UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Administración y Economía

Predicción de Días Restantes Para la Cosecha

.

Elías José Mantilla Ibarra

Economía

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Economista

Quito, 02 de diciembre de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Administración y Economía

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Predicción de Días Restantes Para la Cosecha

Elías José Mantilla Ibarra

Nombre del profesor, Título académico **Ricardo López, MSc**

Quito, 02 de diciembre de 2022

© DERECHOS DE AUTOR

Nombres y apellidos: Elías José Mantilla Ibarra

Código: 00210862

Cédula de identidad: 1722109731

Lugar y fecha: Quito, 02 de diciembre de 2022

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El presente trabajo busca proponer un modelo de predicción para los días restantes a la cosecha para la variedad de rosas Freedom. Se entrenará una serie de modelos cuyas predicciones serán comparadas de manera tal que uno de los algoritmos de aprendizaje se defina como el mejor y resulte seleccionado.

Palabras clave: predicción, modelo predictivo, flores, florícola, cosecha

ABSTRACT

The present work aims at proposing a predictive model for the days to harvest for the rose variety known as Freedom. A series of models will be trained whose prediction will be compared in order to select the best model for actual use.

Key words: prediction, predictive model, flower, floriculture, harvest.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

INTRODUCCIÓN

El sector floricultor ha crecido en importancia durante las últimas décadas para la economía ecuatoriana. La actividad de esta industria se encuentra concentrada en la región Sierra, adquiriendo todavía mayor relevancia para en la provincias de Pichincha y Cotopaxi. Las flores ecuatorianas crecen en alturas de entre 2,200 a 2,800 m.s.n.m lo que permite cosechas en menores periodos, están expuestas a 12 horas de luz al día y una temperatura estable alrededor del año lo que le impacta positivamente sobre la calidad de la flor (Macias & Villalta, 2015). Asi el Ecuador se ha convertido el primer exportador de rosas en el mundo.

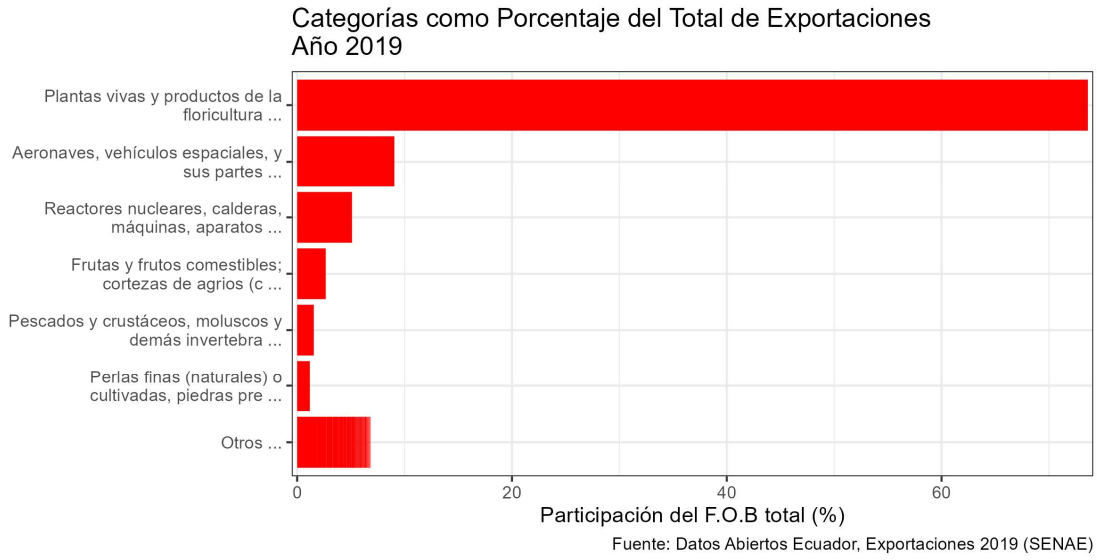

Figura 1: Participación de Categorías de Productos en Exportaciones desde Quito, Año 2019.

La base de datos de SENAE contienen información sobre las exportaciones desde distintas jefatura aduaneras, entre ellas el aeropuerto de Quito. Las exportaciones desde esta jefatura da atisbo sobre la importancia que tiene la horticultura en la provincia de Pichincha en cuanto a exportables. Nótese que la categoría de 'Plantas vivas y productos de la floricultura' no solamentente contiene las exportaciones de flores pero también de tubérculos y demás productos agrícolas. No obstante, esta categoria domina en gran medida el segmento de la

canasta exportadora desde Quito. Las exportaciones de la categoria superan el 70% y el 80% del total en los años 2019 y 2021, respectivamente.

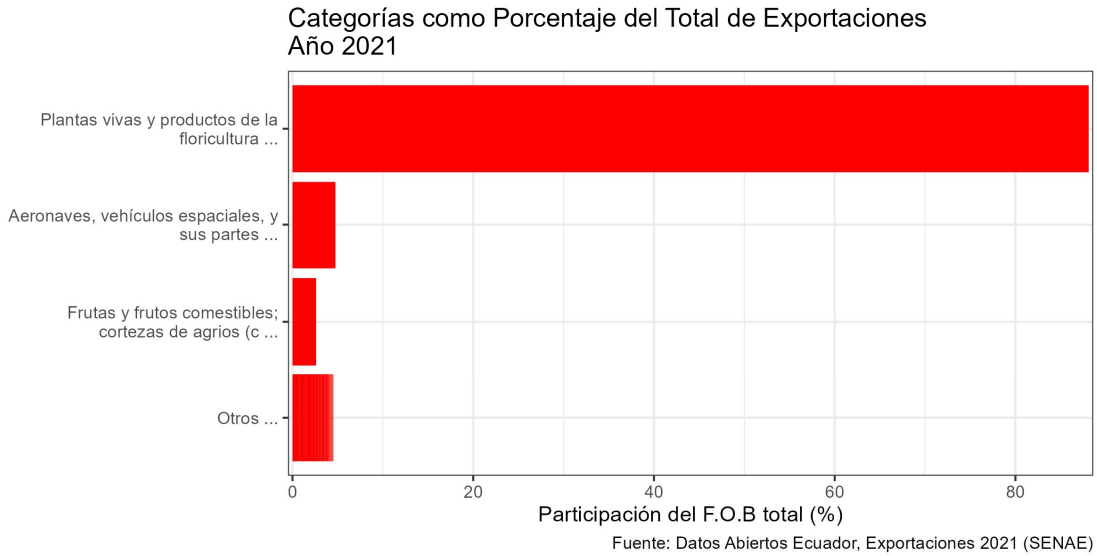

Figura 2: Participación de Categorias de Productos en Exportaciones desde Quito, Año 2021.

Para las fincas florícolas, la previsión del stock de flor para fechas clave es de suma importancia. Variaciones en el stock de producto pueden deberse al número de ciegos entre las flores - plantas que no llegan a brotar, cambios imprevistos en el clima que aceleran o ralentizan la maduración de la flor, intuiciones que erran de parte de los agrónomos a cargo sobre el número de días faltantes para la cosecha. Todos estos elementos pueden confluir y generar incertidumbre sobre el volumen de producción que se tendrá disponible. Este problema no es menor. Subestimar el número de flores disponibles implicaría decepcionar al cliente o forzaría a proveerse de flores de otras fincas, sobre cuya calidad no se ejerce control alguno. Por otro lado, sobreestimar el volumen de producción forzaría a desechar producto o venderlo con un gran descuento.

Mejorar el grado de certidumbre acerca de la producción con la que se contará dentro de un horizonte de tiempo dado es menester del gestor de una finca florícola a razón de la estacionalidad de las ventas. Las ventas de flores alcanzan sus picos mayores en el mes de

febrero a causa de las festividades del amor y la amistad, y uno menor durante mayo a causa del día de las madres.

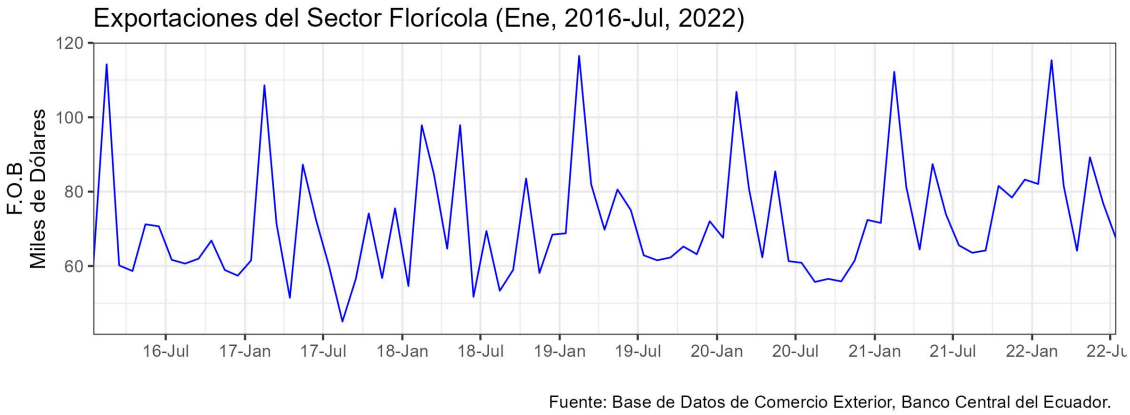

Figura 3: F.O.B Exportaciones

Los datos obtenidos para la realización de este proyecto corresponden a un bloque de uno de los invernaderos de una finca ubicada en el cantón Cayambe, Pichincha. La gerencia y el área técnica de la florícola notaron la necesidad de mejorar la capacidad predictiva de manera tal que se mitige en cierta medida el problema ocasionado por la incertidumbre en la cantidad de stock. En la actualidad, son los agrónomos quienes, sirviéndose de la experiencia e intuición de expertos, estiman la cantidad de rosas que se obtendrán en un horizonte cualquiera de tiempo. La recolección de la data es manual y laboriosa. Por esta última razón, se cuenta con una muestra de 63 flores.

En este proyecto se busca explorar la data disponible, y proponer modelos que eventualmente se implementen en producción para coadyudar en la toma de decisiones a los gestores de la finca.

ANTECEDENTES

El método tradicional para pronosticar el inicio de la cosecha es contar el número desde el primer corte o inicio de florecimiento de la planta (Baptista et al., 2006). Más adelante llegaría a considerarse a la acumulación de grados días como el indicador fenológico predilecto (Carlson and Hancock, 1991). Los grados día son también conocidos como unidades térmicas o de calor, Muñoz et al. (2012) utilizan modelos de series de tiempo para pronosticar los máximos y mínimos de temperatura diaria. A partir de esos pronósticos se calculan los grados días acumulados y se predice el número de días restantes a la cosecha de cultivos de arándano. Para el presente trabajo se hará el cálculo de grados días de manera similar a Pérez et al. (2002). El cálculo es simple y consiste en substraer de la temperatura promedio un número de grado denominados el umbral de activación o desarrollo de la flor. Este umbral es el valor de la temperatura bajo el cual se detiene el desarrollo. Este umbral identificado por los técnicos de la florícola es de 5.5° C.

Pérez et al. (2002) concluyen que el muestreo y la revisión de estados fenológicos son importantes y superiores en capacidad predictiva que el mero conteo de días pinch a brotación. Adicionalmente, reafirman la importancia ya establecida de la temperatura y sugieren controlar esta variable al interior del invernadero. Son varios los trabajos realizados en Ecuador y Colombia en aras de incrementar la certidumbre sobre el periodo de cosecha. Cañizares & Leiva (2014) emplean y ejecutan un diseño experimental para tres variedades de rosa en tres distintas fincas. Utilizaron una regresión lineal para la cual emplearon temperaturas mínima y máxima junto con los grados día. La variables dependientes fueron días de cambio de estado fenológico, la altura del tallo y el diámetro del tallo. Además, ajustan una regresión logística para obtener una estimación de la curva de crecimiento para el largo del tallo. Los autores

generan una regresión lineal para predecir el estado fenológico, variable categórica que fue codificada. Vila (2009), en cambio, utiliza datos de ciclos pasados para estimar la proporción de cosecha obtenida dada una cantidad de tiempo transcurrida. Con datos de ciclos pasados reconstruye la curva de crecimiento - la distribución de rosas cosechadas por día, que empleará en conjunto con el volumen de la cosecha anterior ajustado por el porcentaje de cosecha exitosa esperado. Las predicciones resultantes son útiles pero carecen de la incorporación de más datos pues utilizan solo los datos de una cosecha anterior lo que puede sesgar las estimaciones potencialmente.

Un problema similar a la predicción de días restantes a la cosecha es la predicción del volumen de cosecha. Van Klompenburg et al (2020) realizan una revisión sistemática de la literatura que busca encontrar solución al problema. Los resultados de la revisión concluyen que los trabajos más recientes utilizan algoritmos 5 de aprendizaje automático y aprendizaje profundo pero recurren a la regresión lineal como el punto de referencia que permite evaluar desempeño. En particular fueron mencionadas las redes neuronales. Los estudios revisados, particularmente aquellos que empleaban aprendizaje profundo, contaban con considerables cantidades de datos. La principal alternativa a los modelos estadísticos que explotan datos disponibles son aquellos fundamentados en principios físicos. Rodríguez et al (2015) proponen una serie de modelos para los meso sistemas que tienen lugar dentro del invernadero. Uno de los modelos corresponde a la temperatura dentro del invernadero con lo que se puede visualizar más precisamente la evolución y acumulación de los grados día. Estos modelos son simulados y sus predicciones son insumos para dispositivos de control e información para los gestores de la finca. Modelos basados en la teoría y principios físicos subyacentes pueden ser un soporte a los modelos basados en datos observados, principalmente cuando la cantidad de datos es limitada o su levantamiento es costoso. También es posible recurrir a las dos clases de modelos. Bernsen et al (2014) presentan un modelo de predicción con ecuaciones diferenciales que

describen el proceso de fotosíntesis local combinado con un modelo de entorno global del invernadero para predecir volumen de producción por unidad de área del invernadero. Este modelo es calibrado y evaluado con datos observacionales de alta frecuencia.

METODOLOGÍA

Los datos disponibles consisten de 77 muestras de rosas que alcanzaron la cosecha, siendo la rosa la unidad de análisis a la que le corresponde una observación en los datos, correspondientes a distintos ciclos. Cada uno de los ciclos corresponde a un periodo de alrededor de 3 a 4 meses. El primer ciclo comenzó en el mes de octubre del 2015 y el último ciclo finalizó en enero del 2022. Estas observaciones corresponden a la variedad de rosa Explorer. Para cada muestra es posible observar la sucesión de estados fenológicos, registrados mediante apreciación visual, y la evolución de variables ambientales, que fueron obtenidas de una estación meteorológica. Sin embargo, para el presente trabajo se concentrará en predecir el número de días restantes a partir del estado fenológico inmediatamente anterior a la cosecha. El estado fenológico anterior a la cosecha son las 5 lígulas. El levantamiento de la información no es constante por lo que muchas flores cosechadas no fueron registradas al haber alcanzado las 5 lígulas, por ende, nos quedan 63 observaciones (81 %) del conjunto original. Las variables consideradas para como el conjunto de predictores base son :

Cuadro 1: Descripción de Predictores Base

rawvar	Predictor
dias_des	Días desde Pinch
largo_cm	Largo del Tallo (cm)
jouls_cum	Jouls Acumulados
gd_cum	Unidades Térmicas Acumuladas
irrig_cum	Litros/Metro Cuadrado Acumulados (Irrigación)
c.e_cum	Conductividad Eléctrica Acumulada (Nutrientes)
rain_cum	Lluvía Acumulada (mm)

Estos predictores serán incluidos en los modelos subsiguientes según su grado de relación con la variable dependiente, modelos cuyo propósito principal será el de estimar el número de días restantes a la cosecha, condicional a haber alcanzado el estado fenológico de 5

lígulas. La preferencia por sentar el enfoque en uno de los estados se hace para así poder conformar una base de datos en la cual cada observación corresponda a una unidad de análisis. En esencia, se cuenta con un panel de datos y cada panel corresponde a una muestra observada a través de cada uno de los estados fenológicos. De no ser el caso que las observaciones sean estadísticamente independientes, violaríamos una presuposición de varios de los modelos, en particular de los modelos lineales de regresión.

El número de días restantes a la cosecha da indicios de alta variabilidad como muestra la Figura 4. Podemos ver que un número importante de rosas llegan al momento propicio para la cosecha en menos de 5 días después de que su estado fenológico haya sido declarado como '5 lígulas'. Sin embargo, la distribución presenta una cola alargada con observaciones que tomaron más de 10 días en estar listas para la cosecha, alrededor de un tercio de las rosas (30 %).

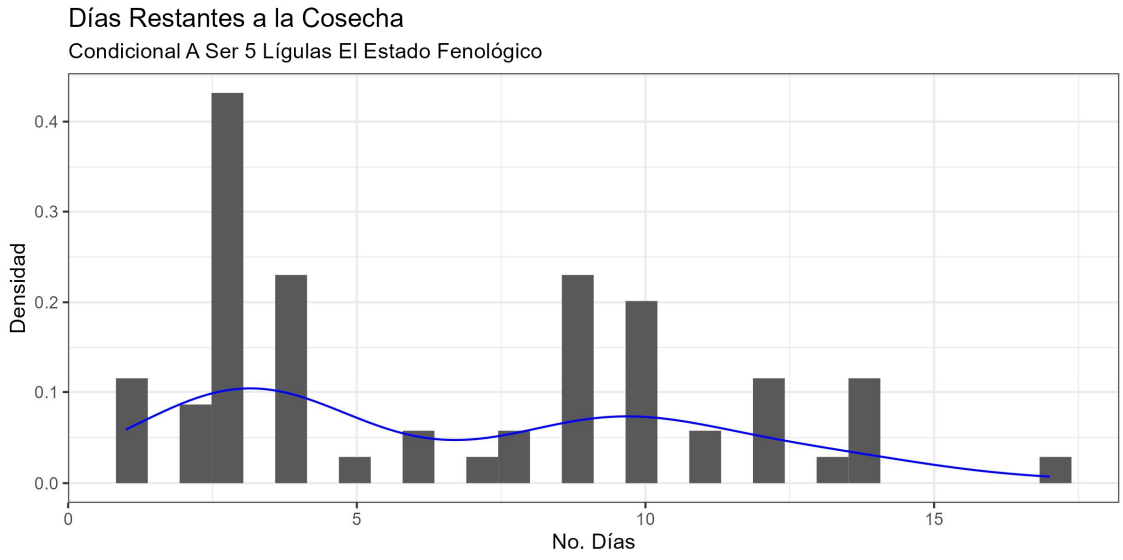

Figura 4: Días Restantes a la Cosecha

La Figura 5 muestra escasa multicolinealidad. Algunas correlaciones son fáciles de entreveer. Por ejemplo, el número de días desde el Pinch con todo el resto de predictores puesto que todas las variables ambientales únicamente incrementan en el tiempo. Entre las potenciales

variables con problemas de correlación están los días desde pinch con los jouls acumulados y los días desde pinch con los litros cuadráticos de fertilizante acumulados.

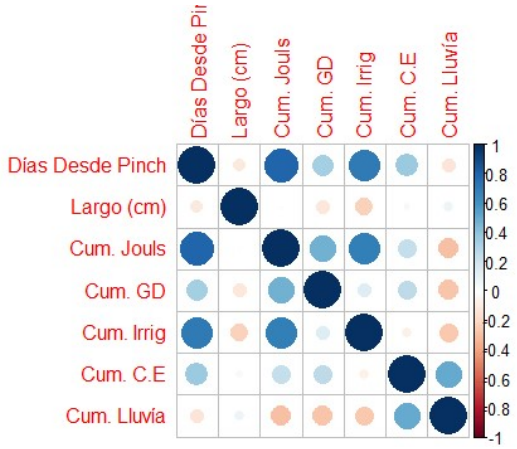

Figura 5: Matriz de Correlaciones

Como muestra en la Figura 6, es posible deducir que la relación entre las variables independientes con la variable dependiente u objetivo puede ser aproximada por una recta o ecuación lineal y este supuesto es válido para la mayor parte de variables. Los grados día acumulados y los Joules acumulados tienen potencialmente efectos no linea les. La figura 6 muestra las variables normalizadas - se substrae la media y se divide para la desviación estándar - para permitir comparación directa.

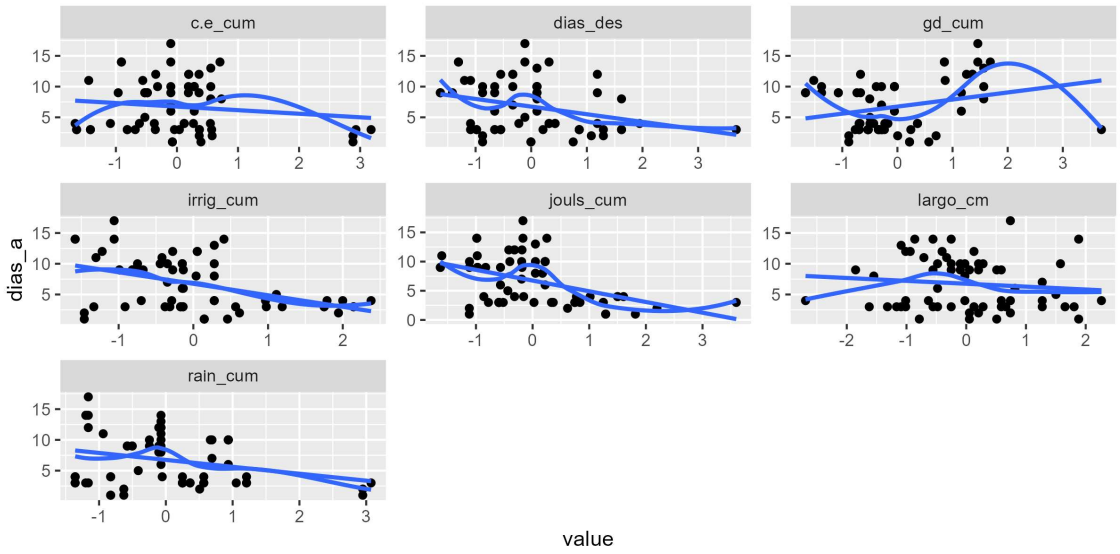

Figura 6: Relación entre predictores y la variable objetivo

Las variables están levemente correlacionadas con la variable objetivo como es evidente en la Figura 7. Para el propósito de generar un modelo predictivo se eligirán únicamente las 3 variables más correlacionadas y los grados día acumulados. Los grados día acumulados no se excluyen dada su notable prevalencia en la literatura. Es importante, sin embargo, notar que las unidades térmicas acumuladas hasta 5 lígulas no aparecen como correlacionadas con la data. Son la única variable correlacionada positivamente con los días restantes a la cosecha pero el coeficiente indica una relación más bien débil (coef. = 0.12). El coeficiente de correlación entre las variables de grados día y los días restantes a la cosecha es positiva lo que constituye un resultado contraintuitivo.

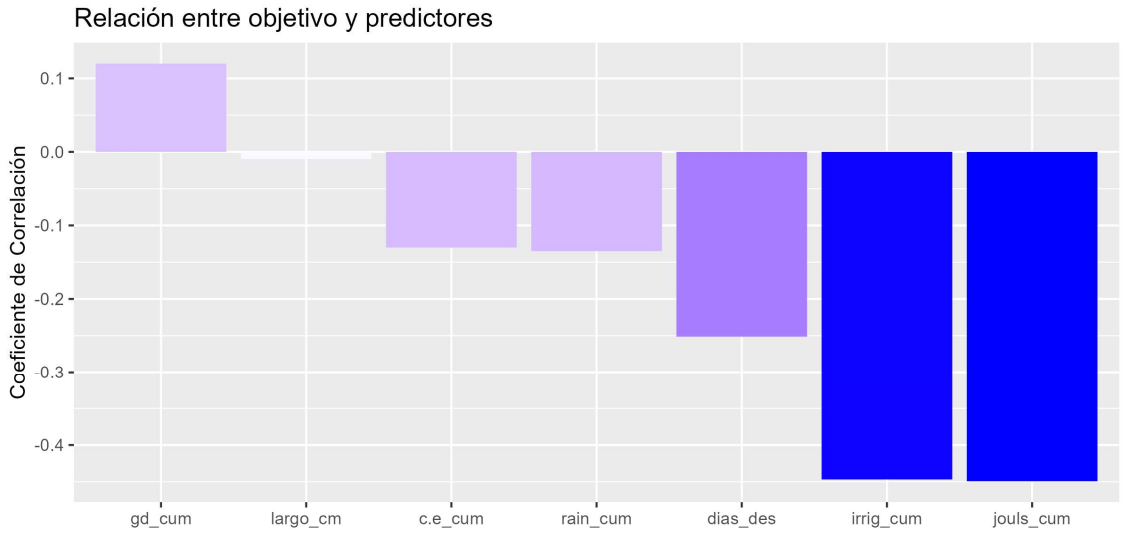

Figura 7 : Relación entre predictores y la variable objetivo

Para la modelación predictiva se emplearán modelos de regresión lineal. Para obtener un modelo predictivo se procederá en dos pasos fundamentales. En primer lugar, se entrenarán regresiones lineales comenzado con los 4 predictores principales y continuando con transformaciones de variables o inclusión de términos cuadráticos de ser necesario. El rasgo principal de la modelación predictiva, y que la distingue de la inferencia causal, es que el objetivo reside en encontrar un modelo que prediga acertadamente para observaciones fuera de muestra (Kuhn, 2013). En la práctica se han conformado consensos sobre en qué consiste la modelación predictiva y los pasos involucrados. Se sugiere una secuencia cronológica que comienza por una escisión del conjunto de datos en dos subconjuntos principales: el set de entrenamiento y el de prueba. A esto también se le conoce como data spending (Kuhn & Silge, 2022).

La construcción del modelo se hará únicamente con el conjunto de entrenamiento. Existen técnicas para evaluar y validar el desempeño del modelo restringiéndose al conjunto de entrenamiento llamadas técnicas de remuestreo. El remuestreo para estimar el desempeño de un modelo consiste en tomar subconjuntos para entrenar el modelo - subconjuntos de

análisis - y calcular métricas de desempeño con los datos restantes -subconjuntos de evaluación-, agregar estimaciones resultantes, usualmente mediante la media, y presentar una sola estadística (Kuhn & Johnson, 2013). Los parámetros de interés, α, corresponden a los siguientes modelos lineales:

$$\alpha \intercal X = \alpha 0 + \alpha 1\ (Joules) + \alpha 2\ (GDD) + \alpha 3\ (L/m2\ ) + \alpha 4\ (\text{Días Pinch})$$

$$\text{Días a Cosecha} = \alpha \intercal X + \epsilon\ (1)$$

$$\text{Días a Cosecha} = \alpha \intercal log(X) + \epsilon\ (2)$$

$$\text{Días a Cosecha} = \alpha \intercal X + \alpha 5 D(T\ rimestrei\ ) + \epsilon\ (3)$$

La técnica de remuestreo aquí utilizada será el bootstrap. Esta técnica consiste en generar conjuntos del mismo tamaño del conjunto de entrenamiento eligiendo de las observaciones allí contenidas, pero con remplazo. Esto implica que una misma muestra puede ser incluida en el nuevo conjunto más de una vez que será el conjunto de análisis. Aquellas muestras que fueron excluidas por azar del conjunto de análisis serán tomadas para el conjunto de evaluación. El lenguaje estadístico R elige 25 nuevas muestras por defecto. Las estadísticas y predicciones dentro del conjunto de entrenamiento son calculadas con 100 nuevos subconjuntos bootstrap.

Adicionalmente a los modelos lineales, una vez se distinga un conjunto de predictores que mejoren de forma definitiva la capacidad predictiva de un modelo lineal simple se recurrirá a una regresión poisson. La regresión poisson nos da el beneficio de obtener predicciones no negativas. Una regresión poisson requiere estimar los parámetros mediante el procedimiento de máxima verosimilitud. El algoritmo necesita de la especificación de una función de verosimilitud logarítmica. Esta función representa la probabilidad conjunta de observar la

distribución presente en los datos. Para cada observación, la regresión de poisson tomará la forma:

$$P(Y_i = y_i|X, \alpha) = \frac{e^{-\exp(X_i,\alpha)} \exp(x_i,\alpha')}{y_i!} \qquad (4)$$

La función de verosimilitud logarítmica se obtiene de multiplicar las funciones de verosimilitud individuales.

$$\log(a; y, X) = \log \Pi_{i=1}^{n} \frac{e^{-\exp\ (X_i\alpha')} \exp(x_i\alpha')}{y_i!} \quad (5)$$

RESULTADOS

El Cuadro 2 muestra los rasgos estadísticos de los predictores. Todas las variables están presentadas en niveles y los joules acumulados son presentados en miles. Nótese que las escalas de cada uno de los predictores son muy distintas. Una transformación puede ser utilizada para volverlos más homogeneos. De los coeficientes de variación sabemos que es nuestra variable objetivo, días a cosecha, la que presenta mayor dispersión alrededor del promedio mientras que días desde pinch es la variable que presenta menos variación alrededor de su media.

Cuadro 2: Variables: Estadísticas Descriptivas

Variable	Promedio	Mínimo	Máximo	Coeficiente de Variación
Días a Cosecha	6.75	1.00	17.00	62.27
Días desde Pinch	96.05	81.00	130.00	9.59
Largo de Tallo	68.30	33.00	98.00	19.21
Joules Acumulados	167.66	131.80	246.80	13.13
Grados Día Acumulados	1012.98	642.20	1836.15	21.96
Conductividad Eléctrica (Nutrientes)	125.28	97.50	178.24	13.30
Lluvía Acumulada (mm)	161.49	78.59	349.05	37.80
Litros/Metro Cuadrado Acumulados (Irrigación)	297.99	196.83	450.10	21.46

La Figura 8 muestra las gráficas de diagnóstico para la regresión lineal que incluye a los predictores: No. de días desde el pinch; Joules acumulados; grados día acumulados; litros por metros cuadrado de riego. Estos predictores fueron provistos a un modelo de regresión lineal en niveles (sin ninguna transformación) y se entrenaron con la data de la muestras bootstrap del conjunto de entrenamiento. Las predicciones corresponden a aquellas realizadas para los subconjuntos de evaluación generados por la técnica de remuestreo bootstrap.

El modelo lineal presenta errores que se distribuyen aproximadamente normal como se muestra en el panel del extremo derecho de la Figura 8. Sin embargo, el promedio R-cuadrado

de las predicciones para los conjuntos de evaluación es bajo (0.34) y parece existir evidencia de relaciones no lineales a partir de lo presentado en la gráficas de residuales vs predicciones. Esto también podría deberse a que contamos con una muestra pequeña, lo que implica estimaciones menos precisas. Observando la gráfica de R-cuadrado es fácil notar que para las muestras que llevan días a cosecha menores a 5 días nuestro modelo sobreestima los días restantes de forma casi sistemática. En cambio, para muestras con más de 10 días restantes el modelo subestima la variable objetivo. Las predicciones resultan de modelos entrenados en 100 conjuntos de análisis generados con la técnica bootstrap. Nótese que estos conjuntos de análisis tienen el mismo número de muestras que el conjunto de entrenamiento (47). El modelo final será entrenado en todo el conjunto de entrenamiento y evaluado en el conjunto de prueba.

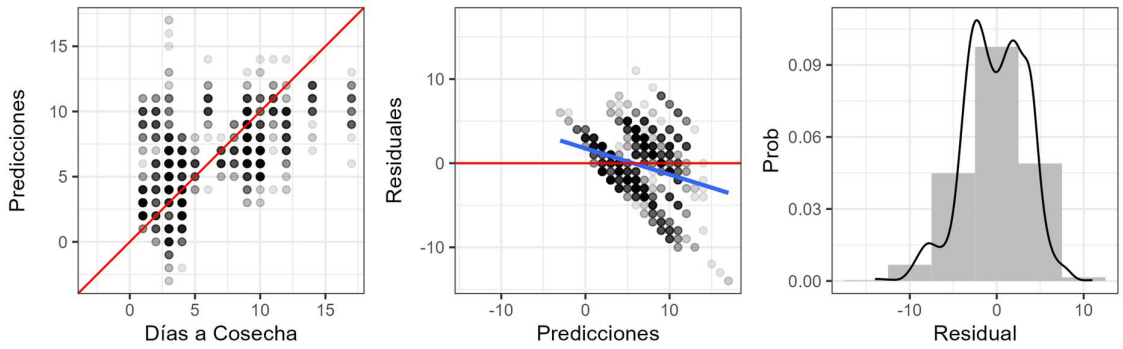

Figura 8: Gráficas de Diagnóstico. Gráfica de R-cuadrado; Gráfica de Residuales; Distribución de Residuales

Para cada uno de los predictores existe al menos una observación que dista más de 1.5 veces el rango intercuartílico al percentil 75 %, como se muestra en la Figura 9. Los valores atípicos tienen errores cuadráticos más grades en magnitud por lo que contribuyen más que proporcionalmente a suma de errores cuadráticos. Tomar logaritmos es útil para disminur la varianza de la data mientras se mantienen el mismo ordenamiento de las muestras. Esta transformación se aplicará a todo el grupo de predictores.

La Figura 10 contiene las mismas gráficas para un modelo de regresión lineal cuyos predictores son los logaritmos de las variables originales. Los residuales tienen una distribución que se apróxima más a una normal. Adicionalmente se reduce la varianza de los errores, aunque muy levemente.

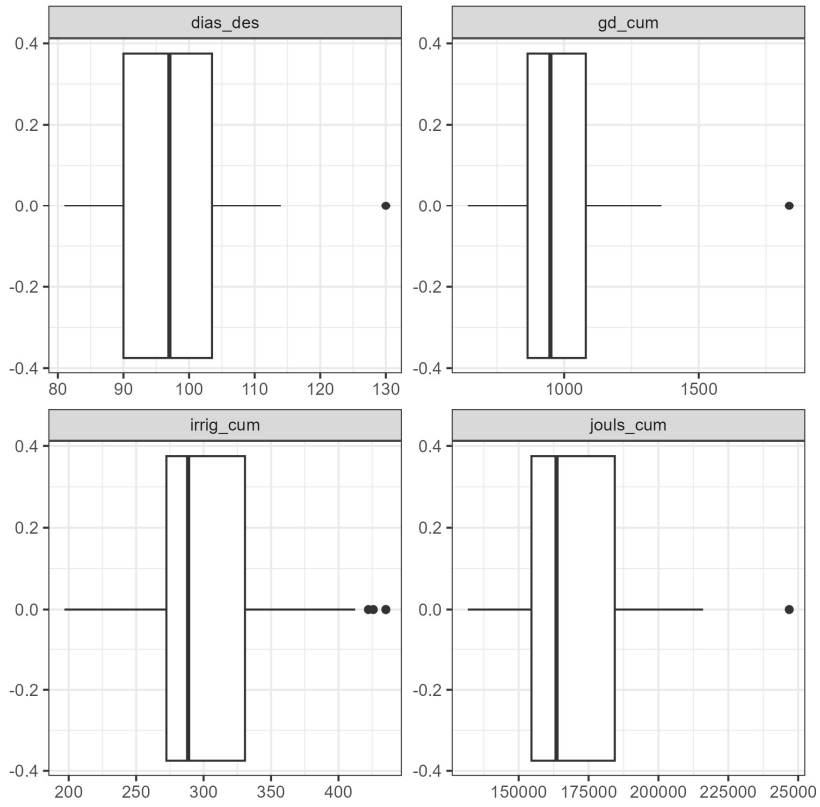

Figura 9: Distribuciones de los Predictores

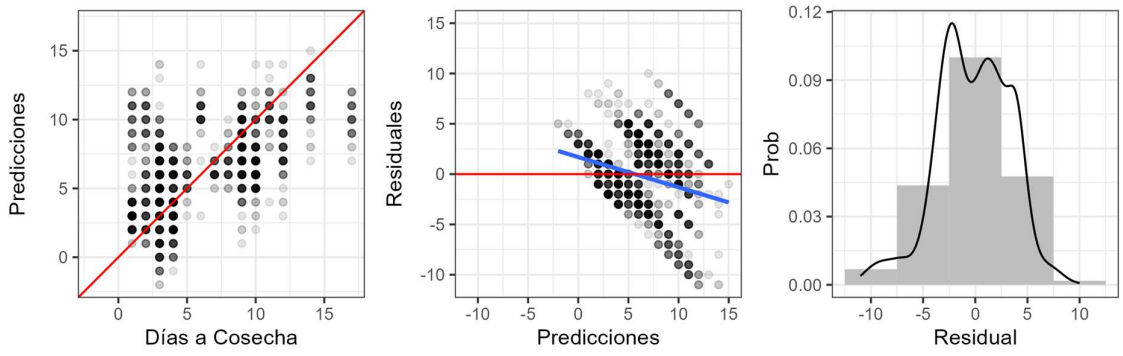

Figura 10: Gráficas de Diagnóstico. Gráfica de R-cuadrado; Gráfica de Residuales; Distribución de Residuales

Las interacciones entre variables son predictores candidatos para incluir en el modelo. Para explorar el potencial de las interacciones posibles se corrieron 12 regresiones lineales: 2 por cada par de variables. Un modelo consistía en únicamente el par de variables originales y captura sus efectos principales mientras que el otro contenía además su interacción. Los modelos se entrenaron en 25 nuevos subconjuntos generados con la técnica bootstrap. Aplicar métodos de remuestro nos permite obtener una distribución y así hacer pruebas de hipótesis. Los valores p reportados en el Cuadro 3 corresponden a una prueba t cuya hipótesis nula consiste en que la diferencia entre el RMSE del modelo con efectos principales y el RMSE del modelo con la interacción son iguales. La prueba es de cola superior y la hipótesis alternativa nos dice que el RMSE del modelo con interacciones es mayor al RMSE del modelo solo con efectos principales. Nótese que el RMSE es la medida que buscamos minimizar, por ende, buscamos rechazar la hipótesis nula para la mayoría de las interacciones de forma tal que las rechazamos y excluimos del modelo. A un nivel de significancia de referencia del 5 %, la hipótesis nula no es rechazada para las interacciones de grados día con irrigación (litros por metros cuadrado) y de irrigación con días desde el corte inicial, sin embargo, la inclusión de este par de interacciones solo redundaría y no mejoraría el poder predictivo. Esto se justifica puesto que todas las estimaciones puntuales de RMSE son menores para los modelos sin interacciones.

Cuadro 3: Cambios en RMSE para modelos simples de interacciones y no interacciones

Interactión	RMSE (Interacción)	RMSE (No interacción)	Cambio en RMSE	P Valor
Jouls:GD	3.74	3.35	-0.39	0.00
Jouls:L/m2	3.75	3.62	-0.14	0.01
Jouls:Días Desde Pinch	4.07	3.57	-0.50	0.01
GD:L/m2	3.81	3.78	-0.03	0.37
GD:Días Desde Pinch	5.32	4.07	-1.24	0.01
L/m2:Días Desde Pinch	3.81	3.75	-0.06	0.07

Los resultados anteriores sugieren que todas las interacciones excepto aquellas entre grados día con los litros irrigados por metro cuadrado y entre los días desde pinch y los litros irrigados por metro cuadrado empeoran el poder predictivo del modelo. Las estimaciones puntuales muestran que los RMSE para los modelos con interacción son siempre mayores a aquellos de los modelos con solo los efectos principales. Todas esas diferencias son significativas excepto para los dos casos ya mencionados. Adicionalmente, para estos casos, la diferencia no es significativa pero la inclusión de estos efectos de interacción no mejora ni empeora el modelo. Por ende, se prescinde de la inclusión de efectos de interacción.

Además de las variables ambientales contamos con las fechas en las que se realizó el Pinch. La Figura 11 muestra los gráficos para el diagnóstico de predicciones para un modelo con las variables en niveles y con dummies para los trimestres en los que comienza el ciclo. Estas dummies se incluyen en el afán de descubrir si existen cambios en las estimaciones controlando por el hecho de que distintos ciclos ocurrieron en distintas épocas del año. Podemos ver que no hay mejoría en la gráfica de R-cuadrado ni en la gráfica de residuales. En cambio, observamos que se comprometió la presunción de normalidad de errores en cierto grado como puede notarse en la distribución de los residuales. Los Cuadros 4 y 5 muestran las métricas obtenidas para cada modelo en el orden en el que fueron estimados. Podemos ver que la introducción de las variables dummy para el trimestre redujeron el RMSE mientras que aumentaron levemente el R-cuadrado. El incremento en el R-cuadrado es menor y es debido a la mera extensión del modelo por sobre nueva varianza explicada. Debido a esto es mejor prescindir de las variables indicador de cada trimestre.

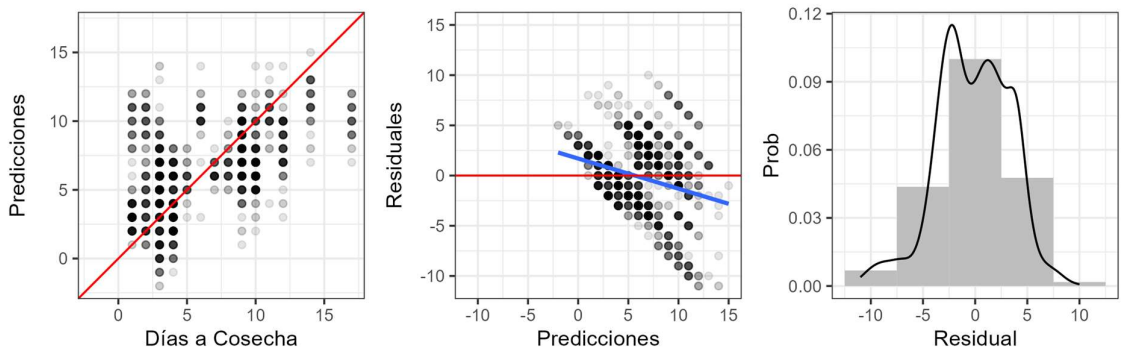

Figura 11: Gráficas de Diagnóstico. Gráfica de R-cuadrado; Gráfica de Residuales; Distribución de Residuales

Cuadro 4: RMSE: Tres Modelos para el Set de Entrenamiento

Modelo	RMSE	Intervalo de Confianza
Lineal en Niveles	3.43	(3.55,3.31)
Lineal Logaritmos	3.43	(3.56,3.31)
Lineal Dummies Trimestres	3.42	(3.55,3.29)

Cuadro 5: R-cuadrado: Tres Modelos para el Set de Entrenamiento

Modelo	R-cuadrado	Intervalo de Confianza
Lineal en Niveles	0.34	(0.37,0.31)
Lineal Logaritmos	0.33	(0.36,0.3)
Lineal Dummies Trimestres	0.35	(0.38,0.32)

Las estimaciones de cada una de las métricas de bondad de ajuste son el promedio de RMSE y R-cuadrado resultantes de las predicciones para los 100 conjuntos bootstrap de evaluación generados.

La regresión lineal presenta el problema de predicciones negativas en un contexto en el que no tienen sentido. Una solución simple es tomar la predicciones negativas y reportar 0 en lugar de los valores negativos. Otra alternativa es utilizar modelos que excluyan esas posibilidad desde el inicio. El modelo poisson es parte de la familia de modelos lineales

generalizados. Se estima mediante la técnica de máxima verosimilitud utilizando la distribución poisson como distribución subyacente y sus resultados se presentan en la Figura 12. Podemos observar enseguida que las predicciones negativas desaparecieron. Sin embargo, persiste el problema ya antes obtenido y que sugiere mala especificación del modelo y es que los residuales aumentan progresivamente cuando el modelo predice más de 10 días restante a la cosecha. El Cuadro 6 muestra las métricas del modelo Poisson cuyo desempeño no parece especialmente mejor tanto para el RMSE como para el R-cuadrado. Sin embargo, obtenemos predicciones únicamente positivas y que, en su mayoría, no son residuales mayores a 5 días. El Cuadro 7 nos muestra las proporciones correspondientes de los residuales del modelo Poisson menores a 5 días y de aquellos mayores o iguales a 5 días. La vasta mayoría se encuentra en la primera categoría (82 %).

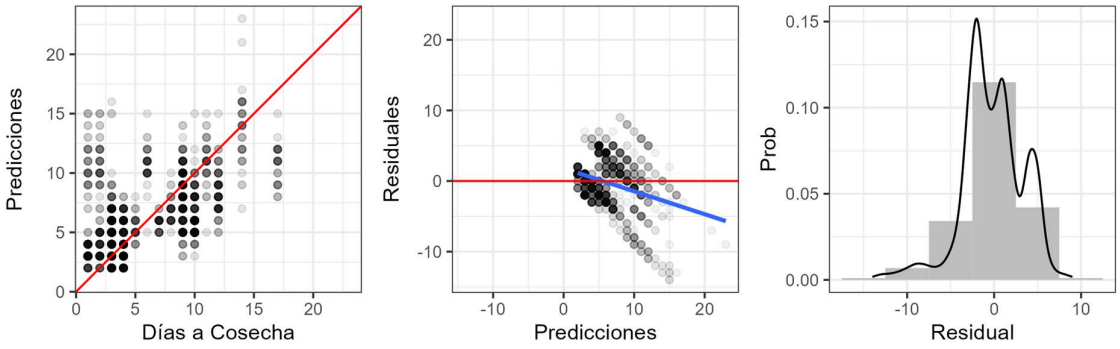

Figura 12: Gráficas de Diagnóstico. Gráfica de R-cuadrado; Gráfica de Residuales; Distribución de Residuales

Si bien el modelo poisson presenta la ventaja de no hacer predicciones negativas, su desempeño para las predicciones realizadas para los conjuntos de análisis, resultantes del remuestreo bootstrap, parece ser peor que el del modelo lineal canónico. Es necesario recordar que por lo pronto se ha trabajado únicamente con el conjunto de entrenamiento de donde se tomaron nuevas muestras y crearon subconjuntos para entrever el desempeño antes de hacer la prueba final con el conjunto de prueba.

Cuadro 6: Métricas del Modelo Poisson

Métrica	Estimación	No. Muestras	Error Std
rmse	3.44	100	0.07
rsq	0.34	100	0.02

Cuadro 7: Proporción de Residuales

Residual (>=5)	Conteo	Proporción
Falso	1426	0.83
Verdadero	294	0.17

La exploración realizada determinó que el mejor modelo contendría alguna transformación de las variables originales sin incluir interacciones y sin incluir términos que controles por efectos de temporalidad. Como paso final se entrenará un modelo de regresión con los 4 predictores en logaritmos y también un modelo poisson. Los modelos cuyos coeficientes fueron estimados con todo el conjunto de entrenamiento serán comparados con el conjunto de prueba. A continuación se presentará el Cuadro 8. Observamos un revés en el desempeño siendo el modelo poisson que presenta mejores predicciones por fuera de muestra. Ambos modelos presentan mejoría con respecto a las predicciones realizadas con subconjuntos del modelo de entrenamiento únicamente.

Cuadro 8: Métricas de Modelos Lineal Clásico y Poisson

Modelo	rmse	rsq	mae	mape
Lineal	2.92	0.57	2.61	43.46
Poisson	2.73	0.62	2.29	38.23

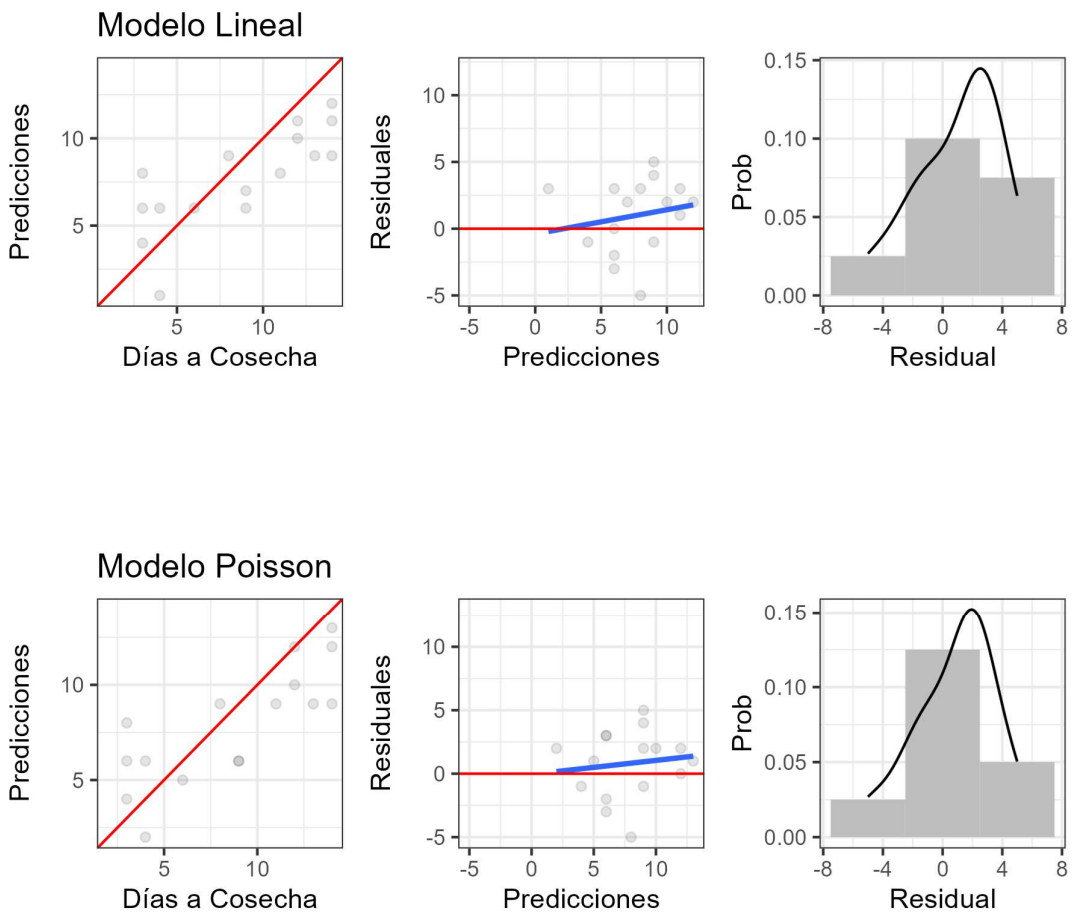

Figura 13: Gráficas de Diagnóstico para Modelos Lineal y Poisson. Gráfica de Rcuadrado; Gráfica de Residuales; Distribución de Residuales

CONCLUSIONES

Los modelos aquí provistos presentan un poder predictivo moderado más no deficiente. Ambos modelos pertenecen a la familia de modelos lineales generalizados. El primero es un modelo lineal clásico mientras que el modelo poisson, aunque todavía lineal en parámetros, estima los parámetros de una función de probabilidad conjunta. El RMSE, que puede ser interpretado como la distancia media entre las predicciones del modelo y la variable dependiente observada, se encuentra dentro del intervalo entre 3.31 y 3.56 días para el modelo lineal y entre 3.30 y 3.57 para el modelo poisson. Las estimaciones puntuales sobre el conjunto de prueba fueron todavía menores (Cuadro 8). Los modelos utilizan un subconjuntos de las variables disponibles pues solo se consideró aquellas lo suficientemente correlacionadas con la variable dependiente. Dados los datos actuales, ninguna interacción entre variables resultó en mejoras en la capacidad predictiva. Tampoco fueron de utilidad las variables dummy que señalaban el trimestre en el que inició el ciclo y con las que se buscaba controlar por la estacionalidad en variables atmosféricas. No obstante, se avizoran mejoras posibles en la medida que mejore la cantidad y calidad de los datos. Por ejemplo, un mayor número de muestras ayuda a disminuir la varianza del modelo y podrían desvelar correlaciones que ahora no están presentes en los datos. También es posible reformular el problema y entrenar un modelo poisson con los mismos datos, pero respondiendo una pregunta distinta. Por ejemplo, el número de rosas que se llegarán a la cosecha dentro de un horizonte de tiempo dado. Adicionalmente, modelos y algoritmos de aprendizaje más flexibles podrían capturar patrones no lineales en los datos. Estas mejoras serán posibles en las fases siguientes del programa actualmente en marcha para encontrar modelos predictivos de los días restantes a la cosecha. Más datos de las mismas variables son importantes para de manera definitiva vislumbrar más claramente la importancia de estas. Por ejemplo, las unidades térmicas o grados día acumulados son un predictor estándar y es siempre

confirmado como uno de los que cuenta con mayor poder predictivo. Sin embargo, en los datos disponibles los grados día constituían de los predictores con menos correlación con la variable objetivo. Esto puede deberse a mera aleatoriedad que se disiparía mediante incrementos en los datos que se posean. Adicionalmente, más variables podrían capturar más de la variación en la variable objetivo. Ejemplos para el problema que nos interesa son el ancho del botón de la rosa o la temperatura referenciada espacialmente dentro del invernadero. El levantamiento de mejores datos es la siguiente fase de proyecto. Monroy et al (2001) muestran que la evolución de los grados día es distinta según su ubicación dentro del invernadero, siendo los puntos centrales los más calientes. Estas son direcciones en las que estos modelos pueden ser mejorados y refinados. Inclusive la especificación misma de los modelos puede estar sujeta a cambios pues a medida que se generan más datos, otros algoritmos más flexibles que requieran de una mayor cantidad de datos pueden 19 ser puestos a prueba.

REFERENCIAS BIBLIOGRÁFICAS

Carlson, J.D. & Hancock Jr. (1991). A methodology for determining suitable heatunits requirements for harvest of highbush blueberry. J. Am. Soc. Hortic. Sci. 116 (5), 774–779.

Baptista, M.C. & Oliveira, P.B. & Lopes da Fonseca L. & Oliveira, C.M. (2006). Early ripening of southern highbush blueberries under mild winter conditions. Acta Hortic. 715, 191–196.

Macias, M, & Villalta, E. (2015). Factibilidad de una integración EcuatorianaColombiana para la comercialización de flores dirigidas al mercado ruso. Tesis de grado previo al título de Ingeniería en Comercio y Finanzas Internacionales Bilingue,Universidad Catolica de Santiago de Guayaquil, Facultad de Especialidades Empresariales, 111-112. http://repositorio.ucsg.edu.ec/handle/3317/4902

Muñoz, C & Salvo, S & Huircán, J. (2012). Prediction of harvest start date in highbush blueberry using time series regression models with correlated errors. Sci. Hortic, 138, 165-170.

Pérez, I., Cure, J. R., & Monroy, N. (2002). Modelo de predicción y manejo de cultivos de rosas. Revista de Ingeniería, (15), 18-22.

Rodríguez, F., Berenguel, M., Guzmán, J.L., Ramírez-Arias, A. (2015). The Greenhouse Dynamical System. In: Modeling and Control of Greenhouse Crop Growth. Advances in Industrial Control. Springer, Cham. https://doi.org/10.1007/978-3-319-11134-6_2

Kuhn, M, & K Johnson. (2013). Applied Predictive Modeling. Springer.

Kuhn, M, & K Johnson. (2020). Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press.

Kuhn, M, & Silge, J. (2022). Tidy Modeling with R: A Framework for Modeling in the Tidyverse. Oreilly Media.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. ´

Vila, J. (2009). Modelo de pr Modelo de proyección par ección para la pr a la producción de r oducción de rosas, basado en las osas, basado en las curvas de crecimiento de las plantas. Retrieved from https://ciencia.lasalle.edu.co/ administracion_agronegocios/200/newline

Zou, H, & Hastie, T. (2005). Regularization and variable selection via the elastic net. Royal Statistical Society, 67, 301–320.