajo indica una menor discrepancia entre las predicciones y los valores reales, lo cual es un indicador positivo de la capacidad del modelo para estimar los kilogramos de aceituna producidos.

Además del RMSE, es importante evaluar el coeficiente de determinación (R cuadrado) del modelo. En este caso, se obtuvo un valor de 0.694, lo que implica que aproximadamente el 69.4 % de la variabilidad de los kilogramos totales de aceituna producidos puede ser explicada por el modelo.

4.2.3. Random Forest

En esta sección, abordaremos la tarea de predecir los kilogramos totales de aceituna producidos por campaña utilizando Random Forest. Al igual que en la sección 4.1.3, continuaremos utilizando el motor de cálculo "ranger" y configuraremos el modo del modelo como "regresión", ya que estamos tratando de predecir una variable numérica continua.

En este escenario, nuevamente tenemos nuestro conjunto de datos dividido en un conjunto de entrenamiento y un conjunto de prueba. Esta división se realiza con el objetivo de evaluar el rendimiento y la capacidad predictiva de nuestro modelo en datos no vistos previamente.

Además de la división en conjuntos de entrenamiento y prueba, también se emplea la técnica de validación cruzada para obtener una evaluación más precisa y robusta del rendimiento del modelo. A continuación, procederemos a mostrar y analizar los resultados.

```
## # A tibble: 1 x 6
##   .metric .estimator     mean     n std_err .config
##   <chr>   <chr>         <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   3178941.    10 760938. Preprocessor1_Model1
```

Al ejecutar el modelo Random Forest, hemos obtenido resultados interesantes que requieren un análisis detallado. El valor promedio del Error Cuadrático Medio (RMSE), obtenido a través de validación cruzada, es de 3.178.941, lo cual indica una discrepancia promedio de aproximadamente 3,178,941 kilogramos entre las predicciones generadas por el modelo y los valores reales de kilogramos totales de aceituna producidos por campaña.

```
## # A tibble: 4 x 5
##   id                    .pred  .row       kg .config
##   <chr>                 <dbl> <int>    <dbl> <chr>
## 1 train/test split 17870862.     8 22615423 Preprocessor1_Model1
## 2 train/test split 17884890.     9 22851016 Preprocessor1_Model1
## 3 train/test split 17889994.    11 19868255 Preprocessor1_Model1
## 4 train/test split 17255457.    13 19268546 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator   .estimate .config
##   <chr>   <chr>            <dbl> <chr>
## 1 rmse    standard   3625187.    Preprocessor1_Model1
## 2 rsq     standard         0.530 Preprocessor1_Model1
```

Asimismo, al calcular el RMSE específico del modelo referido a la evaluación realizada sobre el conjunto de prueba, hemos obtenido un valor de 3.625.187. Esto sugiere que, en promedio, las predicciones del modelo difieren en aproximadamente 3.625.187 kilogramos de los valores reales. Esta diferencia puede ser atribuida a la complejidad y variabilidad de los datos, así como a las limitaciones inherentes al modelo y su capacidad para capturar todas las relaciones y factores relevantes.

Al comparar estos resultados con los obtenidos en el modelo de regresión lineal múltiple mejorado con PCA, podemos observar que las predicciones del modelo Random Forest no son tan precisas en términos de su cercanía a los valores reales. El modelo PCA mostró un mejor desempeño en este aspecto, lo cual sugiere que el modelo Random Forest puede estar siendo afectado por ciertos desafíos o características particulares de los datos.

Adicionalmente, evaluamos el coeficiente de determinación (R cuadrado) del modelo, el cual es de 0.53. Este valor indica que aproximadamente el 53 % de la variabilidad en los kilogramos totales de aceituna producidos por campaña puede ser explicada por las variables predictoras utilizadas en el modelo. Si bien este valor no es tan alto como sería deseable, sugiere que el modelo está capturando parte de la variabilidad y es capaz de proporcionar información relevante sobre los kilogramos totales producidos.

4.2.4. Algoritmo KNN

En esta sección, continuaremos explorando diferentes técnicas de modelado para predecir los kilogramos totales de aceituna producidos por campaña. En esta ocasión, utilizaremos el algoritmo K-Nearest Neighbors (KNN) para construir tres modelos distintos, como hicimos para predicir el rendimiento, usando distintos valores para el parámetro K.

4.2.4.1. Algoritmo KNN (K=1)

Al analizar la ejecución del modelo KNN con 1 vecino, hemos obtenido resultados que nos permiten evaluar su rendimiento.

```
## # A tibble: 1 x 6
##   .metric .estimator     mean     n std_err .config
##   <chr>   <chr>         <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   3765709.    10 565981. Preprocessor1_Model1
```

El RMSE promedio obtenido mediante validación cruzada es de 3.765.709. Este valor nos indica que, en promedio, el modelo tiene un error de aproximadamente 3.765.709 kilogramos al predecir los kilogramos totales.

```
## # A tibble: 4 x 5
##   id                  .pred  .row       kg .config
##   <chr>               <dbl> <int>    <dbl> <chr>
## 1 train/test split 14193736     8 22615423 Preprocessor1_Model1
## 2 train/test split 23147497     9 22851016 Preprocessor1_Model1
## 3 train/test split 19388021    11 19868255 Preprocessor1_Model1
## 4 train/test split 17264291    13 19268546 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator    .estimate .config
##   <chr>   <chr>             <dbl> <chr>
## 1 rmse    standard   4337637.     Preprocessor1_Model1
## 2 rsq     standard         0.0178 Preprocessor1_Model1
```

Además del RMSE promedio, también hemos evaluado el RMSE específico del modelo referido a la evaluación realizada sobre el conjunto de prueba, que se ha calculado en 4.337.637. Este valor nos brinda una medida más precisa del rendimiento del modelo al considerar las diferencias individuales entre las predicciones y los valores reales. Un RMSE específico más alto indica que existen predicciones que difieren significativamente de los valores reales de kilogramos totales.

Al observar las predicciones generadas por el modelo KNN con 1 vecino, hemos encontrado que la mayoría de ellas son bastante acertadas y se acercan a los valores reales. Sin embargo, es importante destacar que ha habido una predicción que difiere en 8 millones de kilogramos.

En general, aunque el modelo KNN con 1 vecino muestra un RMSE promedio aceptable, la discrepancia observada en una de las predicciones resalta la necesidad de una evaluación más exhaustiva y de considerar otras configuraciones de K para mejorar la capacidad de predicción del modelo.

4.2.4.2. Algoritmo KNN (K=3)

Al analizar la ejecución del modelo KNN con 3 vecinos, hemos evaluado su rendimiento utilizando diferentes métricas.

```
## # A tibble: 1 x 6
##   .metric .estimator     mean     n std_err .config
##   <chr>   <chr>         <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   3487310.    10 555783. Preprocessor1_Model1
```

```
## # A tibble: 4 x 5
##   id                   .pred  .row       kg .config
##   <chr>                <dbl> <int>    <dbl> <chr>
## 1 train/test split 15387254.     8 22615423 Preprocessor1_Model1
## 2 train/test split 20696473.     9 22851016 Preprocessor1_Model1
## 3 train/test split 18523867.    11 19868255 Preprocessor1_Model1
## 4 train/test split 16086558.    13 19268546 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator    .estimate .config
##   <chr>   <chr>             <dbl> <chr>
## 1 rmse    standard   4147918.     Preprocessor1_Model1
## 2 rsq     standard         0.0755 Preprocessor1_Model1
```

El RMSE promedio obtenido mediante validación cruzada tiene como valor 3.487.310. Este valor indica que, en promedio, el modelo tiene un error de aproximadamente 3.487.310 kilogramos al predecir los kilogramos totales. Además del RMSE promedio, también hemos evaluado el RMSE específico del modelo referido a la evaluación realizada sobre el conjunto de prueba, que en este caso es de 4.147.918.

Al examinar las predicciones generadas por el modelo KNN con 3 vecinos, hemos observado que la mayoría de ellas son bastante acertadas y se acercan a los valores reales. Sin embargo, también hemos identificado una predicción que difiere en 7 millones de kilogramos.

4.2.4.3. Algoritmo KNN (K=5)

Al analizar el modelo KNN con 5 vecinos, hemos evaluado su desempeño utilizando las siguientes métricas. El valor promedio del Error Cuadrático Medio (RMSE) obtenido mediante validación cruzada ha sido de 3.295.049 kilogramos, lo cual indica una discrepancia promedio entre las predicciones del modelo y los valores reales.

```
## # A tibble: 1 x 6
##   .metric .estimator     mean     n std_err .config
##   <chr>   <chr>         <dbl> <int>   <dbl> <chr>
## 1 rmse    standard   3295049.    10 607297. Preprocessor1_Model1
```

```
## # A tibble: 4 x 5
##   id                    .pred  .row       kg .config
##   <chr>                 <dbl> <int>    <dbl> <chr>
## 1 train/test split 15557070.     8 22615423 Preprocessor1_Model1
## 2 train/test split 19646447.     9 22851016 Preprocessor1_Model1
## 3 train/test split 17791874.    11 19868255 Preprocessor1_Model1
## 4 train/test split 16021915.    13 19268546 Preprocessor1_Model1
```

```
## # A tibble: 2 x 4
##   .metric .estimator    .estimate .config
##   <chr>   <chr>             <dbl> <chr>
## 1 rmse    standard   4328441.     Preprocessor1_Model1
## 2 rsq     standard         0.0969 Preprocessor1_Model1
```

Adicionalmente, hemos calculado un RMSE específico referido a la evaluación realizada sobre el conjunto de prueba del modelo de 4.328.441 kilogramos. Este valor refleja la precisión del modelo al considerar las diferencias individuales entre las predicciones y los valores reales.

En relación a las predicciones generadas por el modelo KNN con 5 vecinos, hemos observado una tendencia consistente en la cual la mayoría de las predicciones son cercanas a los valores reales de los kilogramos de aceituna producidos por campaña. Sin embargo, se ha identificado un caso en particular donde la predicción difiere significativamente del dato real.

4.2.4.4. Comparación de modelos

En esta sección, llevaremos a cabo una comparación de los resultados obtenidos por los modelos KNN que hemos estudiado. Nuestro enfoque se centrará en la evaluación del error cuadrático medio (RMSE) utilizando dos enfoques distintos: la validación cruzada y una muestra específica seleccionada. Analizaremos y contrastaremos los valores de RMSE obtenidos en ambos métodos para determinar la eficacia de cada modelo en la predicción del rendimiento.

Tabla 4.3: Tabla comparativa del RMSE para los modelos KNN basados en los kilogramos producidos

RMSE_CV	RMSE_Muestra	Modelos
3765709	4337637	KNN_1
3487310	4147918	KNN_3
3295049	4328441	KNN_5

Al analizar detenidamente la tabla de resultados, se observa una tendencia clara: a medida que aumentamos el número de vecinos en el modelo KNN, el RMSE obtenido mediante validación cruzada disminuye. Sin embargo, es importante destacar que en el caso del RMSE obtenido en la muestra seleccionada, el modelo con 3 vecinos presenta el valor más bajo. A pesar de esta diferencia, otorgaremos mayor importancia al RMSE

obtenido mediante la validación cruzada, ya que es una medida más confiable y robusta al considerar múltiples particiones del conjunto de datos. Por lo tanto, concluimos que el mejor modelo para predecir el rendimiento sería aquel que utiliza 5 vecinos.

4.2.5. Red Neuronal

En esta sección, se aplicarán modelos de redes neuronales para realizar la predicción de los kilogramos de aceituna recolectados. Específicamente, se utilizará la biblioteca Keras, apoyada en el motor Tensorflow, ampliamente conocida por su eficiencia y flexibilidad en la implementación de redes neuronales.

En este caso, se explorarán y diseñarán diferentes estructuras de redes neuronales con el objetivo de evaluar su rendimiento y determinar cuál de ellas se comporta mejor en la tarea de predicción de los kilogramos de aceituna recolectados.

Se comenzará por definir y probar distintas arquitecturas de redes neuronales, variando el número de capas ocultas, el número de neuronas en cada capa y otras configuraciones relevantes. Esto permitirá evaluar cómo estas variaciones afectan el rendimiento y la capacidad predictiva de la red neuronal.

En primer lugar, se procede a realizar una red neuronal con tres capas ocultas. Para ello llevamos a cabo la partición de nuestro conjunto de datos en conjuntos de entrenamiento y prueba. Esta división se realiza con el propósito de utilizar el conjunto de entrenamiento para ajustar los parámetros del modelo y el conjunto de prueba para evaluar su rendimiento en datos no vistos previamente.

Una vez realizada la partición en entrenamiento y prueba, procedemos al preprocesamiento de normalización de los datos. Esta etapa es esencial para garantizar que todas las variables se encuentren en una escala comparable y facilitar así el entrenamiento y la interpretación del modelo.

```
library(caret)
library(nnet)
library(tidymodels)
datos_prediccion_kg = data.frame(datos_prediccion,kg)
datos_prediccion_kg = datos_prediccion_kg %>%
  mutate_all(as.numeric)
datos_prediccion_kg = datos_prediccion_kg[-c(4,11,12,13,21),-c(1,26)]
attach(datos_prediccion_kg)

set.seed(123)
datos_prediccion_kg_split <- datos_prediccion_kg %>%
  initial_split(prop = 0.75)


preProcValues <- preProcess(training(datos_prediccion_kg_split)[,-37],
                            method = c("center", "scale"))
trainTransformed <- predict(preProcValues,
                            training(datos_prediccion_kg_split))
```

```
testTransformed <- predict(preProcValues,
                           testing(datos_prediccion_kg_split))
```

A continuación, se define el modelo y se compila para incluir la configuración del optimizador.

Las capas Dropout se utilizan como una técnica de regularización para evitar el sobreajuste del modelo. Durante el entrenamiento, se desactivan de forma aleatoria un porcentaje de las neuronas en estas capas, en nuestro caso un 30 %, lo que reduce la interdependencia entre ellas y evita que algunas neuronas dominen el proceso de aprendizaje.

El número de unidades en cada capa de la red neuronal determina la capacidad de representación y la complejidad del modelo. En este caso, se eligen 20 unidades en la primera capa oculta, 10 unidades en la segunda capa oculta y 5 unidades en la tercera capa oculta. Un mayor número de unidades en una capa permite a la red neuronal capturar relaciones más complejas y representar características más detalladas de los datos.

La función de activación ReLU (Rectified Linear Unit) se utiliza en las capas ocultas de la red neuronal. Esta función es no lineal y se define como $f(x) = max(0, x)$. La función ReLU se utiliza ampliamente en redes neuronales debido a su simplicidad y efectividad para introducir la no linealidad en el modelo. Permite que la red neuronal aprenda representaciones más complejas al activar las neuronas cuando la entrada es positiva y desactivarlas cuando la entrada es negativa o cero.

El optimizador Adam se utiliza para ajustar los pesos y los sesgos de la red neuronal durante el proceso de entrenamiento. Adam es un algoritmo de optimización basado en descenso de gradiente estocástico que combina la idea del descenso de gradiente con momentos adaptativos. Esto permite una convergencia más rápida y eficiente durante el entrenamiento de la red neuronal.

```
# Definir la arquitectura de la red neuronal
modelo <- keras_model_sequential() %>%
  layer_normalization(input_shape = shape(ncol(trainTransformed) - 1),
                      axis = NULL) %>%
  layer_dropout(0.3) %>%
  layer_dense(units = 20, activation = "relu") %>%
  layer_dropout(0.3) %>%
  layer_dense(units = 10, activation = "relu") %>%
  layer_dropout(0.3) %>%
  layer_dense(units = 5, activation = "relu") %>%
  layer_dense(units = 1)

#summary(modelo)


# Compilar el modelo
modelo %>% compile(
  loss = "mean_squared_error",
```

```
  optimizer = optimizer_adam(learning_rate = 0.01)
)
```

Luego de definir y compilar el modelo, se procede al entrenamiento de la red neuronal utilizando los datos de entrenamiento. Durante el entrenamiento, se ajustan los pesos y los sesgos de la red neuronal para minimizar la función de pérdida, en este caso, la media del error cuadrático (mean squared error) con una tasa de aprendizaje de 0.01. Se realizan varias épocas de entrenamiento (200) para mejorar gradualmente la capacidad de predicción del modelo.

Para garantizar que los datos de prueba estén en el mismo rango que los datos normalizados y que la función de pérdida sea adecuada, se realiza una división por 10 millones en la variable objetivo de los datos de prueba. Esto se hace para ajustar la escala de los datos y asegurar que estén dentro del rango adecuado para el entrenamiento y la evaluación del modelo. De esta manera, se asegura que los resultados de la función de pérdida reflejen correctamente la discrepancia entre las predicciones del modelo y los valores reales de los datos de prueba.

```
# Entrenar el modelo
history <- modelo %>% fit(
  x = as.matrix(training(datos_prediccion_kg_split)[, -37]),
  y = training(datos_prediccion_kg_split)$kg/10000000,
  epochs = 200,
  validation_split = 0.2
)


# Evaluar el modelo con los datos de prueba
test_loss <- modelo %>% evaluate(
  x = as.matrix(testing(datos_prediccion_kg_split)[, -37]),
  y = testing(datos_prediccion_kg_split)$kg/10000000
)

# Obtener las predicciones para los datos de prueba
predicciones <- modelo %>% predict(
  x = as.matrix(testing(datos_prediccion_kg_split)[, -37])
)

predicciones = as.vector(predicciones*10000000)

predicciones_finales = data.frame(predicciones,
                                  testing(datos_prediccion_kg_split)$kg)

colnames(predicciones_finales) = c("predicciones","kilogramos reales")
```

Una vez entrenado el modelo, se evalúa su rendimiento utilizando los datos de prueba. Se calcula la pérdida del modelo en los datos de prueba y se obtienen las predicciones para

estos datos. Estas predicciones permiten evaluar la capacidad del modelo para realizar predicciones precisas sobre nuevos datos.

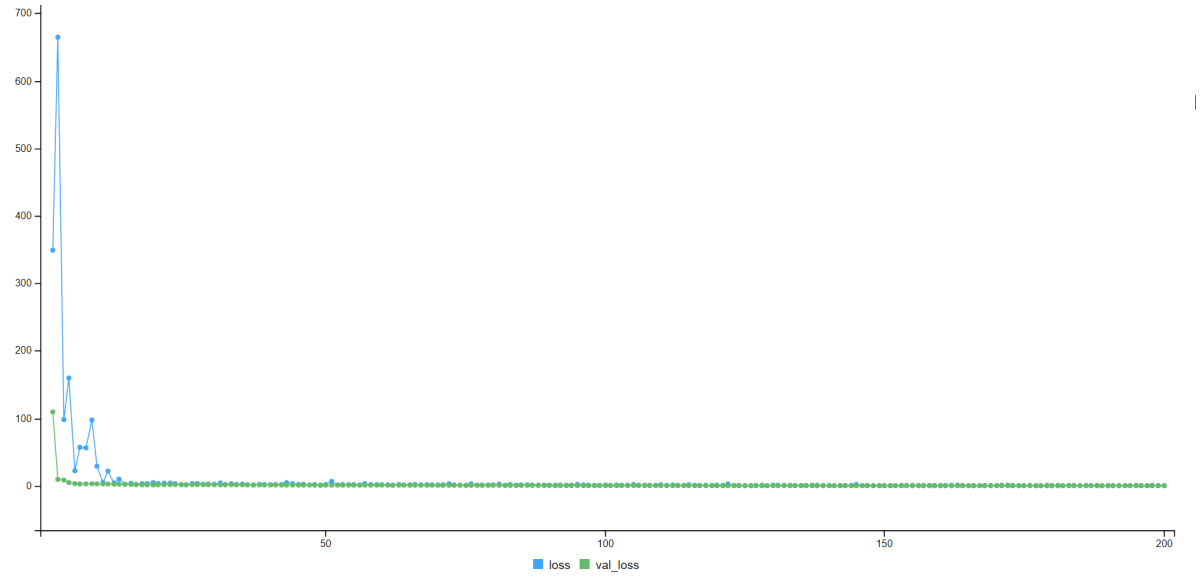


Figura 4.7: Función de pérdida de la Red Neuronal con 3 capas (fuente: elaboración propia)

```
Final epoch (plot to see history):
    loss: 0.2187
val_loss: 0.07766
```

En el gráfico de entrenamiento se observa cómo las funciones “loss” y “val_loss” convergen a lo largo de las épocas. La función “loss” representa la medida del error entre las predicciones del modelo y los valores reales de entrenamiento, mientras que “val_loss” es una medida similar pero calculada en el conjunto de validación.

Al finalizar el entrenamiento, se obtiene un valor de 0.2137 para la función “loss” y un valor de 0.0776 para “val_loss”. Sin embargo, es importante tener en cuenta que estos valores deben ser interpretados considerando que los datos de prueba se dividieron previamente por 10 millones para que estuvieran en el mismo rango que los datos normalizados.

Esto significa que, en realidad, el error promedio entre las predicciones del modelo y los valores reales en los datos de entrenamiento es de aproximadamente 2.137 millones y en los datos de validación es de aproximadamente 776 mil. Estas cifras representan la magnitud del error promedio en la escala original de los datos, y permiten evaluar el desempeño del modelo en términos más interpretables. En este caso, los valores obtenidos indican que el modelo tiene un bajo error promedio en la predicción de los datos, lo que sugiere un buen ajuste y una capacidad predictiva satisfactoria.

A continuación, se presentan las predicciones obtenidas por el modelo para los datos de prueba. Estas predicciones representan las estimaciones del modelo sobre los kilogramos de aceituna recolectados

Description: df [4 × 2]

predicciones <dbl>	kilogramos reales <dbl>
17038549	22615423
14997038	22851016
15701703	19868255
15877582	19268546

4 rows

Por otro lado, procedemos a analizar una red neuronal con dos capas ocultas, manteniendo las mismas configuraciones utilizadas en la red neuronal anterior. Esto implica que se empleará la misma cantidad de unidades en cada capa oculta y se seguirá utilizando la función de activación ReLU. Además, se utilizará el optimizador Adam y se establecerá la tasa de aprendizaje en 0.01. Mediante esta configuración, se busca evaluar cómo el número de capas ocultas afecta el rendimiento y la capacidad predictiva del modelo neuronal en la tarea de predicción de los kilogramos de aceituna recolectados.

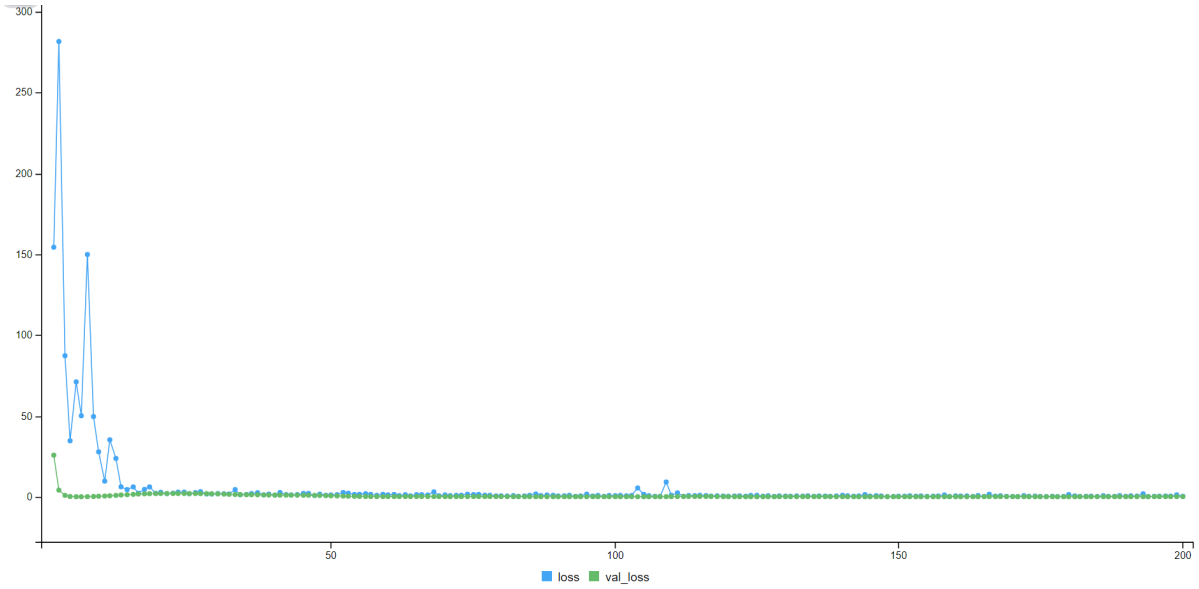


Figura 4.8: Función de pérdida de la Red Neuronal con 2 capas (fuente: elaboración propia)

```
Final epoch (plot to see history):
    loss: 0.318
val_loss: 0.1226
```

En el gráfico se muestra la evolución de la función de pérdida (loss) y la pérdida en el conjunto de validación (val_loss) durante el entrenamiento de la red neuronal con dos capas ocultas. Se observa que ambas métricas disminuyen gradualmente a medida que

avanza el proceso de entrenamiento, alcanzando un valor final de 0.318 para la pérdida y 0.1226 para la pérdida en el conjunto de validación. Es importante tener en cuenta que estos valores deben ser interpretados considerando la escala de los datos y la normalización previa realizada en los kilogramos de aceituna recolectados.

Comparando estos resultados con los obtenidos en la red neuronal de tres capas ocultas, se observa que la pérdida y la pérdida en el conjunto de validación son más altas en la red neuronal de dos capas ocultas. Esto puede indicar que la configuración de tres capas ocultas logró un mejor ajuste a los datos y una mayor capacidad de generalización. Sin embargo, es importante analizar estos resultados en conjunto con otras métricas de evaluación y considerar las características específicas del problema para determinar la mejor configuración del modelo.

A continuación, se presentan las predicciones obtenidas por el modelo para los datos de prueba.

Description: df [4 × 2]

predicciones <dbl>	kilogramos reales <dbl>
18777809	22615423
12145450	22851016
12802038	19868255
13780510	19268546

4 rows

Es importante destacar que los ejemplos presentados son solo dos posibles configuraciones de redes neuronales, y existen numerosas combinaciones y variaciones que se pueden explorar al ajustar los diferentes parámetros. Al modificar la cantidad de capas ocultas, el número de unidades en cada capa, la función de activación, el optimizador y otros aspectos, se pueden obtener resultados y desempeños diferentes. Por tanto, es recomendable realizar pruebas y experimentos con diferentes configuraciones para determinar la mejor arquitectura de red neuronal que se ajuste a los datos y maximice el rendimiento en el problema específico que se está abordando. La elección de la configuración adecuada dependerá de las características de los datos, la naturaleza del problema y las metas del estudio.

4.2.6. Comparación de los modelos predictivos empleados

En esta sección, examinamos y contrastamos el desempeño de varios modelos predictivos utilizados para predecir los kilogramos de aceituna producidos. Con el fin de realizar una comparación confiable y sólida, nos apoyamos en el RMSE promedio obtenido a través de la validación cruzada. Esta medida nos proporciona una perspectiva más sólida debido al tamaño de nuestro conjunto de datos y a la sensibilidad que puede surgir según la partición de los datos utilizada. A continuación, presentamos una tabla que muestra los valores de RMSE correspondientes a cada modelo, los cuales serán analizados y comparados detalladamente.

Tabla 4.4: Tabla comparativa del RMSE de los modelos basados en los kilogramos producidos

RMSE	Modelos
3669865	PCA
3178941	RF
3765709	KNN_1
3487310	KNN_3
3295049	KNN_5
2137000	RN

Al analizar los valores de RMSE promedio de cada modelo, se observa que el modelo de red neuronal tiene el valor más bajo de 2.137.000, lo que indica una menor discrepancia entre las predicciones y los valores reales de los kilogramos totales producidos por campaña. Esto sugiere una mayor precisión en las predicciones y una mejor capacidad para capturar las relaciones en los datos en comparación con los otros modelos..

Además, el modelo Random Forest muestra un valor de RMSE promedio de 3.178.941, lo que indica una discrepancia mayor en comparación con el modelo de red neuronal. Aunque el modelo Random Forest tiene un rendimiento competitivo, el menor valor de RMSE promedio del modelo de red neuronal sugiere una mayor precisión en las predicciones de los kilogramos totales.

El Random Forest también muestra una mejora significativa en comparación con el PCA, que tiene un RMSE promedio de 3.669.865. Esto sugiere que el modelo Random Forest tiene una capacidad superior para capturar las relaciones y patrones subyacentes en los datos, lo que se traduce en predicciones más precisas y cercanas a los valores reales.

En cuanto a los modelos KNN con 1 vecino, KNN con 3 vecinos y KNN con 5 vecinos, sus valores de RMSE promedio son de 3.765.709, 3.487.310 y 3.295.049 respectivamente. Si bien estos modelos pueden ofrecer predicciones razonables, sus RMSE promedio son más altos en comparación con la red neuronal, lo que indica una mayor discrepancia con los valores reales.

En resumen, con base en el análisis de los valores de RMSE promedio, seleccionaríamos la red neuronal como el mejor modelo para predecir los kilogramos totales producidos por campaña. Su menor valor de RMSE promedio sugiere una mayor precisión en las predicciones y una mejor capacidad para capturar las relaciones en los datos.

Sin embargo, además del RMSE promedio, es importante considerar otros factores para evaluar la idoneidad de los modelos, como el coeficiente de determinación. El coeficiente de determinación es una medida que indica la proporción de la variabilidad de la variable objetivo que es explicada por el modelo.

Al examinar los coeficientes de determinación de los modelos, observamos que el modelo de regresión múltiple con PCA tiene el coeficiente más alto, con un valor de 0.694. Esto indica que alrededor del 69 % de la variabilidad de los kilogramos totales producidos por campaña se explica por el modelo PCA. Aunque el modelo RF tiene un RMSE promedio más bajo, su coeficiente de determinación es de 0.53, lo que significa que solo explica aproximadamente el 53 % de la variabilidad de la variable objetivo. Esto implica que

el modelo PCA puede ser más adecuado si se valora en mayor medida la capacidad de explicación y ajuste del modelo.

Capítulo 5

Conclusiones

En este capítulo se presentan las principales aportaciones y hallazgos derivados del estudio realizado. Se discuten los resultados obtenidos, destacando las contribuciones significativas al campo de investigación. Además, se identifican posibles mejoras y dificultades que podrían abordarse en trabajos futuros, con el objetivo de avanzar en el conocimiento y enriquecer la investigación en este ámbito.

5.1. Aportaciones

En este estudio, hemos realizado un buen número de aportaciones significativas que consideramos que pueden contribuir de manera sustancial al conocimiento en el campo de la producción agrícola. A través de un análisis exhaustivo, hemos investigado y explorado las relaciones existentes entre las variables de estudio, revelando tanto las interacciones entre ellas como sus comportamientos individuales. Este análisis detallado nos ha permitido comprender mejor los factores que influyen en la producción agrícola y cómo se relacionan entre sí.

Además de analizar las relaciones entre las variables, hemos llevado a cabo un minucioso estudio del comportamiento de estas variables durante las diferentes campañas agrícolas. Hemos identificado patrones estacionales, tendencias a largo plazo y variaciones interanuales, lo que proporciona información valiosa para comprender la dinámica de la producción agrícola a lo largo del tiempo. Esta comprensión nos permite realizar pronósticos más precisos y tomar decisiones estratégicas informadas en términos de planificación y gestión de cultivos.

Una de las principales contribuciones de este estudio radica en el desarrollo y aplicación de diversos modelos predictivos. Hemos utilizado técnicas estadísticas y de aprendizaje automático para construir modelos que pueden predecir con razonable precisión las dos variables más importantes en la producción agrícola. Estos modelos tienen en cuenta las relaciones y tendencias identificadas en el análisis de datos, lo que les permite generar pronósticos útiles para la toma de decisiones en el sector agrícola.

A continuación, se detallan los modelos utilizados para cada una de las variables objetivo:

Rendimiento

- Serie temporal: se aplicó un modelo de serie temporal para capturar patrones y tendencias en el rendimiento a lo largo del tiempo.
- Regresión lineal múltiple con análisis de componentes principales (PCA): se empleó este modelo para explorar las relaciones lineales entre múltiples variables y el rendimiento. El PCA permitió reducir la dimensionalidad de las variables y capturar las principales componentes explicativas.
- Random Forest: se utilizó este modelo, que es un ensamble de árboles de decisión, para capturar las relaciones no lineales y las interacciones entre las variables predictoras y el rendimiento.
- K-Nearest Neighbors (KNN): se aplicó el algoritmo KNN para encontrar los vecinos más cercanos en función de las características de los datos y predecir el rendimiento en base a ellos.
- Redes neuronales: se empleó un modelo de redes neuronales para capturar patrones complejos y no lineales en los datos y realizar predicciones del rendimiento.

Kilogramos de aceituna producidos

- Serie temporal: se empleó un enfoque especializado en el análisis de datos secuenciales para capturar las fluctuaciones y tendencias en la producción de aceituna a lo largo del tiempo.
- Regresión lineal con análisis de componentes principales (PCA): se utilizó este método estadístico para explorar las relaciones lineales entre múltiples variables y la producción de aceituna. El PCA fue aplicado para reducir la dimensionalidad de los datos y resaltar las componentes más influyentes.
- Random Forest: se empleó un modelo basado en bosques aleatorios para identificar las características más relevantes y capturar las relaciones no lineales entre las variables predictoras y la producción de aceituna.
- K-Vecinos más cercanos (KNN): se utilizó el algoritmo KNN para encontrar los puntos de datos más cercanos y predecir la producción de aceituna en función de las características de esos vecinos más cercanos.
- Modelos de redes neuronales: se aplicaron modelos de redes neuronales para capturar los patrones complejos en los datos y hacer predicciones sobre la producción de aceituna.

También se llevó a cabo una comparación de los diferentes modelos mencionados anteriormente para determinar cuál de ellos proporcionaba las predicciones más precisas para cada variable específica. Se realizaron evaluaciones exhaustivas utilizando métricas de rendimiento como el error cuadrático medio (RMSE) o el coeficiente de determinación (R^2) para analizar el desempeño de cada modelo en relación con la variable objetivo.

Además, nuestras investigaciones han permitido identificar y comprender los principales impulsores y desafíos en la producción agrícola. Hemos examinado factores como la disponibilidad de recursos, las condiciones climáticas y otras variables relevantes para determinar su impacto en la producción y rendimiento agrícola. Esta comprensión más profunda nos brinda información valiosa para desarrollar estrategias y políticas que impulsen la sostenibilidad y la eficiencia en el sector agrícola.

En resumen, este estudio ha realizado valiosas aportaciones al campo de la producción agrícola, mejorando nuestra comprensión de las relaciones entre variables, proporcionando conocimientos sobre el comportamiento de las variables durante las campañas agrícolas y desarrollando modelos predictivos precisos. Estas contribuciones tienen implicaciones tanto teóricas como prácticas, ya que pueden ayudar a optimizar la producción agrícola, mejorar la toma de decisiones y fomentar la sostenibilidad en el sector agrícola.

5.2. Hallazgos

En esta sección se presentan los principales resultados obtenidos en este estudio, revelando la influencia de las variables de temperatura y precipitación en las dos variables de interés en la producción agrícola. Se ha observado que tanto las temperaturas como las precipitaciones desempeñan un papel crucial en el rendimiento y los kilogramos de producción.

En base al estudio realizado, se han obtenido conclusiones significativas que pueden influir en la toma de decisiones de los agricultores. Se ha observado una correlación negativa entre los rendimientos de aceituna y la humedad del fruto a lo largo de la campaña. Al inicio de la campaña, cuando los rendimientos son más bajos, la humedad del fruto es mayor. Esto tiene beneficios tanto en términos de peso de la aceituna, ya que se obtiene una mayor cantidad de materia prima para la extracción de aceite de oliva, como en la calidad del aceite producido, al reducir la oxidación de los compuestos presentes en la aceituna.

Por lo tanto, considerar la humedad del fruto al tomar decisiones relacionadas con el inicio de la campaña es fundamental para maximizar tanto la cantidad como la calidad del aceite producido. Los agricultores pueden utilizar esta información para determinar el momento más adecuado para comenzar la recolección de aceitunas, teniendo en cuenta los beneficios asociados a una mayor humedad en las etapas iniciales de la campaña. Esto puede contribuir a optimizar la producción y obtener un aceite de oliva de mejor calidad.

En este estudio, se ha logrado desarrollar modelos predictivos para estimar las variables de rendimiento y producción en kilogramos. Es importante resaltar que estos modelos han demostrado un desempeño sólido en términos de precisión.

Para predecir el rendimiento, se evaluaron varios modelos y se utilizó el criterio de error cuadrático medio (RMSE) para determinar cuál ofrecía la mejor estimación. El modelo que presentó el menor RMSE y, por lo tanto, se considera el más preciso, fue la regresión lineal múltiple con componentes principales, con un RMSE de 1,50.

De cerca, el segundo modelo con mejor desempeño fue el random forest, con un RMSE de 1,56. Esto indica que existe una pequeña diferencia de 0,06 unidades en el RMSE entre el modelo de regresión lineal múltiple con componentes principales y el random forest.

En términos porcentuales, el margen de error del modelo de regresión lineal múltiple con componentes principales es del 7 % al predecir el rendimiento. Esto significa que, en promedio, existe un error del 7 % al utilizar este modelo para estimar el rendimiento de manera precisa.

Además, en relación a la predicción de los kilogramos de aceituna producidos, se observó que la red neuronal presentó el menor RMSE, con un valor de 2137000. Esto indica que la red neuronal logró estimar de manera precisa la cantidad de kilogramos de aceituna producidos, con un margen de error mínimo en comparación con los otros modelos evaluados.

Sin embargo, es importante destacar que el modelo de regresión lineal múltiple con componentes principales mostró el mayor coeficiente de determinación (R^2) entre todos los modelos.

Aunque la red neuronal presentó el menor RMSE en la predicción de los kilogramos de aceituna producidos, el modelo de regresión lineal múltiple con componentes principales se destaca al tener un mayor R^2. Esto sugiere que este modelo es más efectivo en términos de explicar la variabilidad de los datos y proporcionar una estimación confiable de los kilogramos de aceituna producidos.

5.3. Mejoras y dificultades de cara a trabajos futuros

Durante el desarrollo de esta investigación, nos hemos enfrentado a diversos desafíos y limitaciones que es importante tener en cuenta para futuros trabajos. A continuación, se describen detalladamente algunas de estas dificultades y se plantean posibles mejoras y áreas de investigación adicionales:

- Limitaciones en los datos disponibles: uno de los principales problemas encontrados fue la falta de datos completos y enriquecidos para algunas variables relevantes. Por ejemplo, solo disponemos de datos completos de todas las variables a partir de la campaña 2017/2028. Esto implica que no podemos realizar un análisis exahustivo a largo plazo, ya que el historial de datos es limitado y no podemos considerar tendencias a largo plazo. Sería beneficioso contar con un conjunto de datos más completo y abarcador, que incluya información desde años anteriores con todas las variables, lo que permitiría realizar análisis comparativos más sólidos.

- Ausencia de información sobre tratamientos del olivar: otra limitación importante es la falta de datos detallados sobre los tratamientos aplicados a los olivares. Esta información es costosa de recopilar y suele ser difícil de determinar, lo que limita nuestra capacidad de comprender el impacto de diferentes prácticas agrícolas en la producción de aceituna. Sería valioso contar con registros detallados de los tratamientos aplicados en cada olivar, lo que permitiría realizar estudios más exhaustivos y precisos sobre la relación entre los tratamientos y la producción de aceituna.

- Aperturas de nuevas fábricas y cambios en la producción: durante el desarrollo de este estudio, se han observado aperturas de nuevas fábricas en la zona de estudio. Esto ha llevado a que algunos socios de la cooperativa no entreguen la totalidad de su producción a la misma, lo que dificulta la predicción precisa de la producción

total. Esta situación introduce cierta incertidumbre y puede afectar la calidad y precisión de las predicciones. En trabajos futuros, se podría considerar el estudio de métodos o modelos específicos para abordar estos cambios en la producción y adaptarse a las nuevas dinámicas de las cooperativas y fábricas.

- Ampliar el alcance de la investigación a otras cooperativas y zonas de cultivo: para futuros estudios, sería beneficioso ampliar el alcance de la investigación y aplicar las técnicas desarrolladas en este trabajo a otras cooperativas y zonas de cultivo. Esto permitiría validar y generalizar los resultados obtenidos, así como identificar posibles diferencias o patrones específicos de cada región. Para lograr esto, sería necesario contar con un historial de datos más amplio y representativo de diversas cooperativas y zonas de cultivo.

En resumen, a pesar de los desafíos y limitaciones encontrados en esta investigación, se identifican oportunidades claras para mejorar y ampliar el estudio. Estas mejoras incluyen la obtención de datos más completos y enriquecidos, la recopilación de información detallada sobre tratamientos del olivar, el desarrollo de modelos adaptados a los cambios en la producción y la expansión de la investigación a otras cooperativas y zonas de cultivo. Al abordar estas dificultades y realizar investigaciones adicionales, se podrán obtener resultados más sólidos y aplicables en el ámbito de la predicción de la producción de aceitunas en el sector cooperativo.

Bibliografía

JJ Allaire and François Chollet. *keras: R Interface to 'Keras'*, 2023. URL https://CRAN.R-project.org/package=keras. R package version 2.11.1.

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2022. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.19.

Paul S.P. Cowpertwait Andrew V. Metcalfe. *Introductory Time Series with R.* Springer-Verlag New York, 2009.

Kung-Sik Chan and Brian Ripley. *TSA: Time Series Analysis*, 2022. URL https://CRAN.R-project.org/package=TSA. R package version 1.3.1.

Uriel Ezequiel. *Análisis de Series Temporales. Modelos Arima.* Paraninfo, 2009.

Hannah Frick, Fanny Chow, Max Kuhn, Michael Mahoney, Julia Silge, and Hadley Wickham. *rsample: General Resampling Infrastructure*, 2022. URL https://CRAN.R-project.org/package=rsample. R package version 1.1.1.

Stefan Fritsch, Frauke Guenther, and Marvin N. Wright. *neuralnet: Training of Neural Networks*, 2019. URL https://CRAN.R-project.org/package=neuralnet. R package version 1.44.2.

Miguel González Velasco and Maria Inés del Puerto García. *Series temporales.* Cáceres, Universidad de Extremadura, 2009.

Faraway J.J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* Chapman and Hall-CRC., 2006.

Christoph A. Keller and Mat J. Evans. Application of random forest regression to the calculation of gas-phase chemistry within the geos-chem chemistry model v10. 12(3), 2019.

Max Kuhn and Davis Vaughan. *parsnip: A Common API to Modeling and Analysis Functions*, 2023. URL https://CRAN.R-project.org/package=parsnip. R package version 1.0.4.

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL https://www.tidymodels.org.

Max Kuhn and Hadley Wickham. *recipes: Preprocessing and Feature Engineering Steps for Modeling*, 2023. URL https://CRAN.R-project.org/package=recipes. R package version 1.0.5.

Rahman MA, Muniyandi Rc, Albashish D, Rahman MM, and Usman OL. Artificial neural network with taguchi method for robust classification model to improve classification accuracy of breast cancer. PeerJ Computer Science 7:e344, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. URL http://www.rstudio.com/.

Techopedia. "definition - what does business intelligence (bi) mean?". Disponible en https://www.techopedia.com/definition/345/business-intelligence-bi, 2017.

Adrian Trapletti and Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*, 2022. URL https://CRAN.R-project.org/package=tseries. R package version 0.10-52.

Davis Vaughan and Simon Couch. *workflows: Modeling Workflows*, 2023. URL https://CRAN.R-project.org/package=workflows. R package version 1.1.3.

Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2023. URL https://CRAN.R-project.org/package=readxl. R package version 1.4.2.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2023a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.4.1.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.0.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023c. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.0.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2022. URL https://yihui.org/knitr/. R package version 1.41.