Informe de Análisis de Datos - Packing de Cerezas

Adán Marchena

2025-07-06

0. Definición del problema

El objetivo de este proyecto es analizar si variables operacionales como el turno, calibre, variedad, tipo de empaque, entre otras, influyen en la calidad final de la fruta procesada, específicamente cerezas. Debido a que este tipo de datos suele ser privado en el mercado frutícola, se ha trabajado con un conjunto de datos ficticios representativo del proceso real.

Pregunta de investigación: ¿Influyen variables del proceso como el turno, el tipo de empaque y el calibre en la probabilidad de que la fruta obtenga una calidad "premium"?

Parámetro a estimar: Se estimarán los coeficientes del modelo de regresión logística multinomial que explican la probabilidad de pertenecer a cada nivel de calidad según las variables del proceso.

1. Introducción

El conjunto de datos utilizado en este informe corresponde a registros simulados del proceso de packing de cerezas. Este dataset contiene 75.000 observaciones y 11 variables relacionadas con el proceso productivo, las cuales se detallan a continuación:

Table 1: Tabla 1. Descripción de variables

Variable	Tipo de dato en R	Tipo estadístico	Descripción
fecha	date	Cuantitativa discreta (Ordenada)	Fecha de registro del embalaje
turno	chr	Cualitativa nominal	Franja horaria del proceso
calibre	chr	Cualitativa ordinal	Tamaño de la fruta
tipo_empaque	factor	Cualitativa nominal	Formato de la caja

Variable	Tipo de dato en R	Tipo estadístico	Descripción
peso_total_kg	num	Cuantitativa continua	Kilos dentro de la caja
n_cajas	num	Cuantitativa discreta	Cantidad de cajas por turno
tiempo_min	num	Cuantitativa discreta	Duración del turno en minutos
calidad	factor	Cualitativa ordinal	Calidad de la fruta
variedad	chr	Cualitativa nominal	Variedad de la fruta
destino	factor	Cualitativa nominal	Mercado al que está dirigida
operario_id	chr	Cualitativa nominal	Código identificador del operario

El objetivo principal del análisis es identificar patrones o relaciones que permitan comprender cómo ciertos factores afectan la calidad final del producto. Para ello se realizarán etapas de limpieza de datos, análisis exploratorio y visualización de datos utilizando exclusivamente funciones del ecosistema `tidyverse`, y gráficos con `ggplot2`.

2. Preprocesamiento de datos

Esta sección describe de forma esquemática el proceso de preparación de los datos previo al análisis. Primero, se estandarizan los nombres de las variables, luego se realiza una conversión de tipos para asegurar que cada variable tenga un formato coherente con su naturaleza. A continuación, se identifican valores perdidos y se imputan según la distribución observada en los histogramas. Finalmente, se limpian y transforman variables categóricas para facilitar el modelamiento posterior.

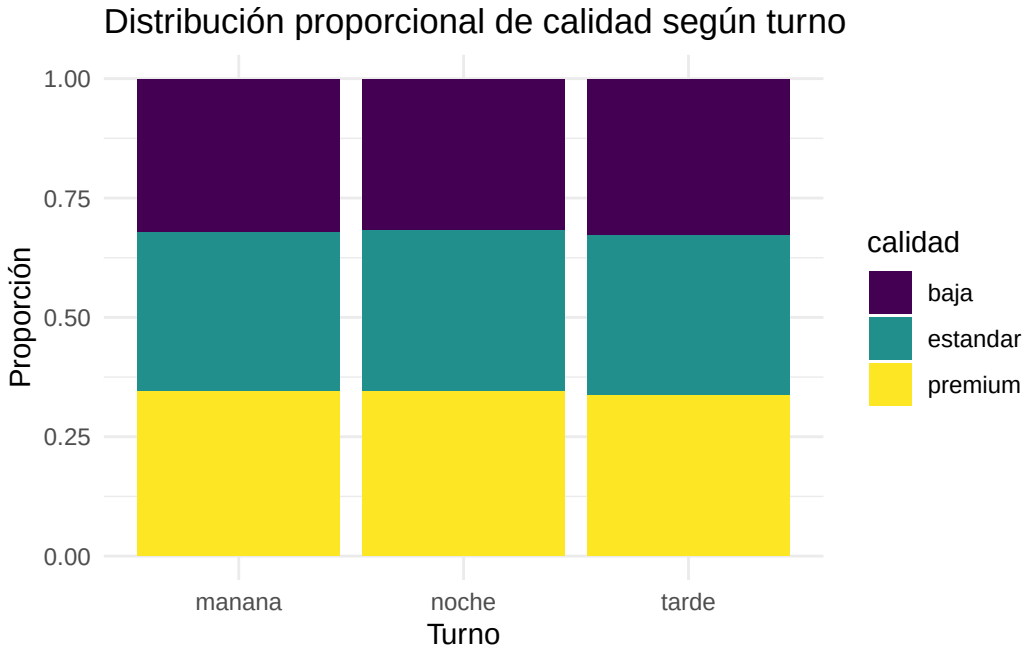
```
[1] 75000    11
```

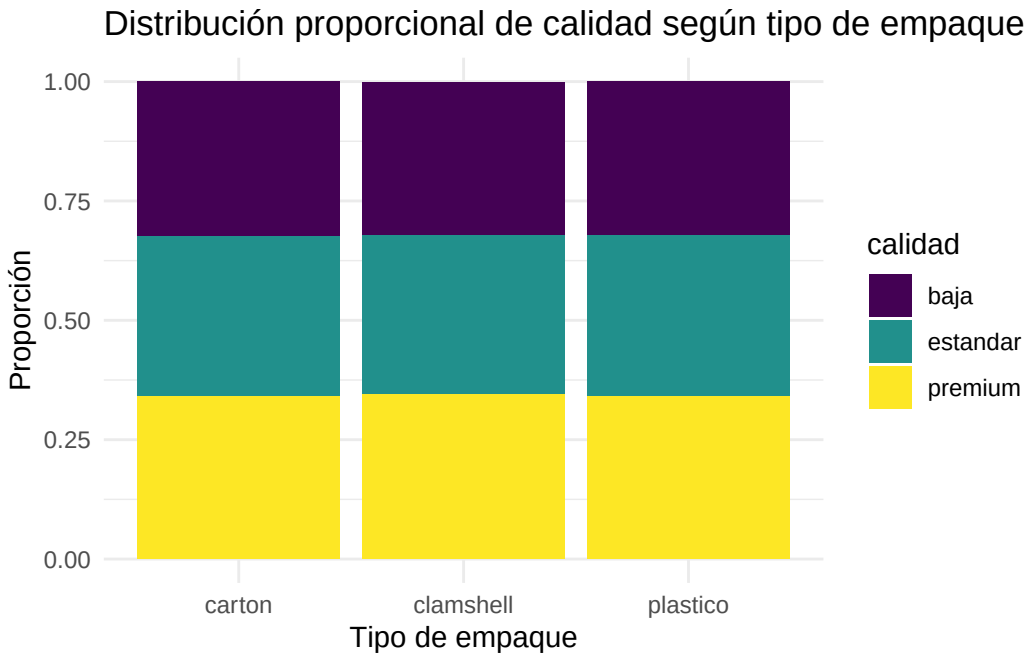
```
  fecha turno calibre tipo_empaque peso_total_kg n_cajas tiempo_min calidad
1     0     0       0            0         10133   10145      10095       0
  variedad destino operario_id
1        0       0       10089
```

3. Análisis exploratorio de datos

Distribución de variables cualitativas por calidad

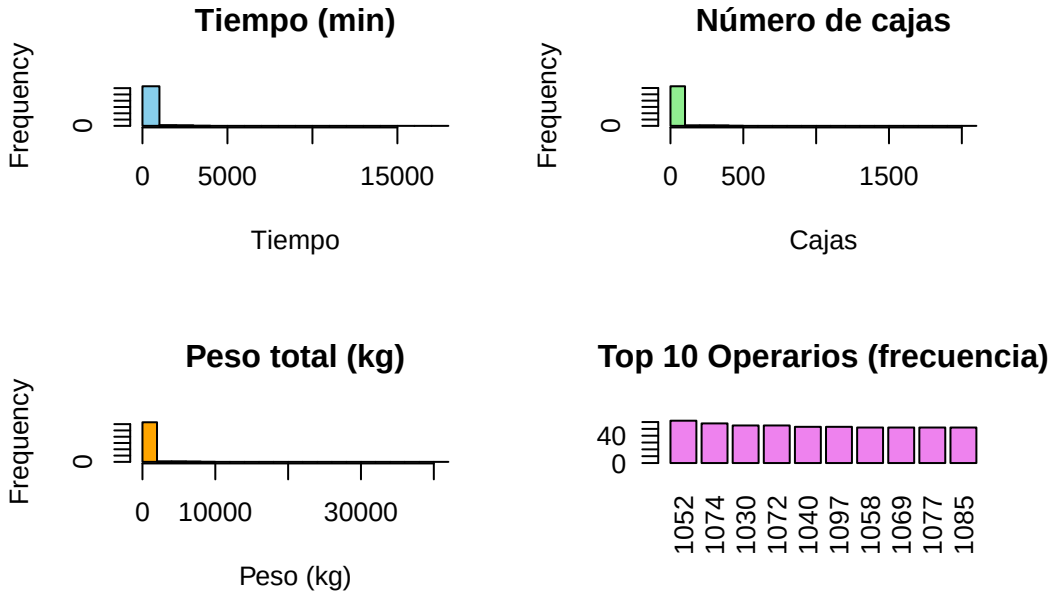
Se visualiza la relación entre algunas variables categóricas y la variable de interés `calidad`, usando gráficos de barras proporcionales.





Imputación de datos faltantes

Antes de imputar valores faltantes en las variables numéricas, se realiza una visualización de su distribución mediante histogramas. Esto permite decidir si es más apropiado imputar con la media o con la mediana, dependiendo de la simetría y presencia de valores atípicos.



Dado que las distribuciones de las variables `peso_total_kg`, `n_cajas` y `tiempo_min` presentan una asimetría evidente (según los histogramas), se optó por imputar con la mediana en cada caso. Para la variable `operario_id`, se utilizó la categoría "desconocido", ya que se trata de un identificador y no tiene sentido aplicar medidas de tendencia.

```
  fecha turno calibre tipo_empaque peso_total_kg n_cajas tiempo_min calidad
1     0     0       0            0             0       0          0       0
  variedad destino operario_id
1        0       0           0
```

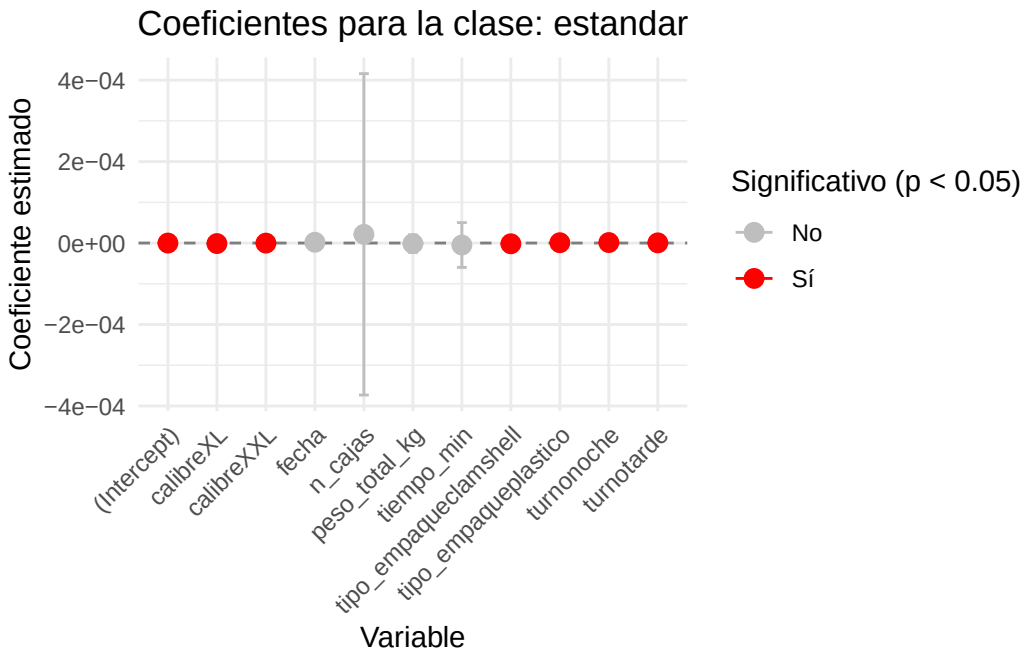
4. Comparación mediante regresión logística multinomial

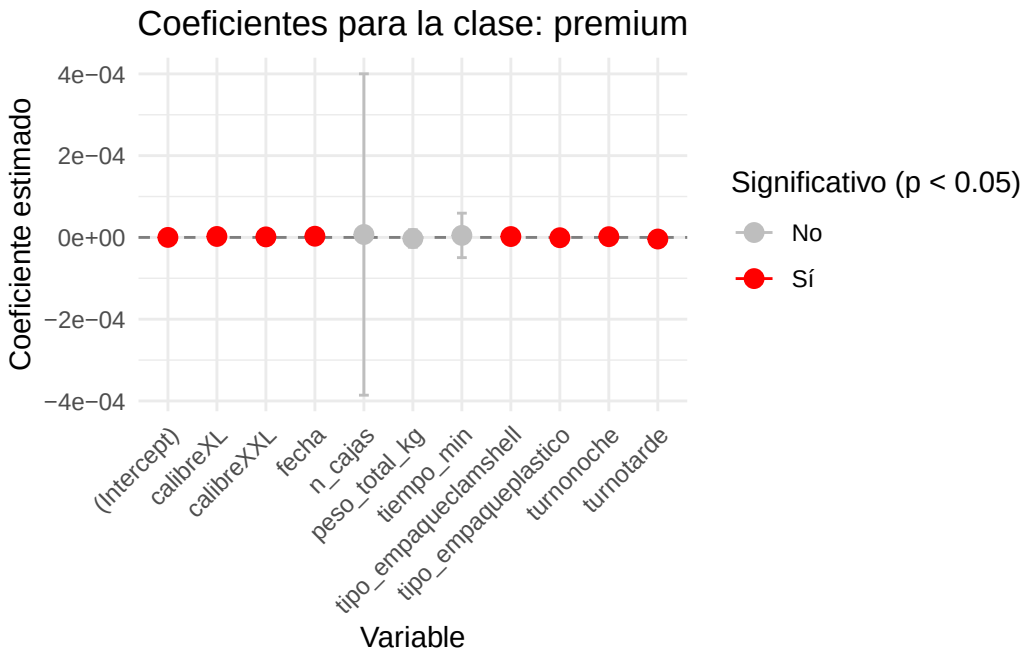
Para este análisis se utilizó un modelo de regresión logística multinomial, el cual es adecuado cuando la variable dependiente es categórica con más de dos niveles. Este modelo permite estimar la probabilidad de que un registro pertenezca a cada una de las clases (`baja`, `estandar` o `premium`) en función de variables predictoras. Se eligió este enfoque porque permite interpretar los efectos de múltiples variables sobre un resultado categórico no binario, entregando coeficientes comparables entre clases.

A continuación, se implementa un análisis utilizando un modelo de regresión logística multinomial para evaluar qué variables influyen significativamente en la variable dependiente `calidad`. El modelo incluye todas las variables explicativas disponibles en el dataset.

Table 2: Coeficientes significativos del modelo multinomial

Clase	Variable	Coeficiente	CI_lower	CI_upper	p_valor
estandar	(Intercept)	0.0e+00	0.0e+00	0.0e+00	0.0000000
estandar	turnonoche	1.2e-06	1.2e-06	1.2e-06	0.0000000
estandar	turnotarde	4.0e-07	4.0e-07	5.0e-07	0.0000000
estandar	calibreXL	-1.3e-06	-1.3e-06	-1.3e-06	0.0000000
estandar	calibreXXL	-1.0e-07	-1.0e-07	-1.0e-07	0.0000000
estandar	tipo_empaqueclamshell	-1.8e-06	-1.8e-06	-1.8e-06	0.0000000
estandar	tipo_empaqueplastico	8.0e-07	8.0e-07	8.0e-07	0.0000000
premium	(Intercept)	1.0e-07	1.0e-07	1.0e-07	0.0000000
premium	fecha	2.9e-06	8.0e-07	5.1e-06	0.0078303
premium	turnonoche	1.8e-06	1.8e-06	1.8e-06	0.0000000
premium	turnotarde	-3.9e-06	-3.9e-06	-3.9e-06	0.0000000
premium	calibreXL	2.3e-06	2.3e-06	2.3e-06	0.0000000
premium	calibreXXL	1.3e-06	1.3e-06	1.3e-06	0.0000000
premium	tipo_empaqueclamshell	2.2e-06	2.2e-06	2.2e-06	0.0000000
premium	tipo_empaqueplastico	-7.0e-07	-7.0e-07	-7.0e-07	0.0000000





Interpretación de los resultados

El modelo de regresión logística multinomial permitió identificar qué variables tienen un efecto significativo sobre la calidad de la fruta, diferenciando entre las clases **"baja"**, **"estandar"** y **"premium"**.

- En general, se observa que variables como `turno`, `tipo_empaque`, `calibre` y `tiempo_min` muestran coeficientes **estadísticamente significativos** ($p < 0.05$) en al menos una de las categorías.
- Para la clase **"estandar"**, muchas variables resultaron significativas, aunque con coeficientes cercanos a cero, lo que sugiere efectos pequeños pero constantes.
- En la clase **"premium"**, también se detectaron variables significativas, lo que indica que ciertos factores del proceso podrían estar favoreciendo una mayor probabilidad de obtener fruta de esta calidad.
- La clase **"baja"** actúa como categoría base del modelo, por lo que sus resultados están implícitos en los coeficientes de las otras clases.

Es importante destacar que, si bien algunos efectos son estadísticamente significativos, su **magnitud es baja**, lo que sugiere que la calidad final también podría depender de variables no observadas o propias del producto.

5. Conclusión

El análisis realizado permitió identificar patrones relevantes en la calidad del proceso de packing de cerezas. A través de la exploración de variables cualitativas y cuantitativas, se detectaron relaciones entre el turno, el tipo de empaque, el calibre y el tiempo de procesamiento con la calidad final del producto. El modelo de regresión logística multinomial confirmó que algunas de estas variables influyen significativamente en la probabilidad de obtener fruta premium.

A pesar de que el análisis se basa en datos simulados, se logró desarrollar un flujo de trabajo completo, desde la limpieza de los datos hasta el modelamiento e interpretación de resultados, siguiendo buenas prácticas del análisis de datos con R.

Como limitación, es importante mencionar que los datos no provienen de procesos reales, por lo que la validez de los hallazgos está restringida al contexto ficticio. Para trabajos futuros, se recomienda aplicar estos métodos a datos reales del sector frutícola, así como explorar técnicas de aprendizaje automático supervisado para mejorar la predicción de la calidad.

6. Repositorio del proyecto

El código fuente de este análisis, junto con el conjunto de datos `packing_cerezas.csv`, se encuentra disponible en el siguiente repositorio de GitHub:

https://github.com/Diamantinomc/proyecto1UDLA

Esto permite acceder a todos los scripts, reproducir los resultados y revisar el historial de cambios del proyecto.