Universidad del Valle de Guatemala
Minería de Datos
S20

Hoja de Trabajo #3

Clasificación

Javier Chavez 21016
Andres Quezada 21085

Parte 1 – Análisis exploratorio de datos

1. Haga una exploración rápida de sus datos para ello, haga un resumen del conjunto de datos.

   Descripción de las variables:

   ID: Identificador único para cada manzana.
   Tamaño: Tamaño de la manzana (normalizado).
   Peso: Peso de la manzana (normalizado).
   Dulzura: Nivel de dulzura de la manzana (normalizado).
   Textura: Textura de la manzana (normalizada).
   Humedad: Nivel de humedad de la manzana (normalizado).
   Madurez: Nivel de madurez de la manzana (normalizado).
   Acidez: Nivel de acidez de la manzana (normalizado).
   Calidad: Calidad de la manzana, puede ser "buena" o "mala". (Variable objetivo)

   Describe:

```
Resumen estadístico de las variables numéricas:
                ID       Tamaño         Peso      Dulzura      Textura  \
count  4000.000000  4000.000000  4000.000000  4000.000000  4000.000000
mean   1999.500000    -0.503015    -0.989547    -0.470479     0.985478
std    1154.844867     1.928059     1.602507     1.943441     1.402757
min       0.000000    -7.151703    -7.149848    -6.894485    -6.055058
25%     999.750000    -1.816765    -2.011770    -1.738425     0.062764
50%    1999.500000    -0.513703    -0.984736    -0.504758     0.998249
75%    2999.250000     0.805526     0.030976     0.801922     1.894234
max    3999.000000     6.406367     5.790714     6.374916     7.619852
```

```
           Humedad      Madurez
count  4000.000000  4000.000000
mean      0.512118     0.498277
std       1.930286     1.874427
min      -5.961897    -5.864599
25%      -0.801286    -0.771677
50%       0.534219     0.503445
75%       1.835976     1.766212
max       7.364403     7.237837
```

Info:

```
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   ID       4000 non-null   float64
 1   Tamaño   4000 non-null   float64
 2   Peso     4000 non-null   float64
 3   Dulzura  4000 non-null   float64
 4   Textura  4000 non-null   float64
 5   Humedad  4000 non-null   float64
 6   Madurez  4000 non-null   float64
 7   Acidez   4001 non-null   object 
 8   Calidad  4000 non-null   object 
dtypes: float64(7), object(2)
```

Shape:

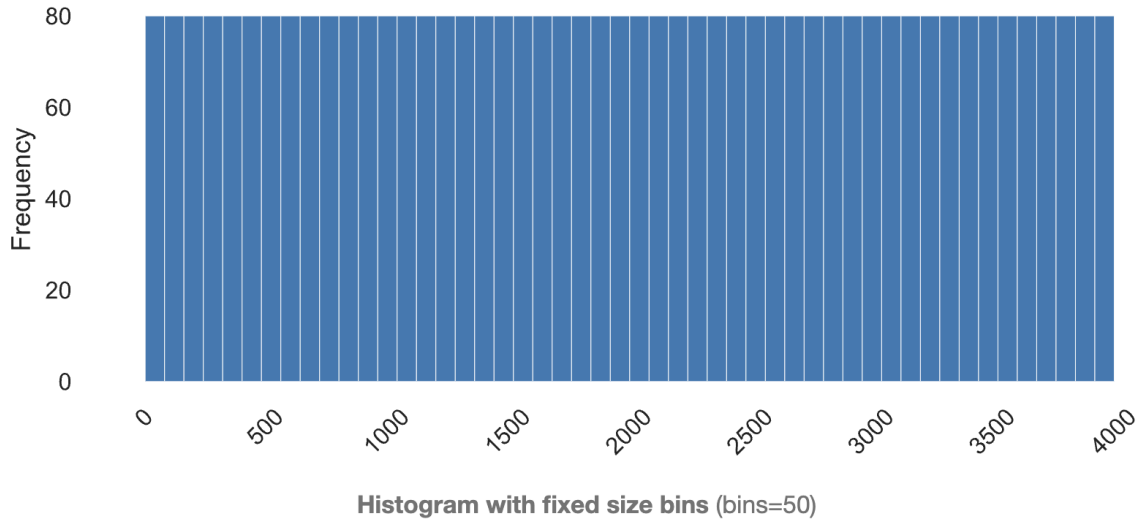
```
Forma del conjunto de datos:
(4001, 9)
```

2. Determine el tipo de cada una de las variables.

Variable	Tipo
ID	Cualitativa o categórica
Tamaño	Cuantitativa continua
Peso	Cuantitativa continua
Dulzura	Cuantitativa continua
Textura	Cuantitativa continua
Humedad	Cuantitativa continua
Madurez	Cuantitativa continua
Acidez	Cuantitativa continua
Calidad	Cualitativa o categórica

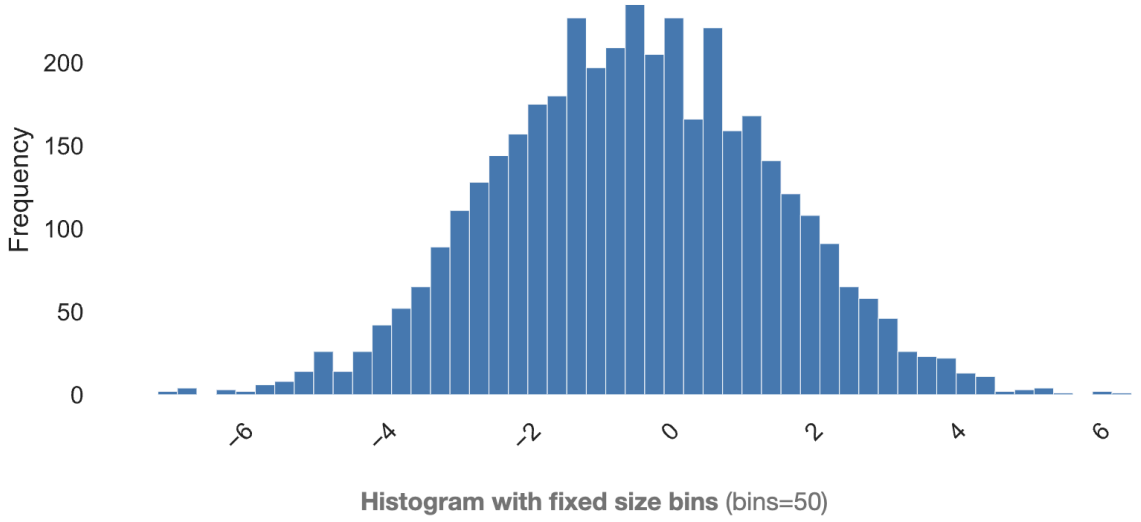
3. Incluya los gráficos exploratorios siendo consecuentes con el tipo de variable que están representando.

ID:



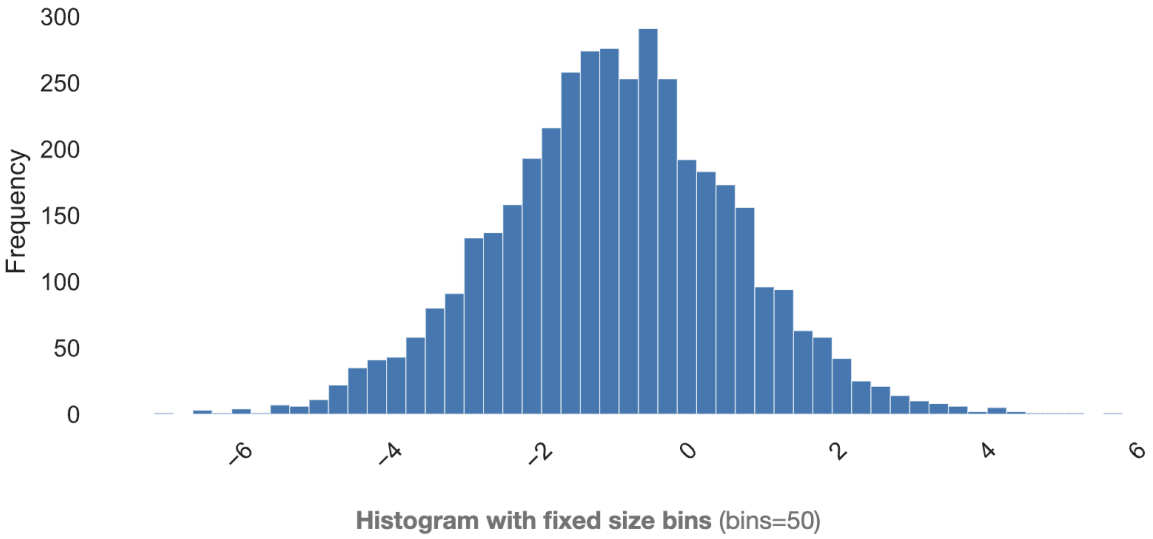
Esto es un gráfico uniformemente distribuido por ser un id incremental.

Tamaño:



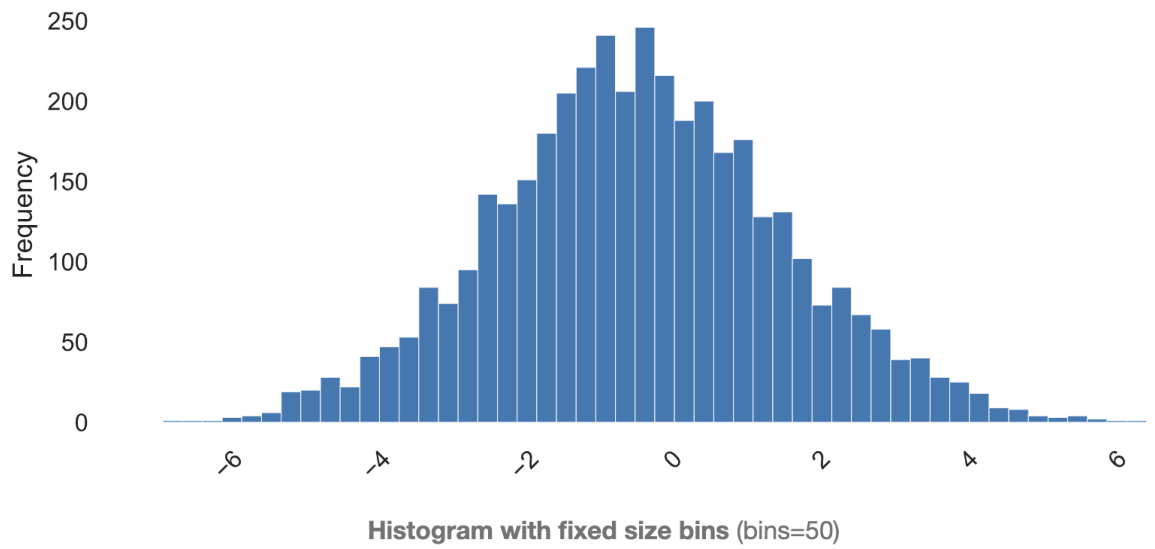
Observamos que la distribución es simétrica alrededor de su punto central. Esto significa que la misma cantidad de datos se encuentra tanto a la derecha como a la izquierda de la media.

Peso:



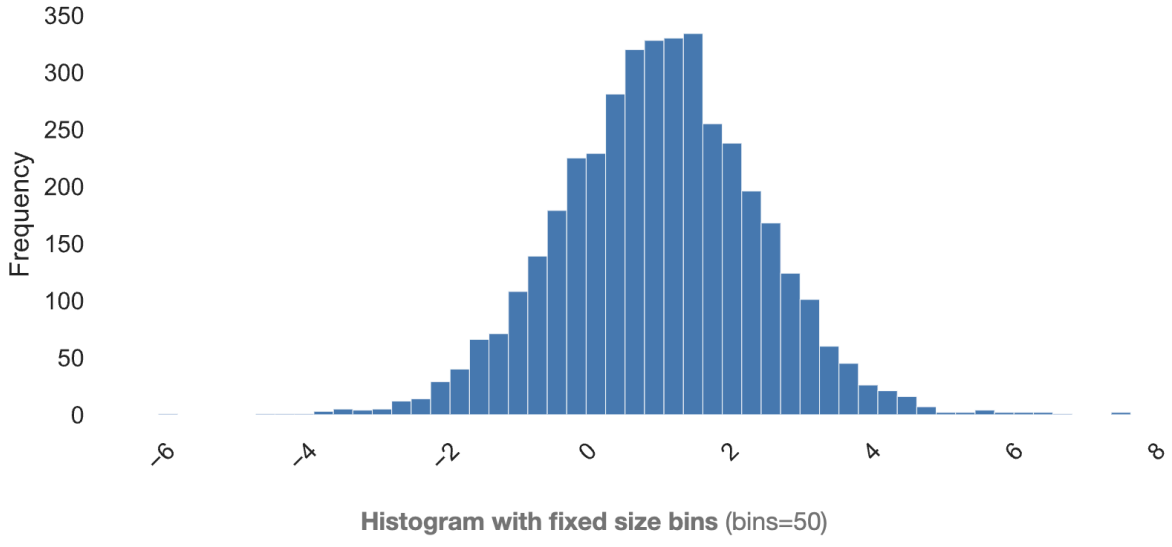
Este sugiere que la mayoría de las personas tendrían pesos cercanos al promedio, lo que nos parece interesante es observar que hay pesos negativos y el promedio es -0.98

Dulzura:



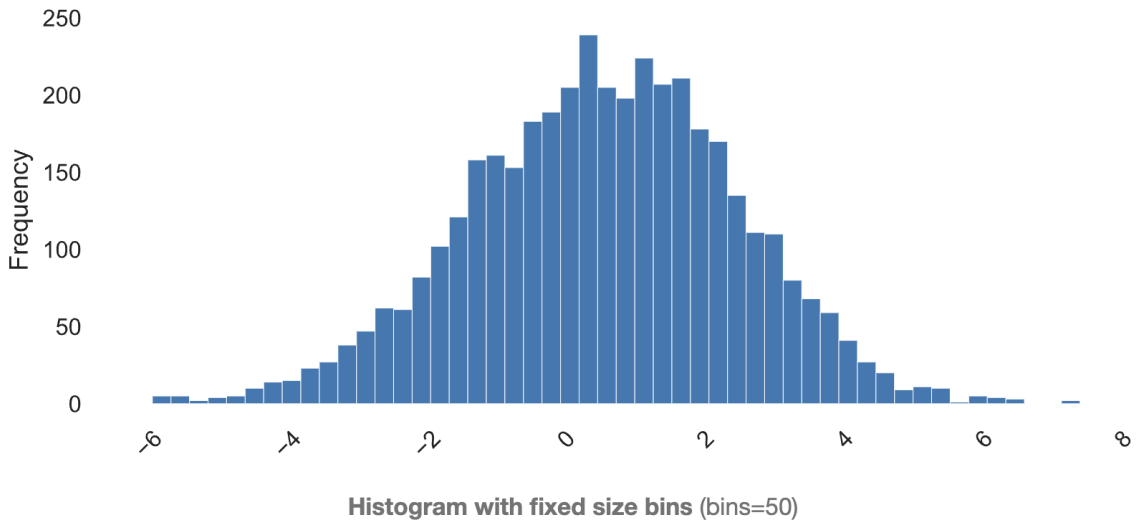
Lla distribución normal también proporciona información sobre la variabilidad de los niveles de dulzura, ya que la mayoría de los productos caen dentro de un rango predecible en torno a la media, mientras que los extremos de la campana representan los productos que se desvían significativamente de la norma en términos de dulzura.

Textura:



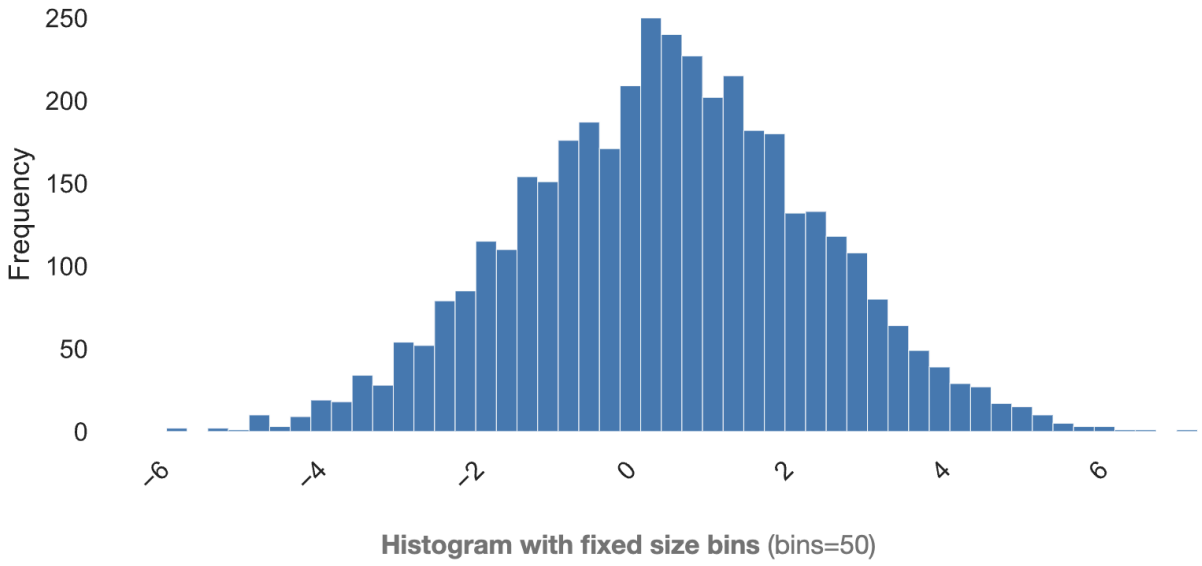
La distribución de textura de las manzanas sugiere que la mayoría de las manzanas en el conjunto de datos exhiben una textura típica, con una suave curva de campana que se centra alrededor de la media. Esto implica que la mayoría de las manzanas tienen una textura común, mientras que las que se desvían de esta norma son menos frecuentes.

Humedad:



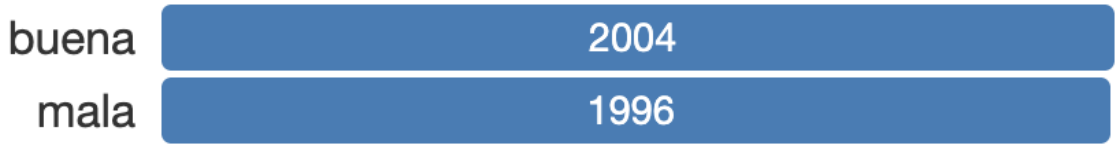
En el gráfico se puede observar que la mayoría de las manzanas tienen niveles de humedad que se agrupan alrededor de un valor central, como indica la forma de campana característica de la distribución normal. Esto indica que la humedad en las manzanas sigue un patrón predecible, con la mayoría de las manzanas presentando niveles de humedad cercanos a la media.

Madurez:



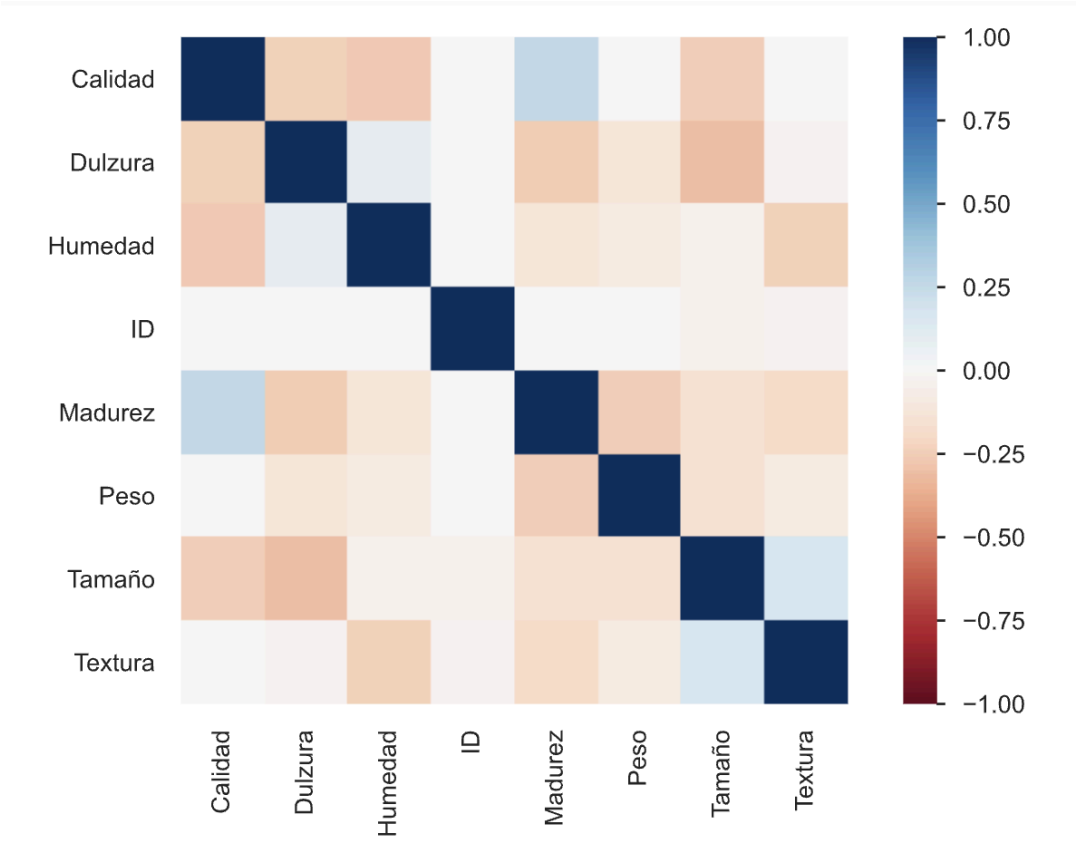
Este gráfico revela que la mayoría de las manzanas en el conjunto de datos tienen un nivel de madurez típico, como lo sugiere la forma de campana en el gráfico. Esto indica que la madurez de las manzanas tiende a seguir una distribución uniforme alrededor de la media, con menos manzanas mostrando niveles de madurez extremadamente bajos o altos.

Calidad:



Este gráfico nos indica que las manzanas están teniendo algún problema de calidad ya que se esperaba un mejor resultado siendo la mayoría en calidad buena.

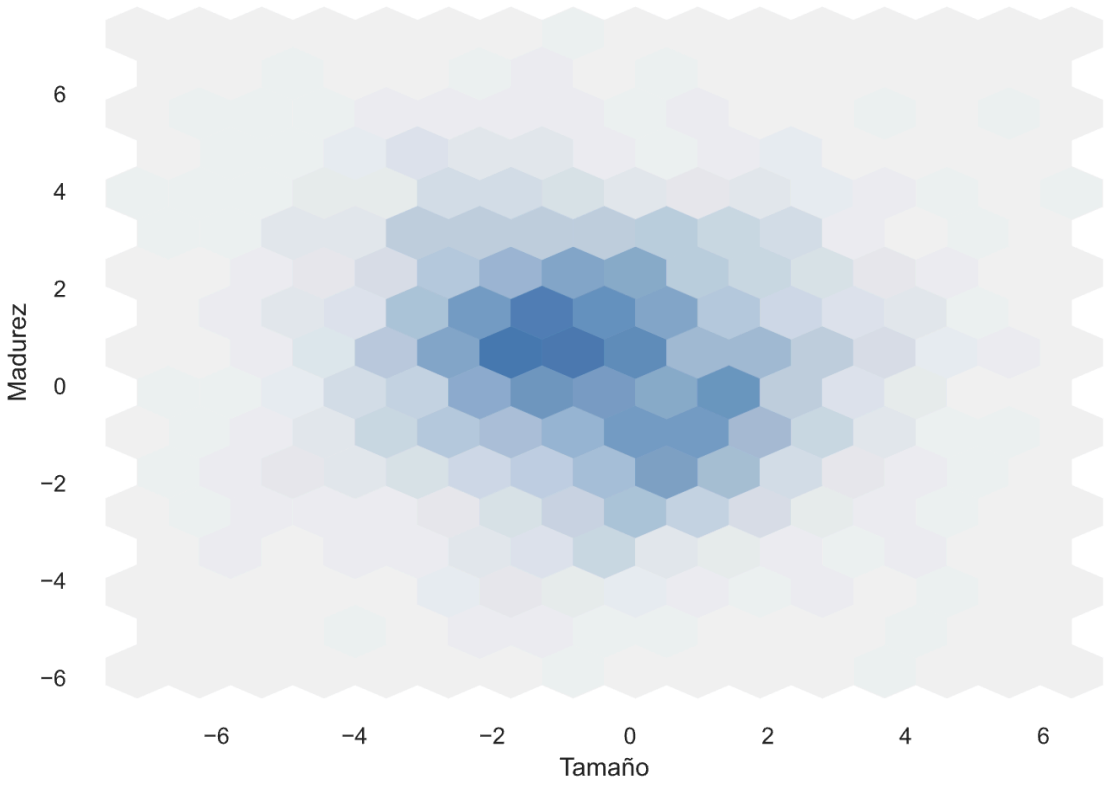
4. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.



En esta matriz podemos observar la correlación entre todas las variables, y como podemos observar los casos más fuertes, donde el coeficiente es mayor, son en la calidad vs la madurez y el tamaño con la textura. Lo que indica que estas dos parejas son las de mayor correlación. Sin embargo siguen siendo valores cercanos a 0.1 - 0.2, lo que indica una relación muy débil.

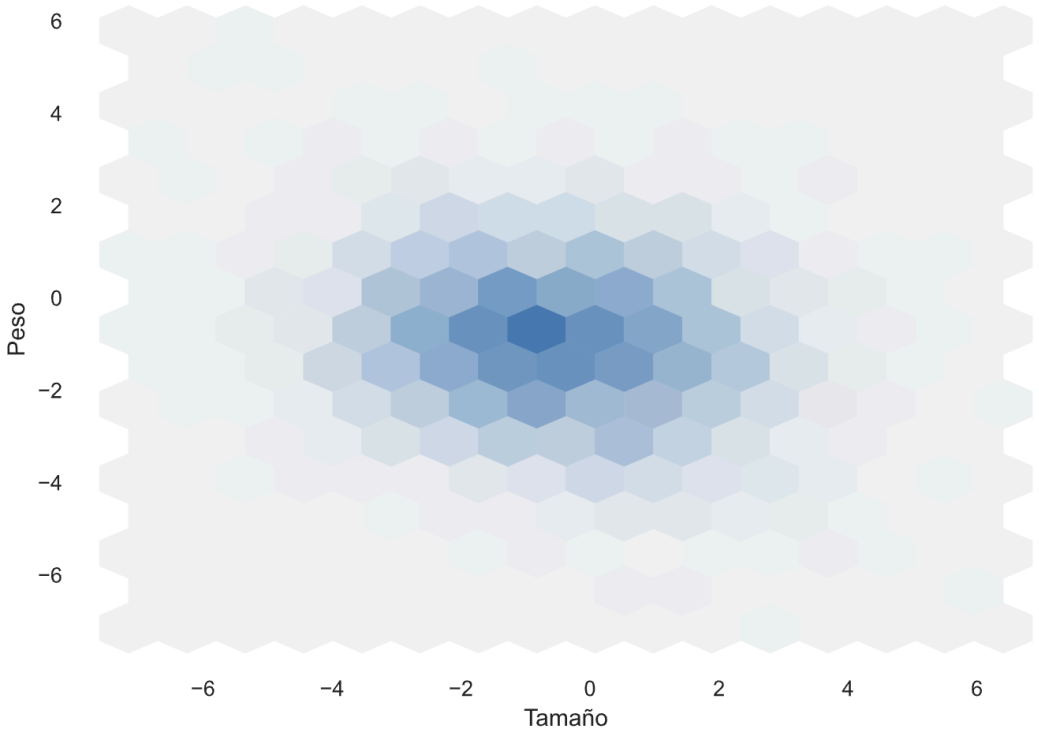
	Calidad	Dulzura	Humedad	ID	Madurez	Peso	Tamaño	Textura
Calidad	1.000	-0.237	-0.258	-0.005	0.260	-0.002	-0.245	-0.000
Dulzura	-0.237	1.000	0.098	0.002	-0.255	-0.120	-0.310	-0.017
Humedad	-0.258	0.098	1.000	0.004	-0.124	-0.091	-0.032	-0.237
ID	-0.005	0.002	0.004	1.000	-0.005	-0.007	-0.033	-0.016
Madurez	0.260	-0.255	-0.124	-0.005	1.000	-0.244	-0.155	-0.184
Peso	-0.002	-0.120	-0.091	-0.007	-0.244	1.000	-0.144	-0.087
Tamaño	-0.245	-0.310	-0.032	-0.033	-0.155	-0.144	1.000	0.172
Textura	-0.000	-0.017	-0.237	-0.016	-0.184	-0.087	0.172	1.000

Tamaño - Madurez



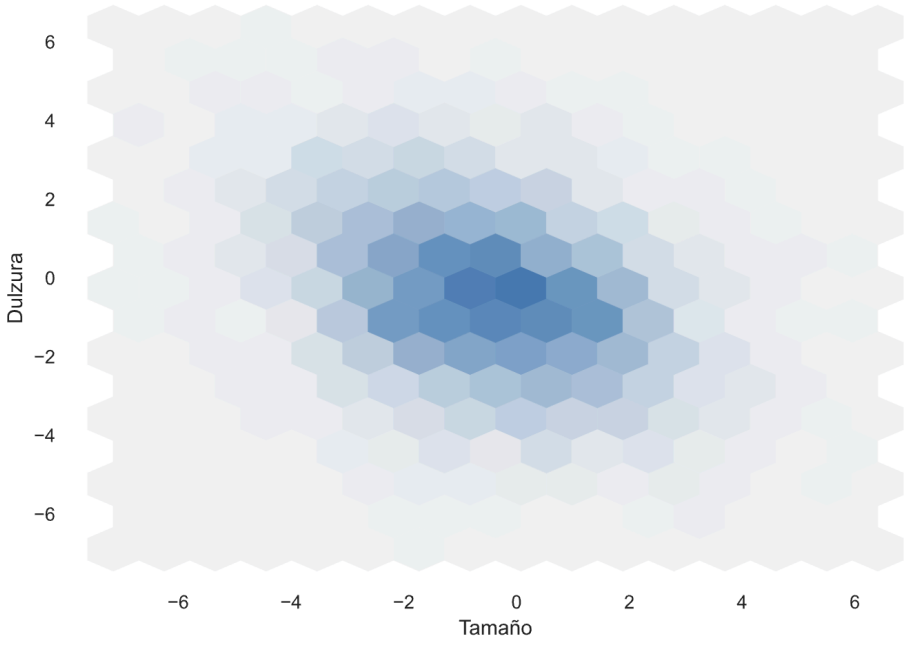
Correlación no existente, con promedio de -0.155.

Tamaño - Peso



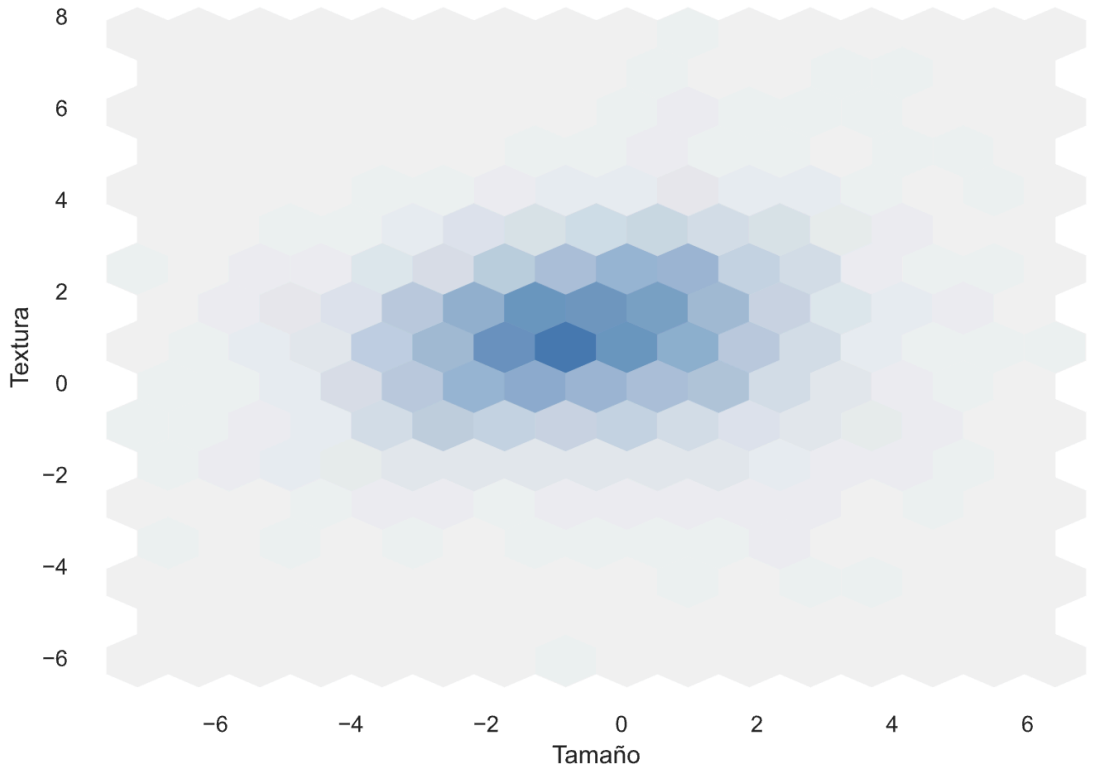
Correlación no existente, con valores en promedio alrededor del -0.144

Tamaño - Dulzura



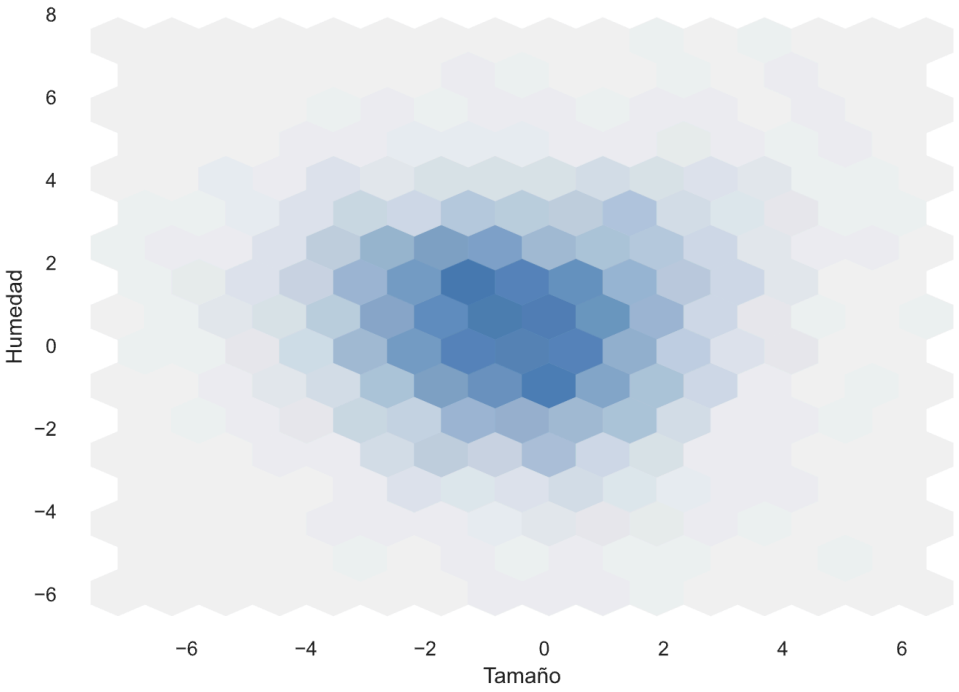
Correlación no existente, con valores en promedio alrededor del -0.310

Tamaño - Textura



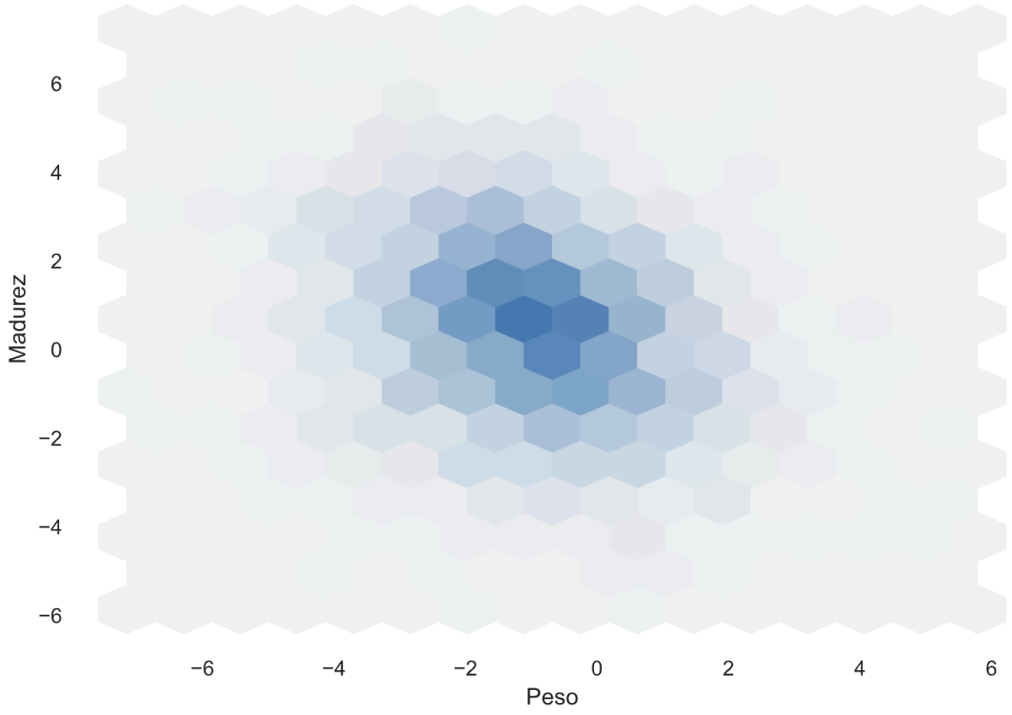
Correlación muy débil, con valores en promedio alrededor del 0.172.

Tamaño - Humedad



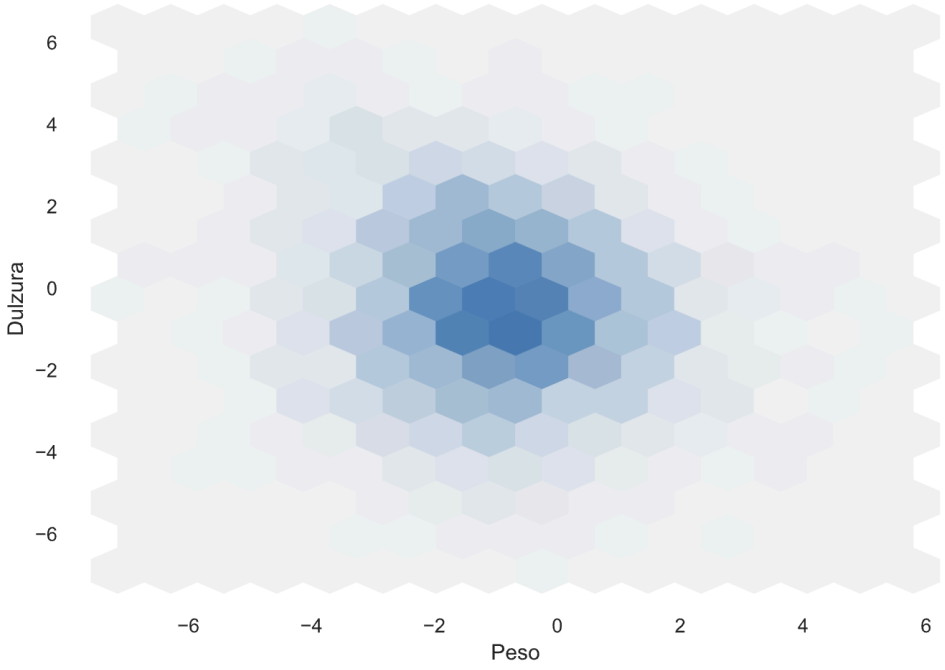
Correlación no existente, con valores en promedio alrededor del -0.032

Peso - Madurez



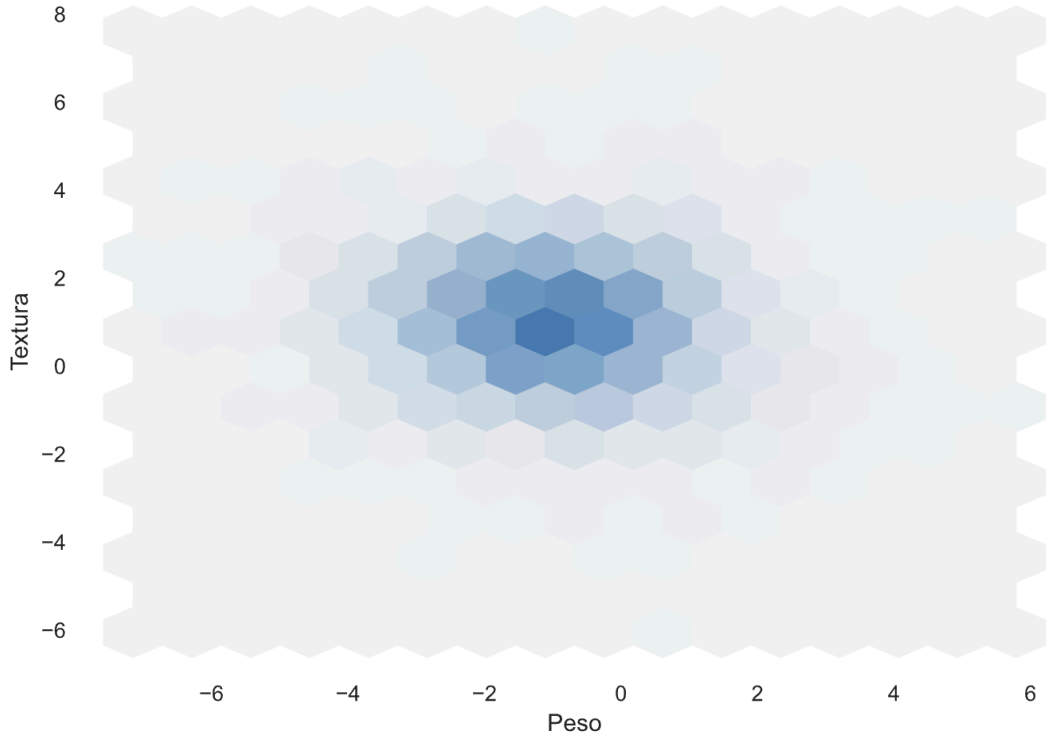
Correlación no existente, con valores en promedio alrededor del -0.244

Peso - Dulzura



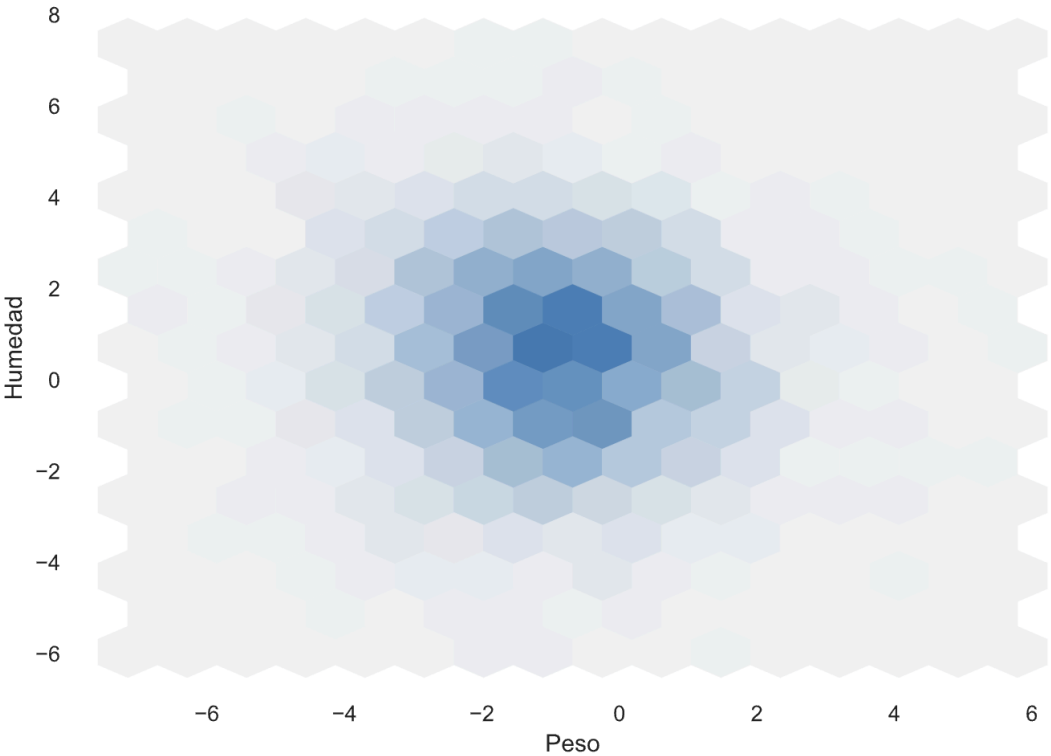
Correlación no existente, con valores en promedio alrededor del -0.12.

Peso - Textura



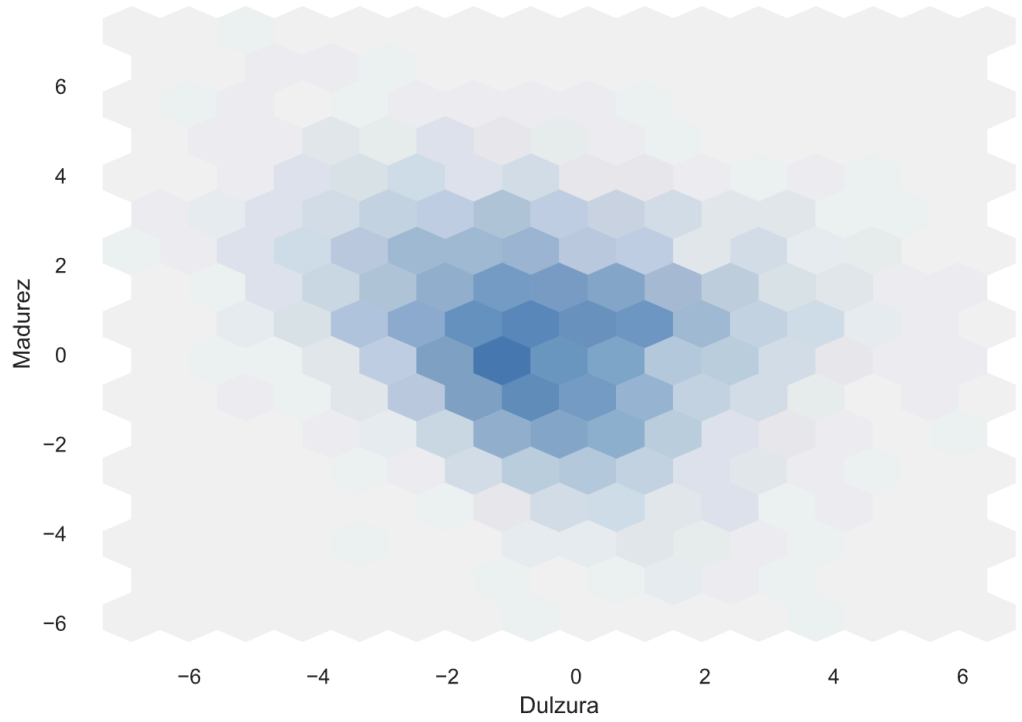
Correlación no existente, con valores en promedio alrededor del -0.087

Peso - Humedad



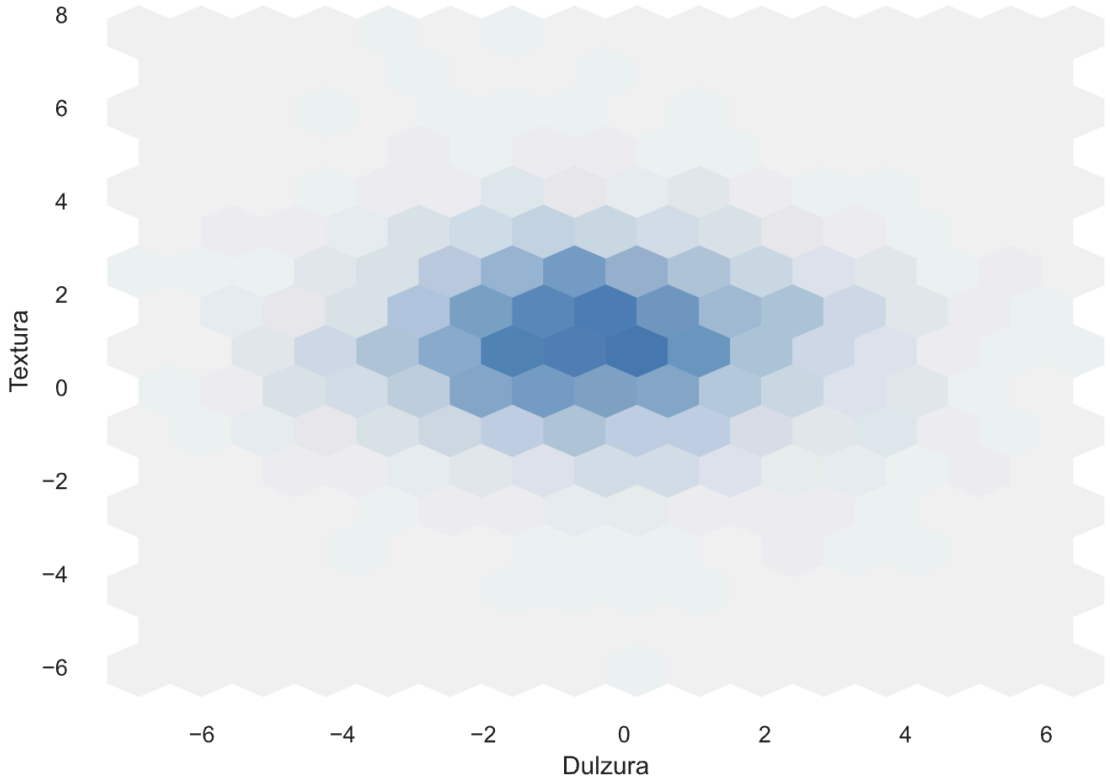
Correlación no existente, con valores en promedio alrededor del -0.091

Dulzura - Madurez



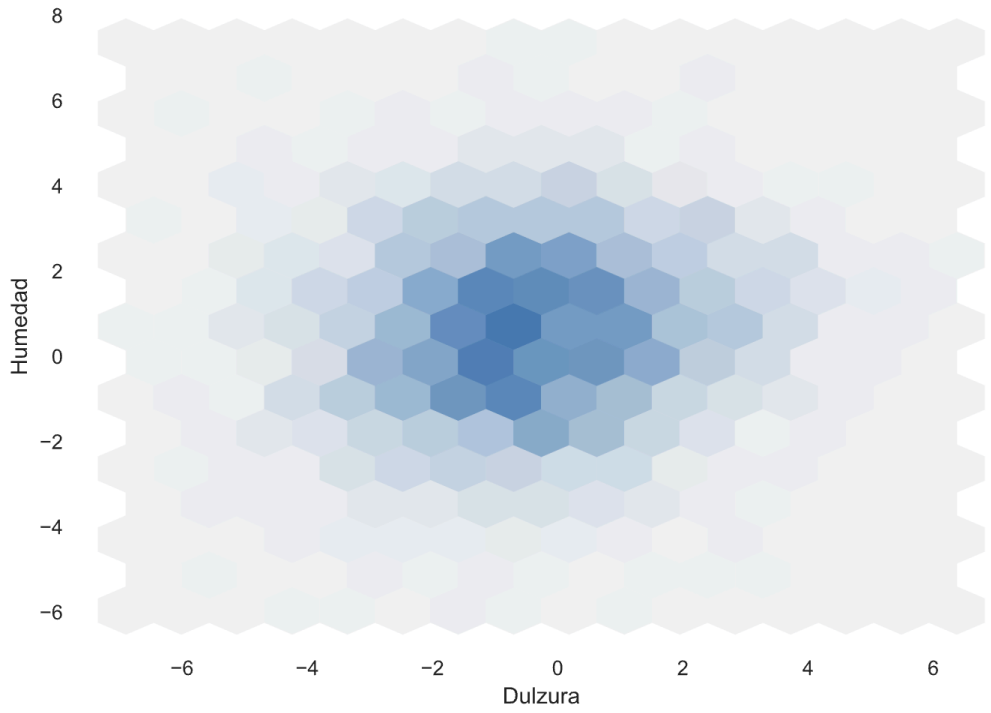
Correlación no existente, con valores en promedio alrededor del -0.255

Dulzura - Textura



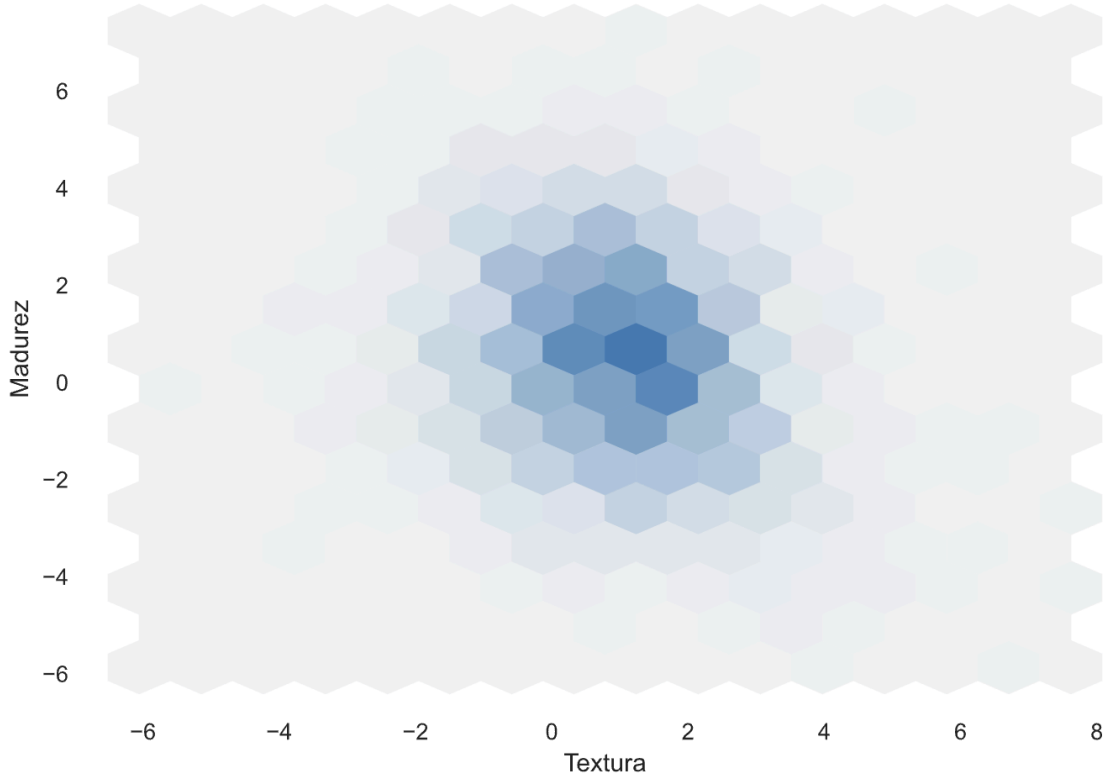
Correlación no existente, con valores en promedio alrededor del -0.017.

Dulzura - Humedad



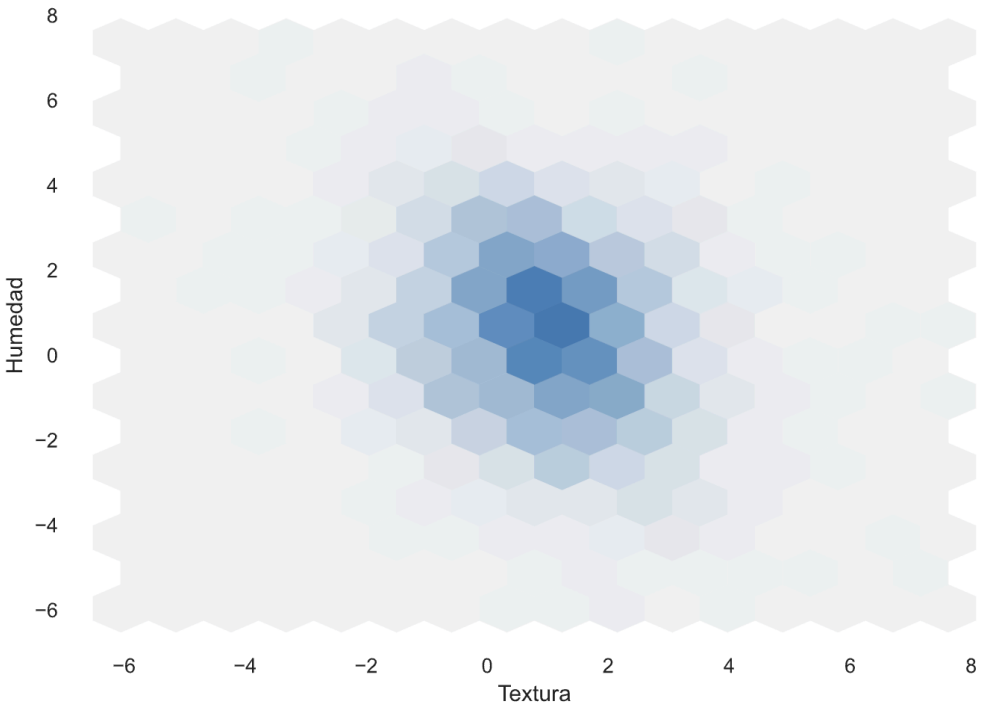
Correlación muy débil, con valores en promedio alrededor del 0.098

Textura - Madurez



Correlación no existente, con valores en promedio alrededor del -0.184

Textura - Humedad



Correlación no existente, con valores en promedio alrededor del -0.237

5. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos.

    - Listo, se realizó en los gráficos de arriba.

6. Realice la limpieza de variables utilizando las técnicas vistas en clase, u otras que piense pueden ser de utilidad.

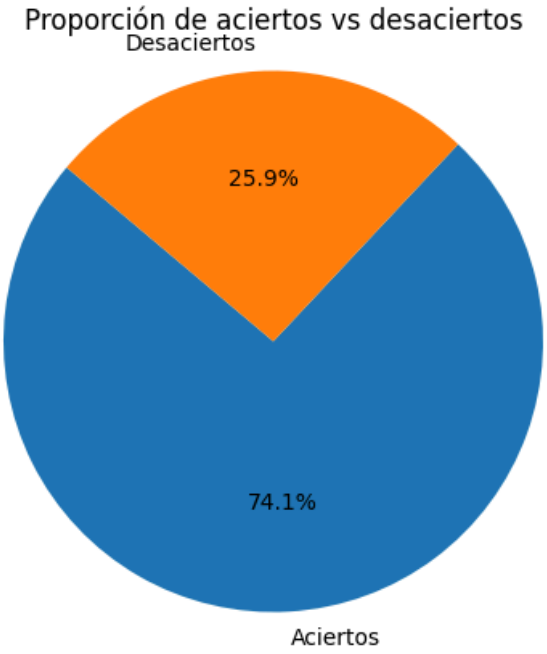
    En este caso los datos ya están normalizados, por lo que realmente no había mucha limpieza por realizar. Lo único que hicimos fue quitar la última fila del data frame ya que eran los créditos de la creadora del archivo.

Parte 2 – Desarrollo de un modelo de clasificación para determinar si la calidad de una manzana es buena o mala.

2.1. Prueben todos los modelos que se vieron en clase. Sí hay alguno que no consideran pertinente, expliquen porqué.

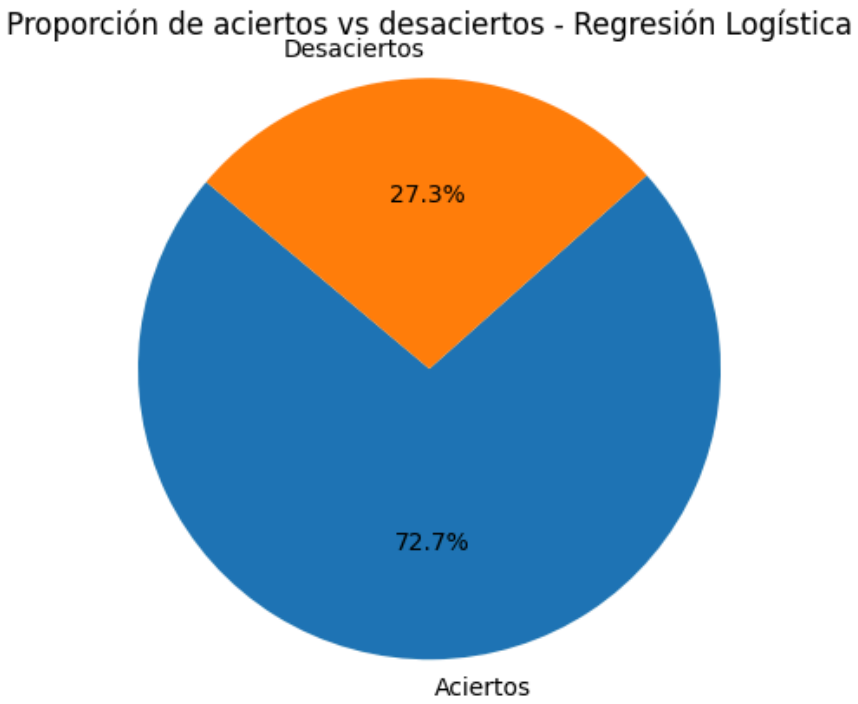
Naive Bayes:

```
Naive Bayes
Predicciones vs. Valores reales:
[['mala' 'mala']
 ['mala' 'buena']
 ['buena' 'buena']
 ...
 ['mala' 'buena']
 ['buena' 'buena']
 ['buena' 'mala']]
Matriz de Confusión:
[[374 120]
 [139 367]]
Precisión: 0.741
```



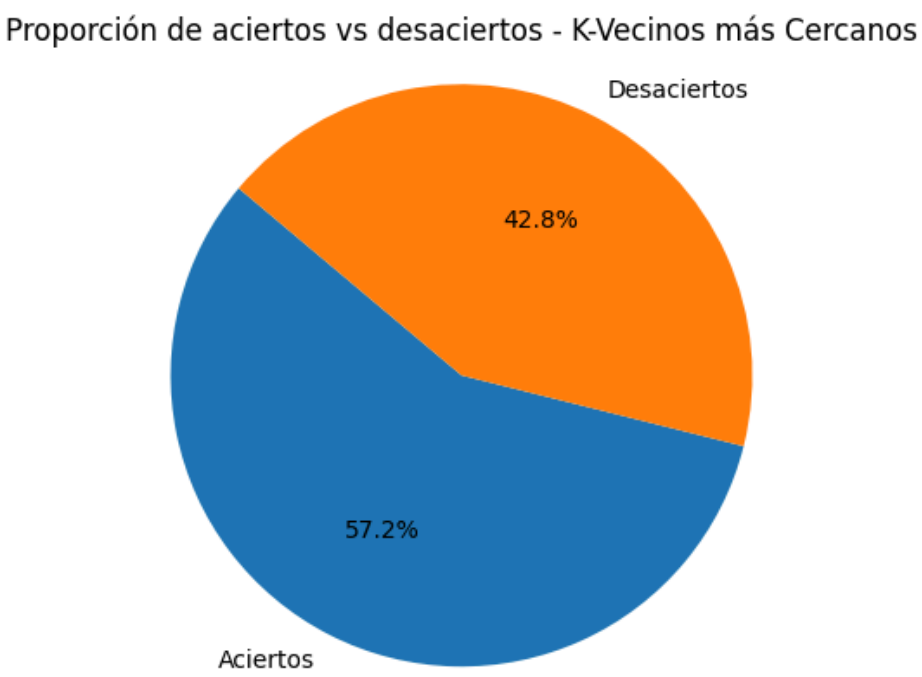
Regresión Logística:

```
Regresión Logística
Predicciones vs. Valores reales:
[['mala' 'mala']
 ['buena' 'buena']
 ['buena' 'buena']
 ...
 ['buena' 'buena']
 ['buena' 'buena']
 ['mala' 'mala']]
Matriz de Confusión:
[[367 127]
 [146 360]]
Precisión: 0.727
```



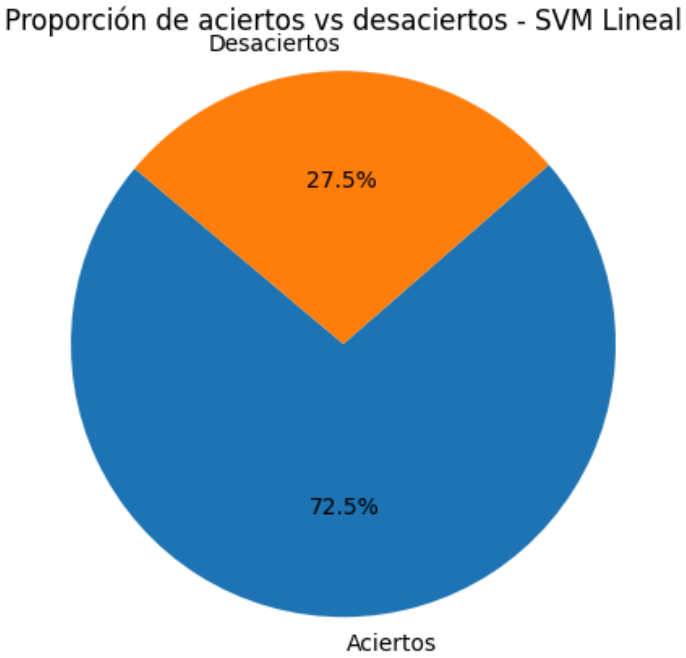
KNN

```
K-Vecinos más Cercanos
Matriz de Confusión:
[[259 235]
 [193 313]]
Precisión: 0.572
```



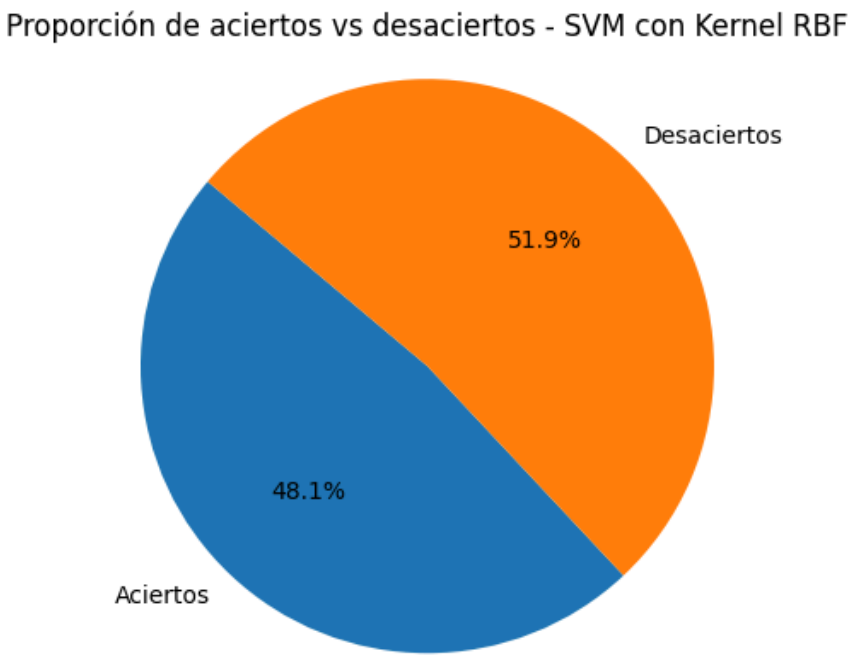
SVM Lineal

```
SVM Lineal
Matriz de Confusión:
[[368 126]
 [149 357]]
Precisión: 0.725
```



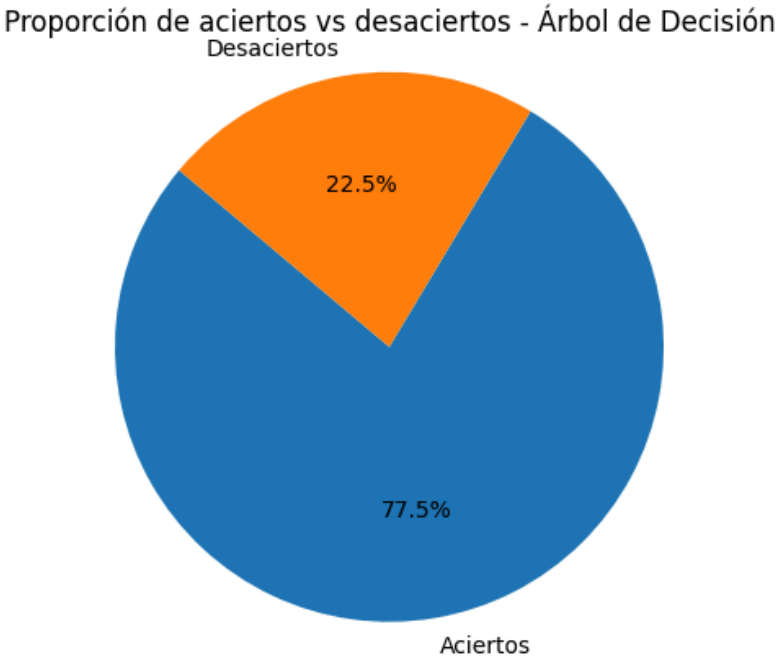
SVM con Kernel

```
SVM con Kernel RBF
Matriz de Confusión:
[[255 239]
 [280 226]]
Precisión: 0.481
```



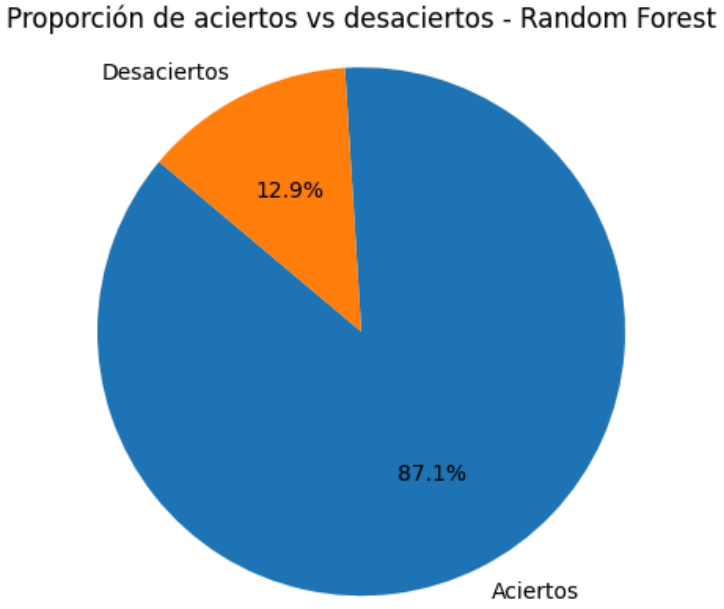
Decision Tree

```
Árbol de Decisión
Matriz de Confusión:
[[391 103]
 [122 384]]
Precisión: 0.775
```



Random Forest

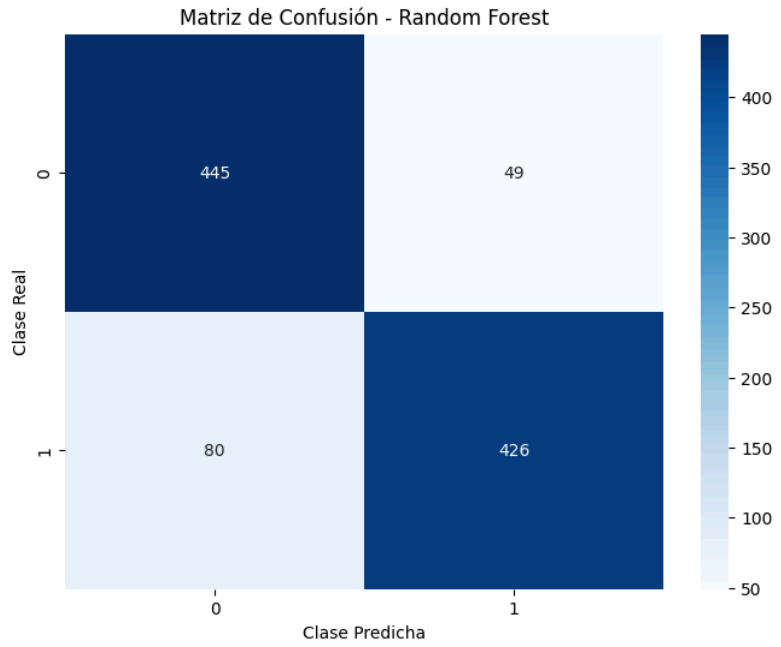
```
Random Forest
Matriz de Confusión:
[[445  49]
 [ 80 426]]
Precisión: 0.871
```



2.2. ¿Cuál es el modelo con mejor rendimiento? Utilice las métricas vistas en clase para dar respaldo a su respuesta. Recuerden que también pueden afinar los hiperparámetros.

El modelo con mejor rendimiento fue el Random Forest, con una precisión de 0.871. Es decir que acertó el 84% de los casos de prueba.

Y al evaluar su matriz de confusión:



Podemos observar que se encontraron:

- 426 verdaderos positivos
- 445 verdaderos negativos
- 49 falsos positivos
- 80 falsos negativos

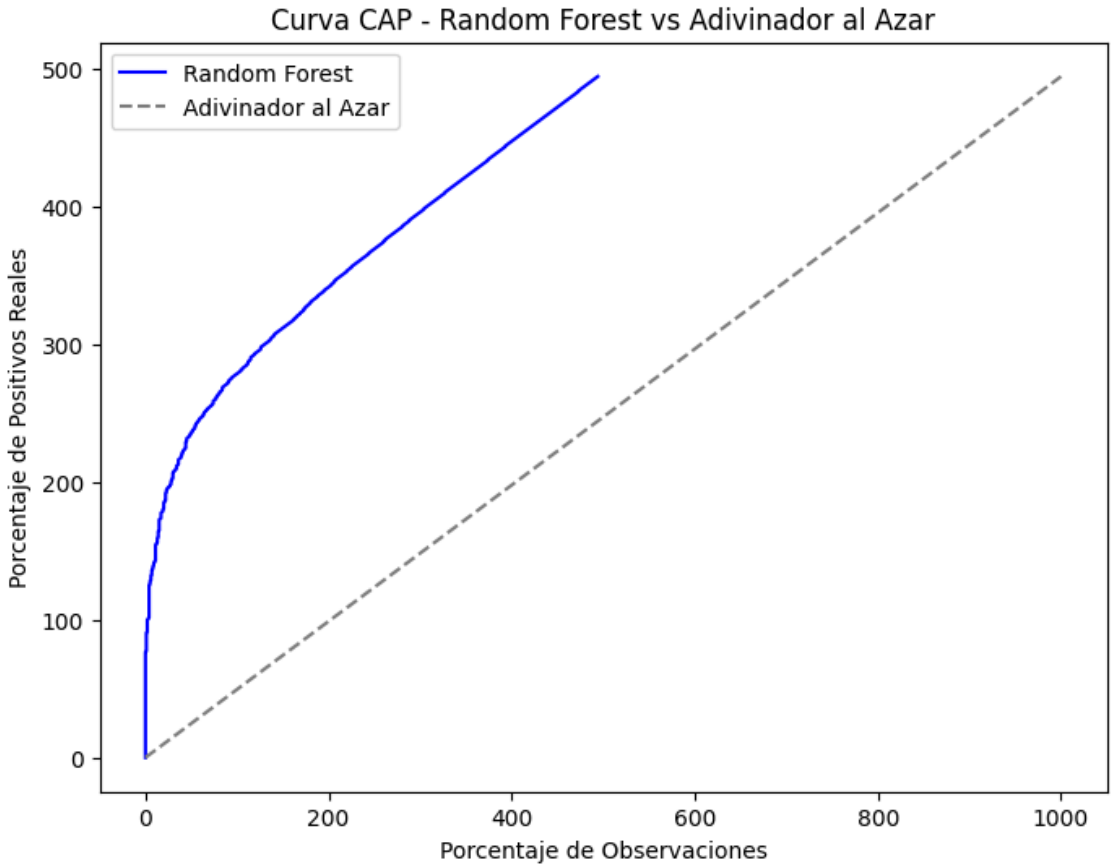
Lo que demuestra que el modelo si comete errores de tipo 1 y de tipo 2, mayormente de tipo 2. Ya que los falsos negativos representan las muestras que fueron clasificadas incorrectamente como negativas cuando en realidad pertenecen a la clase positiva.

Por lo que podemos calcular la Tasa de Exactitud:

$$T_{ex} = \frac{Tn + Tp}{Total} = \frac{871}{1000} = 0.871$$

Y este valor indica que es el modelo con mayor precisión y mejor rendimiento para clasificar las manzanas. También la precisión calculada por el modelo es del 87%, indicando que el modelo es muy bueno ya que se encuentra por arriba del 75% y también muestra que no está sobreajustado ya que no llega por arriba de los 90%.

Para visualizar el rendimiento de nuestro modelo, decidimos crear la gráfica del CAP:



Y como se puede observar, el modelo sigue validando su buen rendimiento al estar por encima de la recta del adivinador al azar. Es decir que es mejor que si clasificamos de manera al azar las manzanas.Lo que demuestra que la precisión del modelo es muy buena, estando con un 87% de precisión . Lo que lo clasifica como bueno pero no sobre ajustado.

Conclusiones

- El modelo con mejor rendimiento fue el Random Forest, con una precisión del 87.1%. Esto significa que acertó el 87.1% de los casos de prueba. Sin embargo, el modelo aún comete errores, especialmente de tipo 2, como se observa en la matriz de confusión.

- Al evaluar la matriz de confusión, encontramos 426 verdaderos positivos, 445 verdaderos negativos, 49 falsos positivos y 80 falsos negativos. Estos resultados indican que el modelo comete principalmente errores de tipo 2, clasificando incorrectamente algunas muestras que pertenecen a la clase positiva como negativas.

- La tasa de exactitud del modelo es del 87.1%, lo que indica que es capaz de clasificar correctamente la mayoría de las muestras. Esta precisión muestra que el modelo es efectivo, pero no está sobreajustado, ya que no alcanza un valor extremadamente alto que podría indicar sobreajuste.

- La curva de CAP confirma el buen rendimiento del modelo, ya que se encuentra por encima de la línea del adivinador al azar. Esto sugiere que el modelo es significativamente mejor que la clasificación al azar de las muestras, lo que respalda su precisión del 87.1%.

- La evaluación cuidadosa de modelos de clasificación permite seleccionar el más adecuado para maximizar la precisión de las predicciones y minimizar los errores. Entender en qué casos cada modelo funciona mejor ayuda a evitar el sobreajuste y la subestimación del rendimiento, garantizando predicciones precisas y decisiones correctas.

Referencias

ScikitLearn. (2024) Classifier comparison. Recuperado de: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Equipo de desarrollo de Matplotlib. (2023). Matplotlib: A Comprehensive Library for Creating Static, Animated, and Interactive Visualizations in Python [Software]. Disponible en https://matplotlib.org

McKinney, W., & otros colaboradores de Pandas. (2023). pandas: Powerful data structures for data analysis, time series, and statistics [Software]. Disponible en https://pandas.pydata.org

Waskom, M. (2023). Seaborn: Statistical Data Visualization [Software]. Disponible en https://seaborn.pydata.org