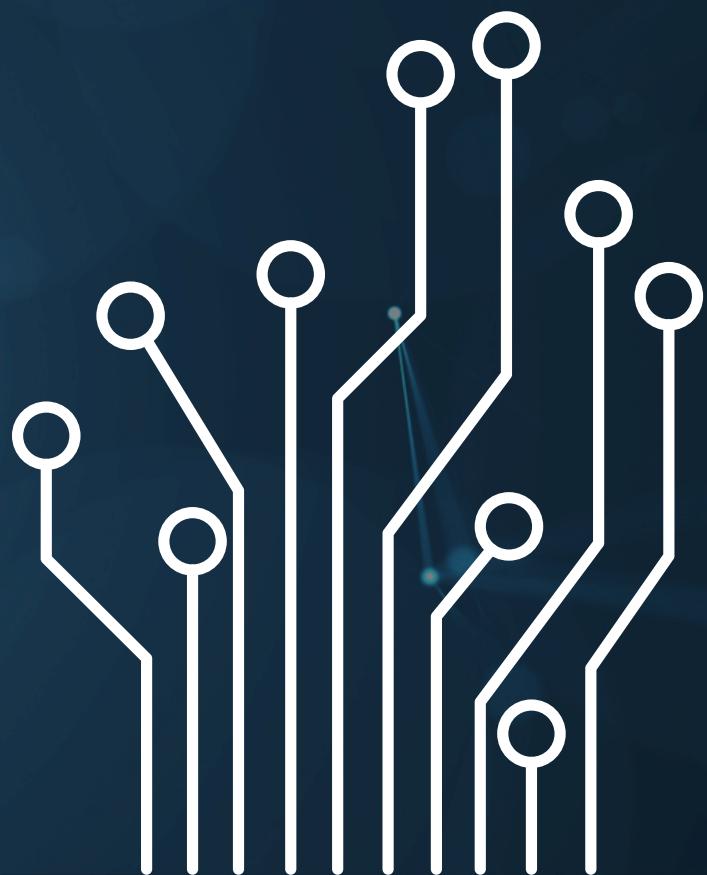


Comision 71945

# PROYECTO FINAL

Seguridad alimentaria

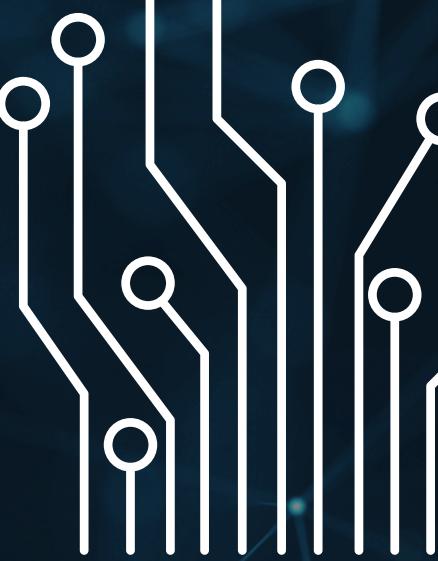
Ramiro Mata



# PRIMERA PARTE

# Abstracto: Motivación y Audencia

---



La seguridad alimentaria es un tema de importancia mundial ya que afecta directamente a la salud pública. En el Reino Unido, la Agencia de Normas Alimentarias (FSA) proporciona calificaciones de higiene alimentaria a diferentes empresas para informar a los consumidores y garantizar los estándares de calidad en la industria de alimentos y bebidas.

Este análisis se centra en explorar y evaluar estas calificaciones de higiene para:

Determinar patrones de calidad de la salud en diferentes tipos de instituciones.

Examine qué factores afectan la calificación de una empresa.

Proporcionando información útil para decisiones de salud pública, regulatorias y comerciales.

# Audiencia Beneficiada



## Consumidores

Personas que quieran tomar decisiones informadas sobre dónde comer o comprar alimentos de forma segura.

## Analistas de Datos y Científicos de Datos

Profesionales interesados en el análisis de datos relacionados con la seguridad y calidad de los alimentos en la industria alimentaria.

## Autoridades de Regulación

Organismos de salud pública y organismos de control que buscan mejorar las condiciones sanitarias en los locales de alimentación.

## Dueños de Negocios en la Industria Alimentaria

Restaurantes, cafeterías y supermercados que buscan mejorar sus estándares de higiene y cumplir con regulaciones.

# Resumen de Metadata



## Información General del Dataset

El conjunto de datos analizado proviene de la API de la Agencia de Normas Alimentarias del Reino Unido y contiene información de calificación de higiene para diferentes empresas. A continuación se muestra un resumen de su estructura:

**Número de filas:** 1,303 registros (establecimientos analizados).

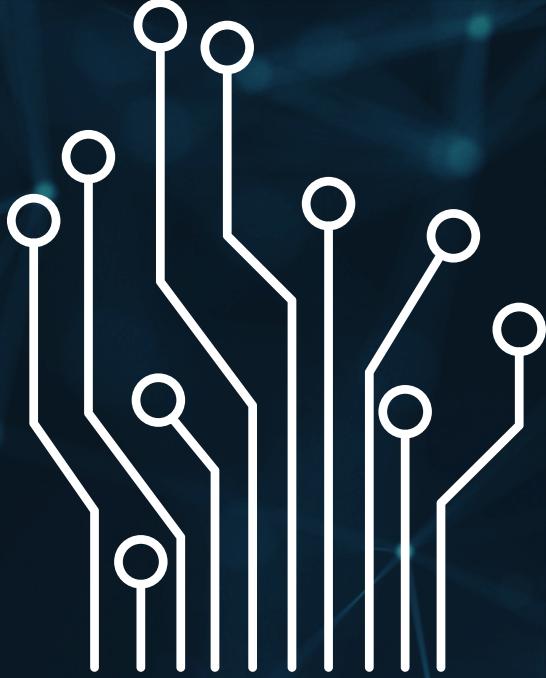
**Número de columnas:** 25 características sobre cada establecimiento.

## Tipos de Variables en el Dataset

- **Variables Categóricas:** incluyen identificadores, nombres de negocios, tipos de establecimientos, esquemas de calificación, y otros campos de texto o booleanos relevantes para clasificar y describir cada entidad.
- **Variables Numéricas:** abarcan identificadores únicos, puntuaciones otorgadas durante la inspección (como higiene, estructura y confianza en la gestión), así como coordenadas geográficas.
- **Variables de Fecha y Localización:** representan la fecha de inspección y componentes de la dirección del establecimiento, útiles para análisis temporales y espaciales.

# Valores nulos en el dataset:

- RatingDate tiene **181 valores nulos** (faltan fechas de inspección en algunos establecimientos).
- Geocode tiene **207 valores nulos** (coordenadas geográficas faltantes en algunos registros).
- Scores tiene **232 valores nulos**, ya que algunas inspecciones no incluyen puntuaciones detalladas.
- Las columnas AddressLine3 y AddressLine 4 van a ser eliminadas ya que tienen muchos valores nulos y no son necesarias



# Preguntas e Hipótesis a Responder

01

- ¿Cuál es la distribución de calificaciones de higiene en los establecimientos?
- Hipótesis: La mayoría de los establecimientos tienen una calificación alta de higiene.

02

- ¿Qué tipo de establecimientos tienden a recibir las mejores y peores calificaciones de higiene?
- Hipótesis: Restaurantes y tiendas minoristas tienen una mayor variabilidad en sus calificaciones de higiene.

03

- ¿Cómo han evolucionado las inspecciones de higiene en los últimos años?
- Hipótesis: El número de inspecciones ha aumentado debido a regulaciones más estrictas.

04

- ¿Existe una relación entre la calificación de higiene y la confianza en la gestión del establecimiento?
- Hipótesis: Establecimientos con alta confianza en la gestión suelen obtener mejores calificaciones de higiene.

# Visualizaciones ejecutivas que responden nuestras preguntas 1 y 2.

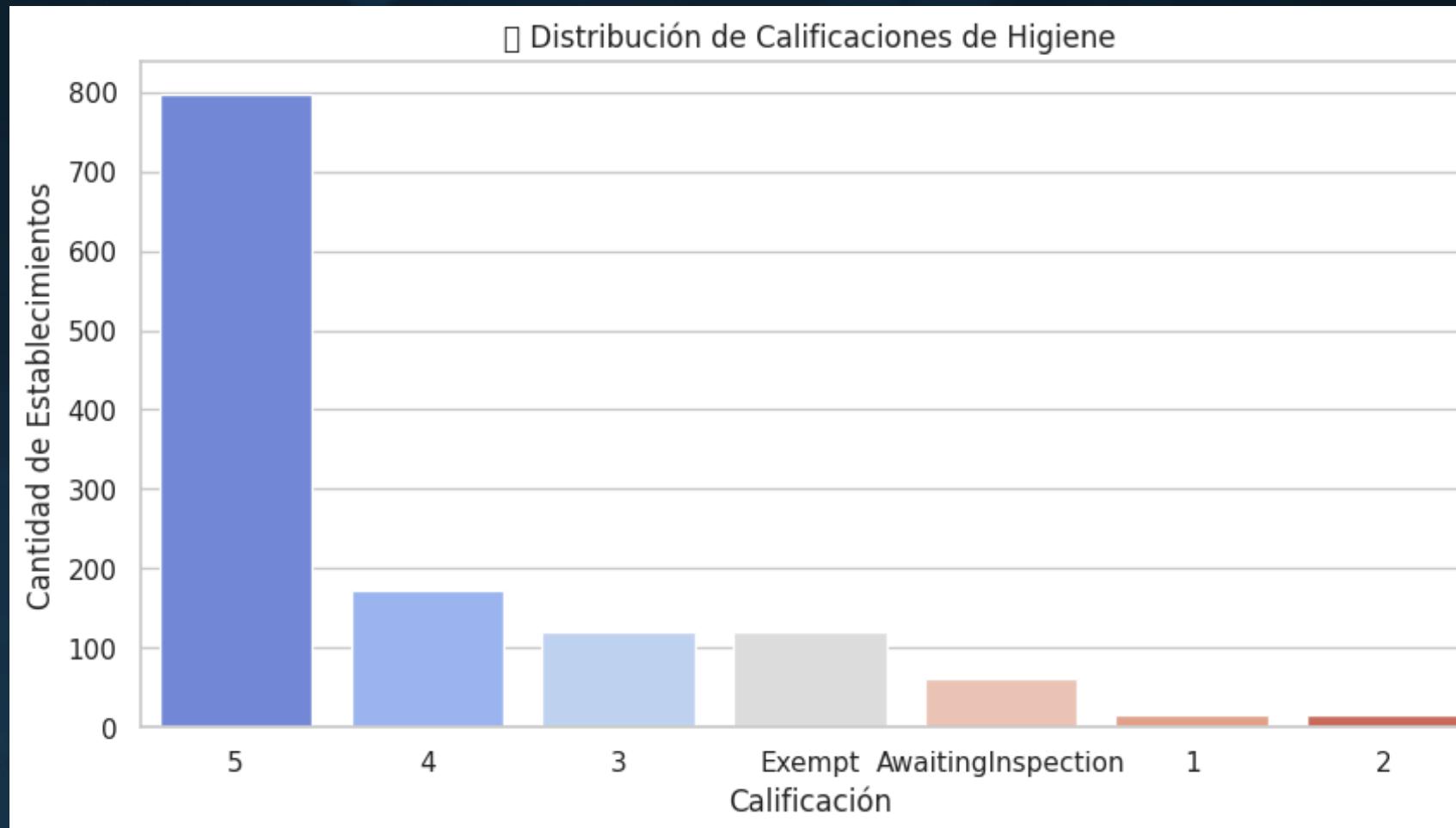


Gráfico de barras: Representa la frecuencia de cada calificación.

Hallazgo: La mayoría de los establecimientos tienen una calificación de 5.

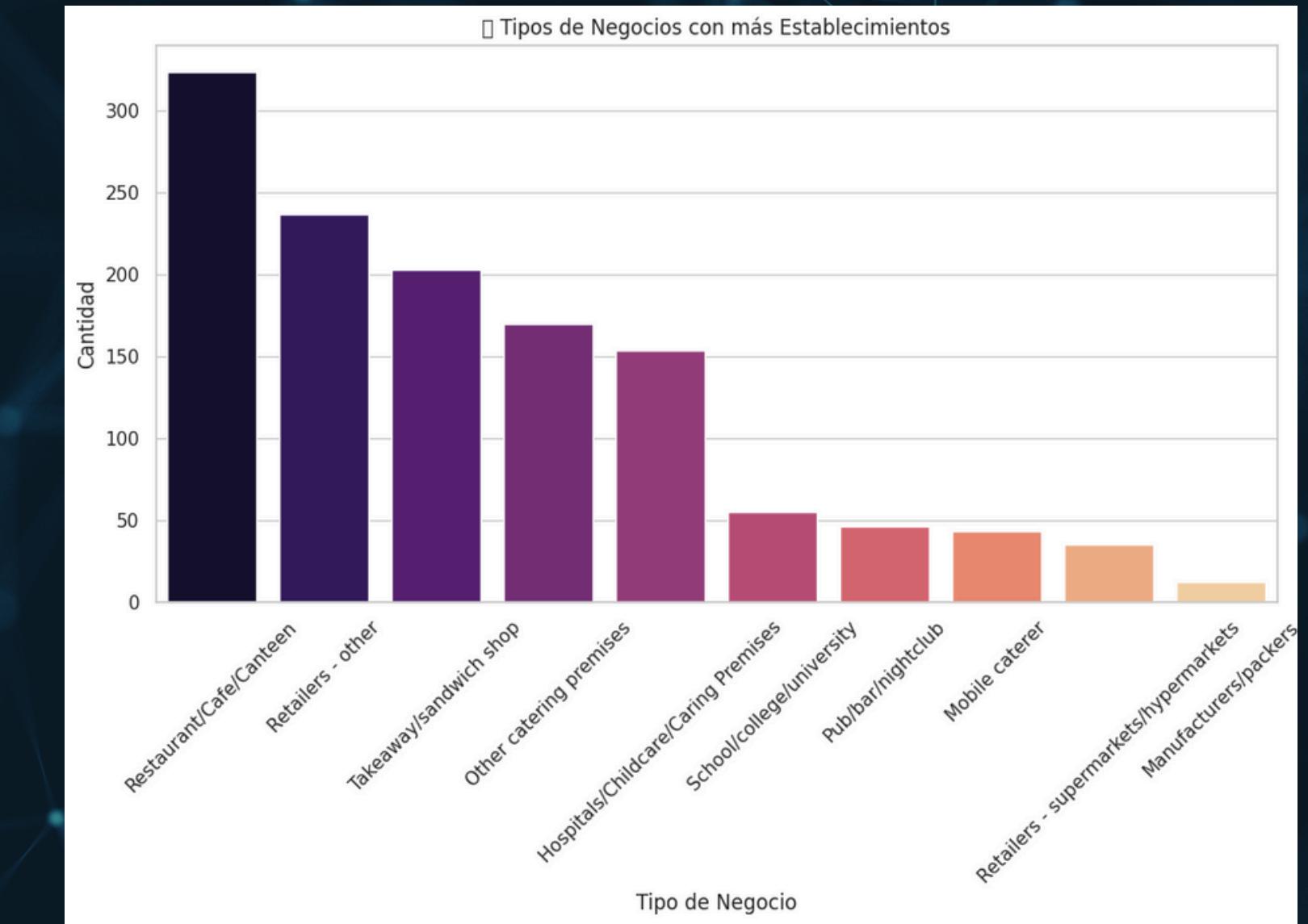


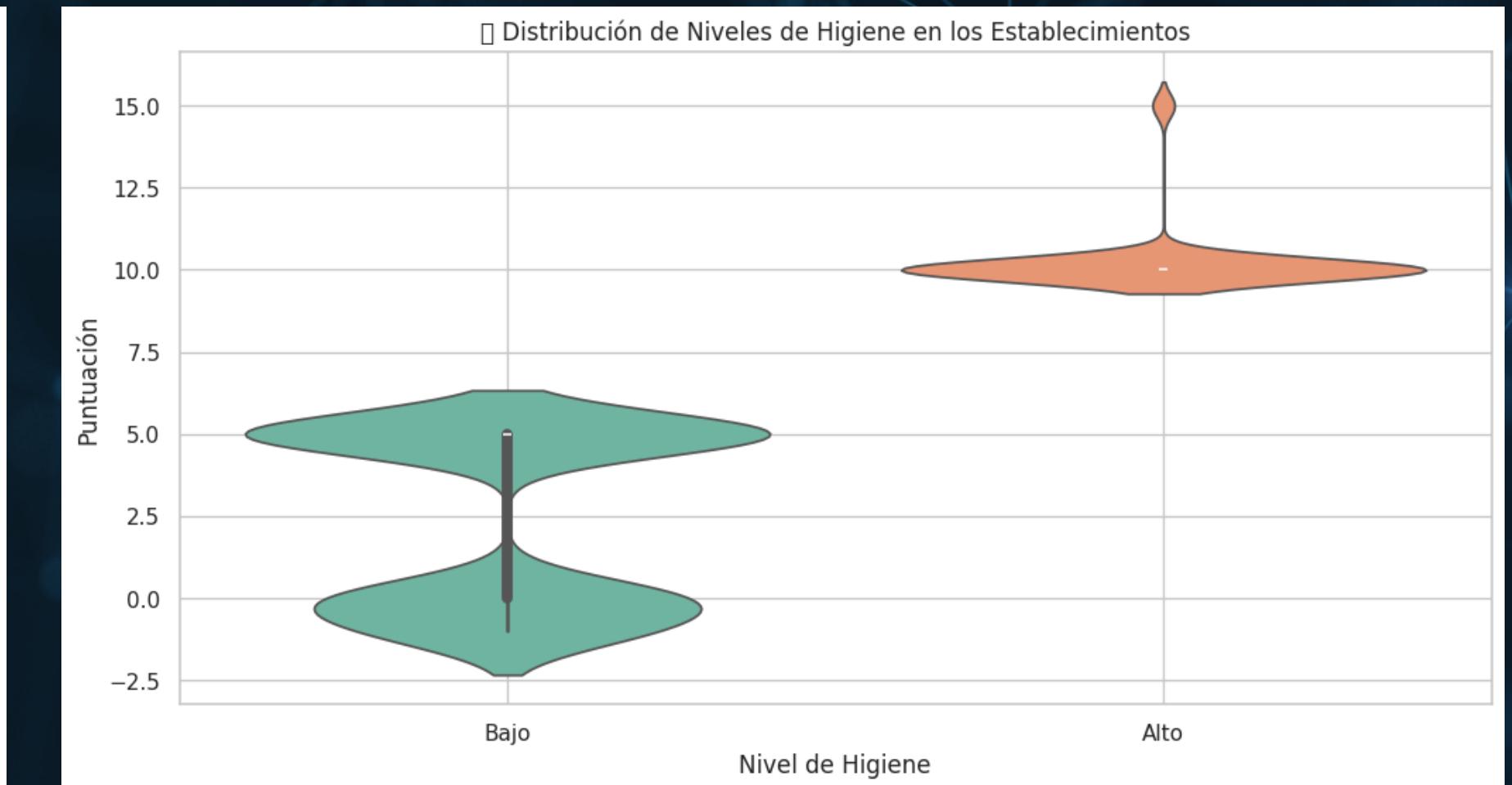
Gráfico de barras: Se observan los tipos de negocio más comunes en la base de datos.

Hallazgo: Los restaurantes y minoristas dominan la lista.

# Visualizaciones ejecutivas que responden nuestras preguntas 3 y 4.



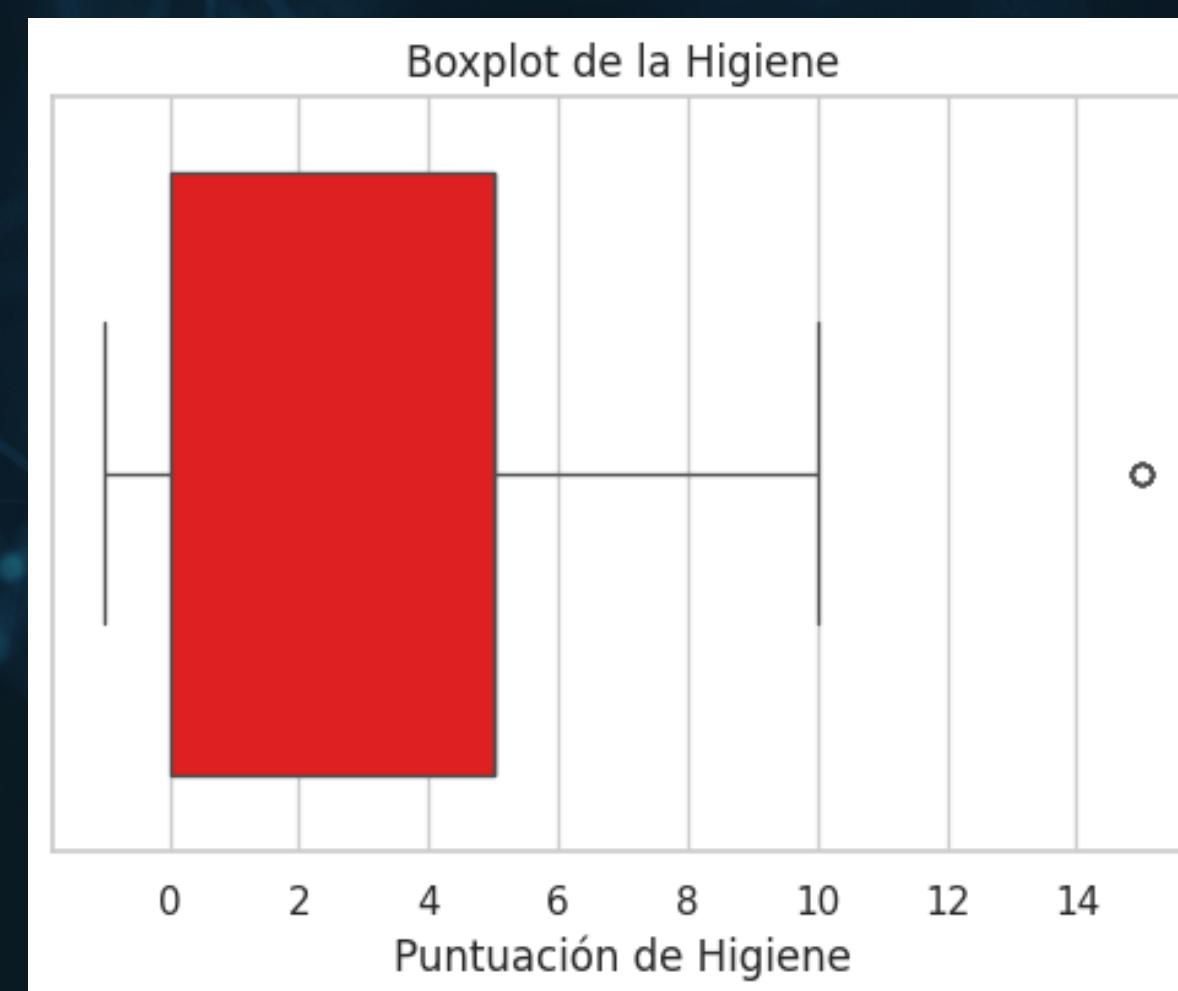
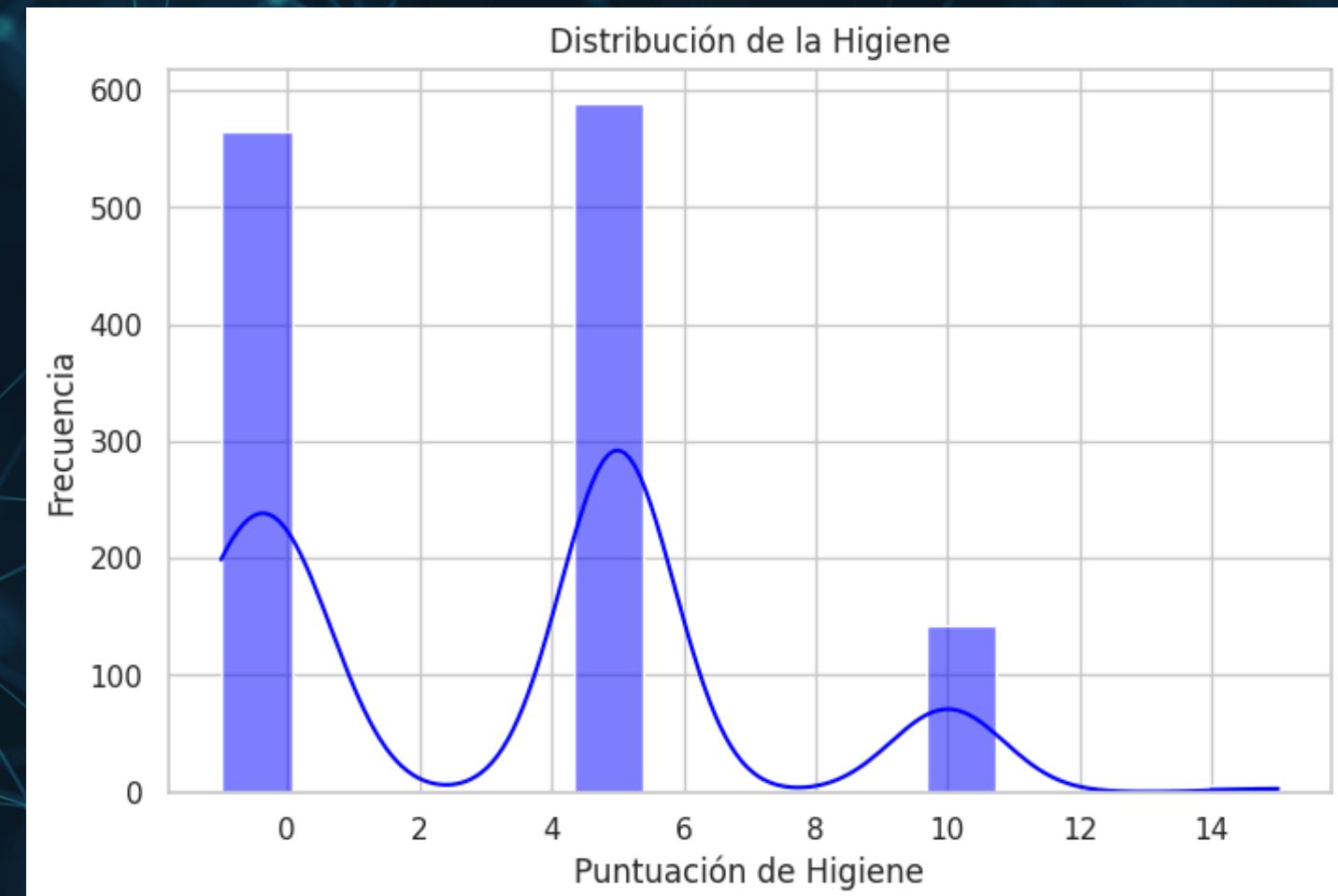
Hallazgo: Aumento en la cantidad de inspecciones en los últimos años.



Hallazgo: Existe una relación positiva entre ambas variables.

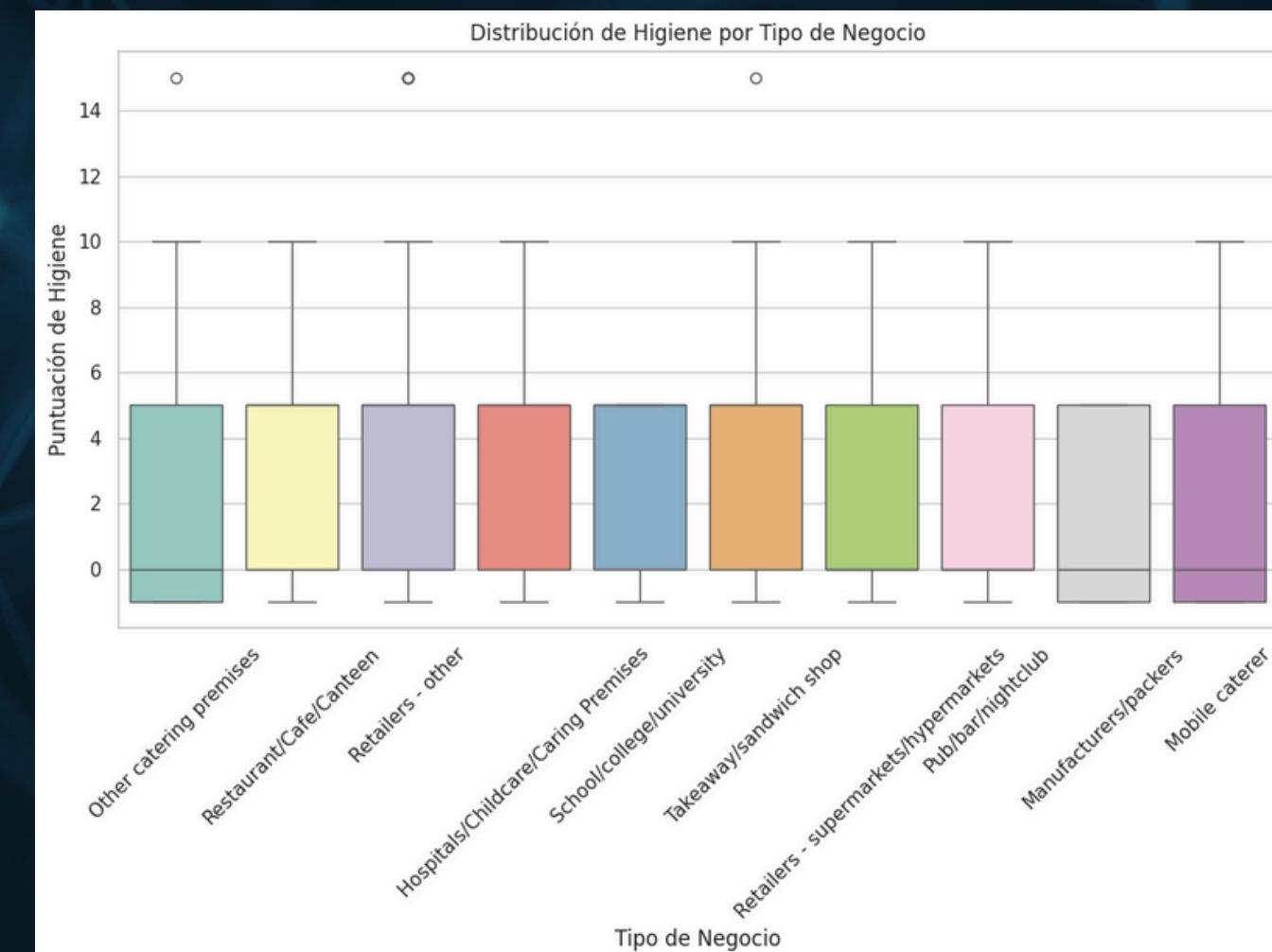
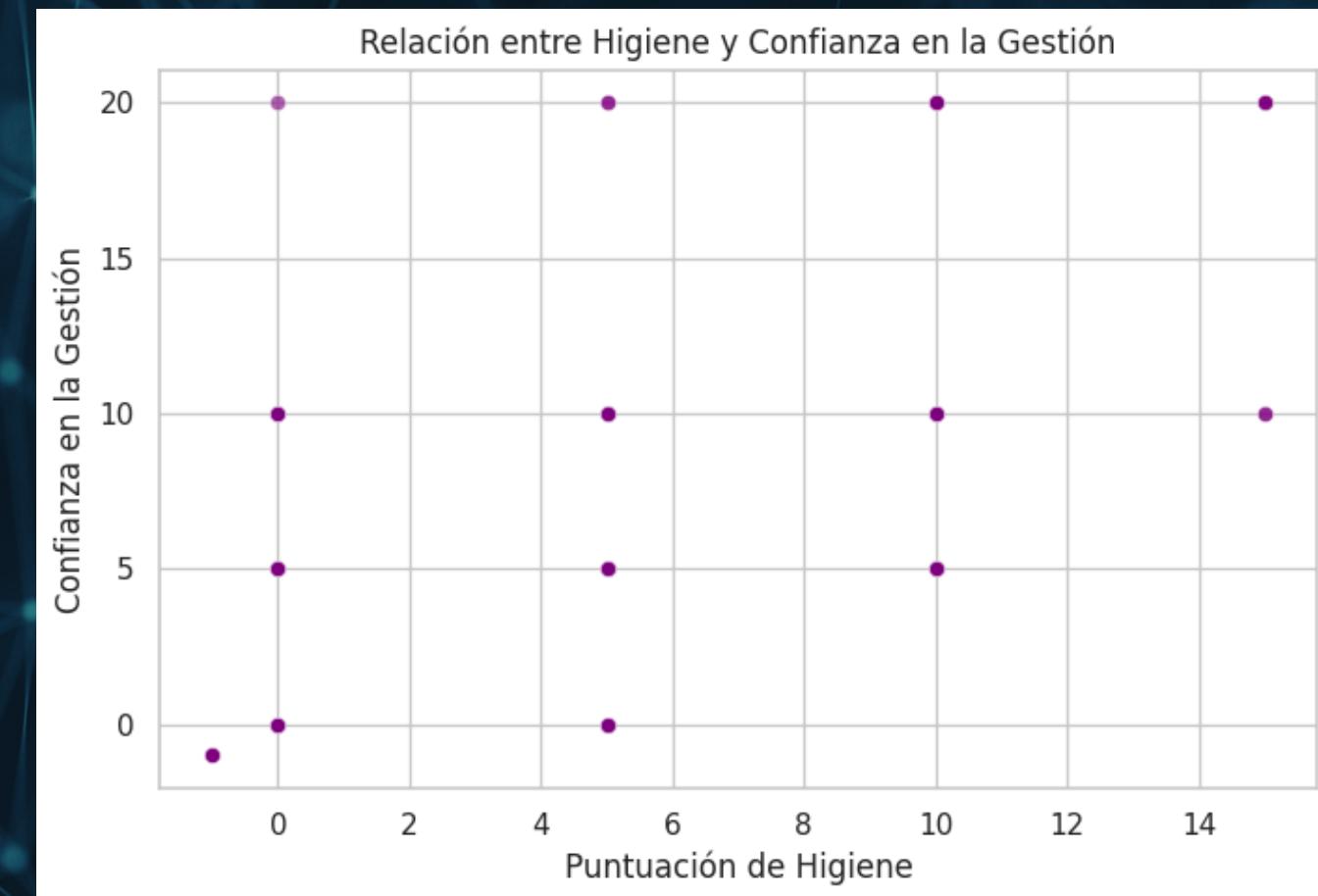
# Análisis Univariado

- Estadísticas destacadas:  
Media: 3.25 | Mediana: 5 | Moda: 5  
Desviación estándar: 3.60  
Asimetría: 0.46 → Hay valores altos extremos.  
Curtosis: -0.57 → Distribución más plana que la normal.
- Principales observaciones:  
La mayoría de los establecimientos tienen buena higiene (modo y mediana en 5).  
Hay una leve inclinación hacia valores altos, pero también se detectan puntuaciones bajas.  
El boxplot revela outliers con valores extremos, posiblemente por errores o casos críticos.



# Análisis Bivariado

- ◆ Relación entre Higiene y Confianza en la Gestión:  
El gráfico de dispersión muestra una tendencia positiva: los establecimientos con mayor confianza en la gestión tienden a tener mejores puntajes de higiene.
- ◆ Diferencias según Tipo de Negocio:  
El boxplot indica que rubros como restaurantes y supermercados presentan más variabilidad en higiene, mientras que hospitales y escuelas suelen mantener estándares más consistentes.



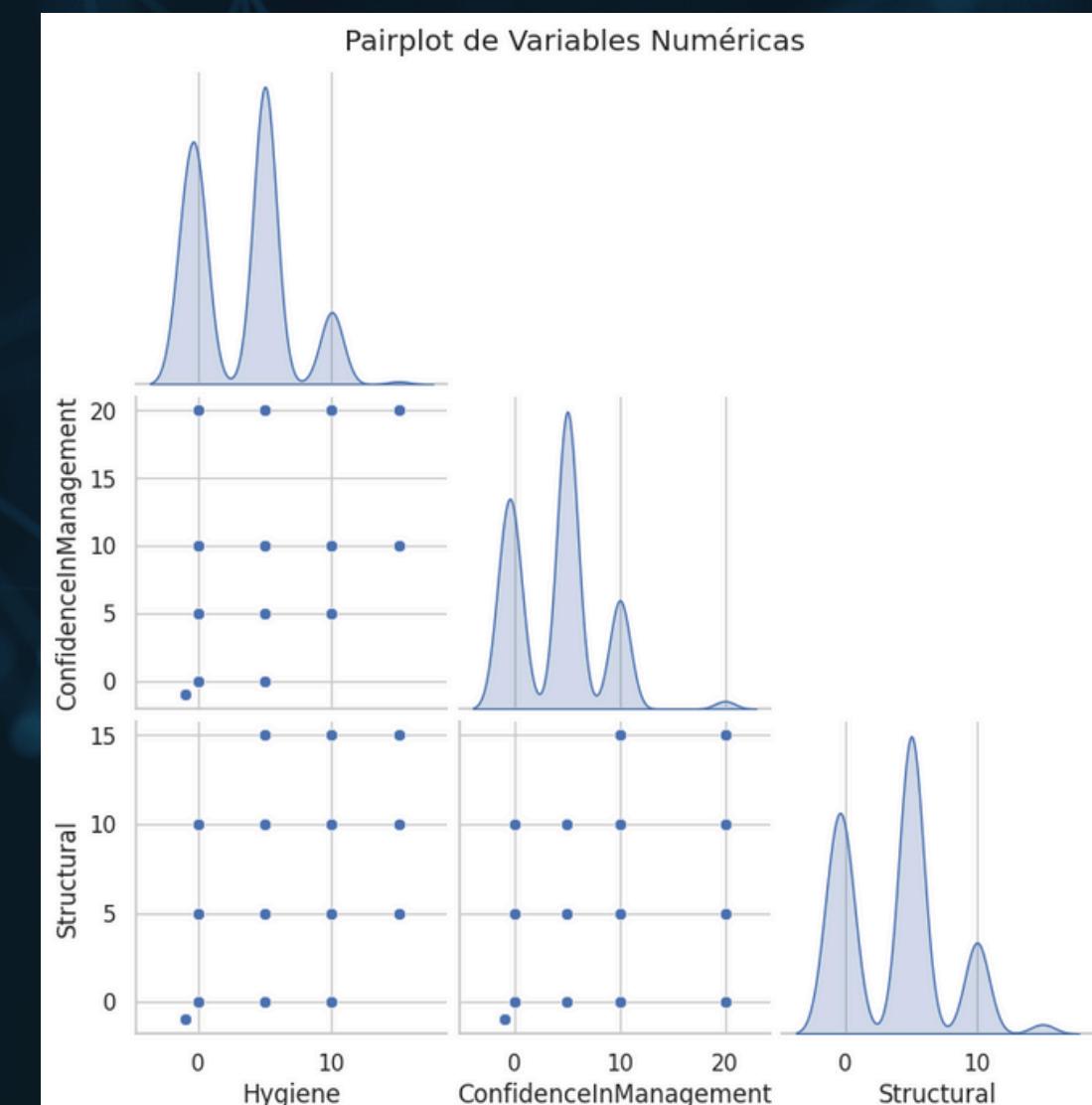
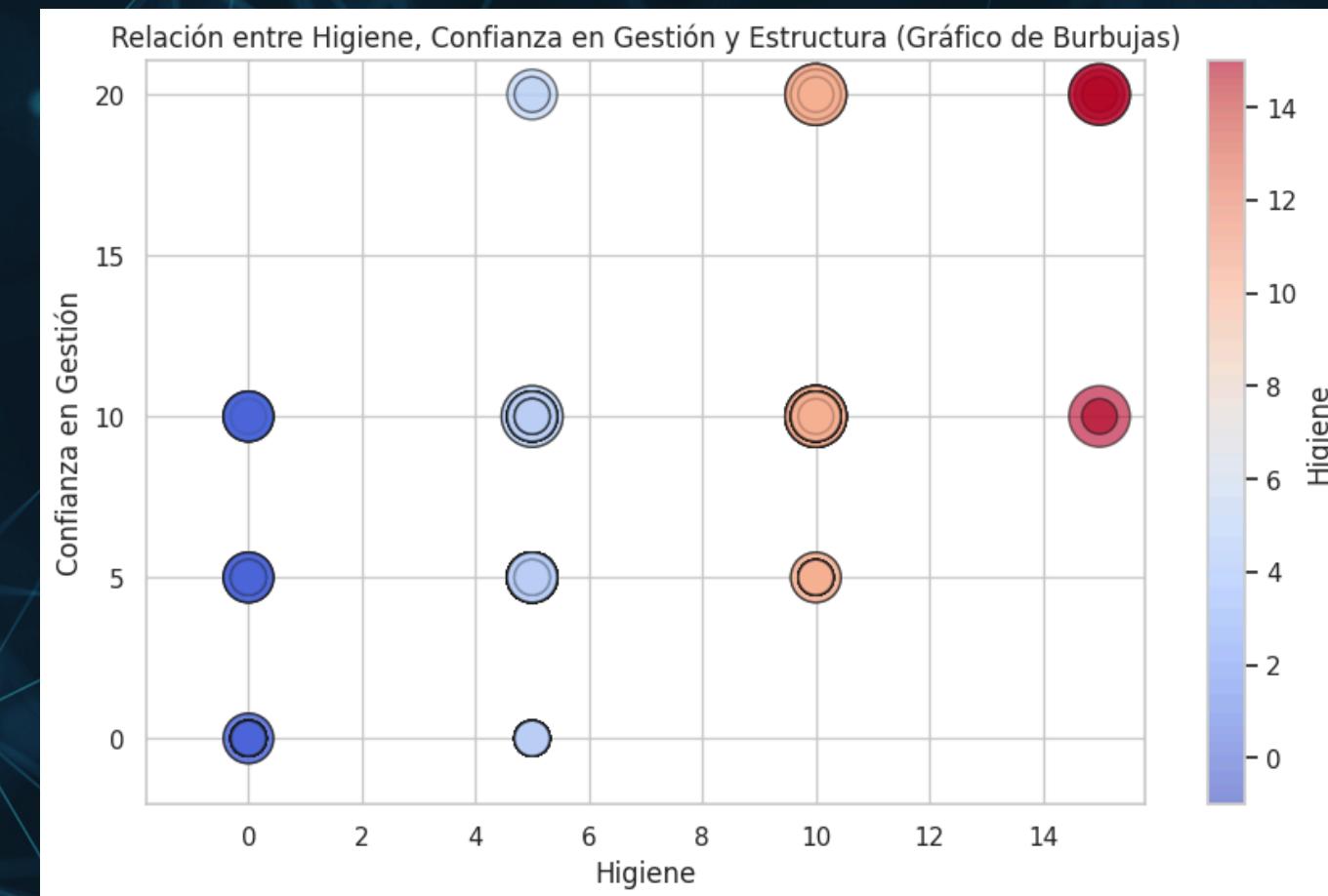
# Análisis Multivariado

## Gráfico de Burbujas:

Se observa una relación general entre mayor higiene y mayor confianza en la gestión, aunque con excepciones. Además, la variable estructural influye, ya que el tamaño de la burbuja (estructura) parece estar relacionado con mejores prácticas.

## Pairplot:

Las variables muestran distribuciones multimodales, lo que indica que las puntuaciones se asignan en categorías discretas. No hay correlaciones fuertes entre las variables, pero sí patrones agrupados por niveles específicos.



# Insights de las preguntas e hipótesis

01

Mayoría con alta higiene: La calificación más común es 5, lo que indica que la mayoría de los establecimientos cumplen con altos estándares de higiene.

02

Restaurantes y minoristas predominan: Son los tipos de negocios más evaluados, lo que resalta su impacto en la seguridad alimentaria.

03

Aumento en inspecciones recientes: Posiblemente debido a normativas más estrictas o una mayor supervisión gubernamental.

04

Relación entre higiene y gestión: Establecimientos con mayor confianza en la gestión suelen tener mejores calificaciones de higiene.

# Insights de los análisis univariado, multivariado y bivariado

## 01 Insights del Análisis Univariado

- La mayoría de los establecimientos tienen buena higiene (modo = 5).
- Hay algunos con muy baja higiene (sesgo a la derecha).
- Se detectaron outliers con valores extremos.

## 02 Insights del Análisis Bivariado

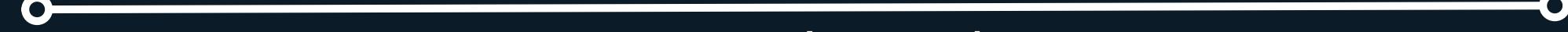
- Mayor confianza en la gestión se asocia a mejor higiene.
- Restaurantes y tiendas muestran mayor variabilidad en higiene.
- Algunos tipos de negocios presentan diferencias marcadas en las calificaciones.

## 03 Insights del Análisis Multivariado

- Las variables clave (higiene, estructura, gestión) se distribuyen en niveles discretos.
- No hay correlaciones lineales fuertes entre las variables.
- Hay negocios con baja higiene pero alta confianza, lo que sugiere otros factores influyentes.

# SEGUNDA PARTE

# Modelo 1: Random Forest Classifier

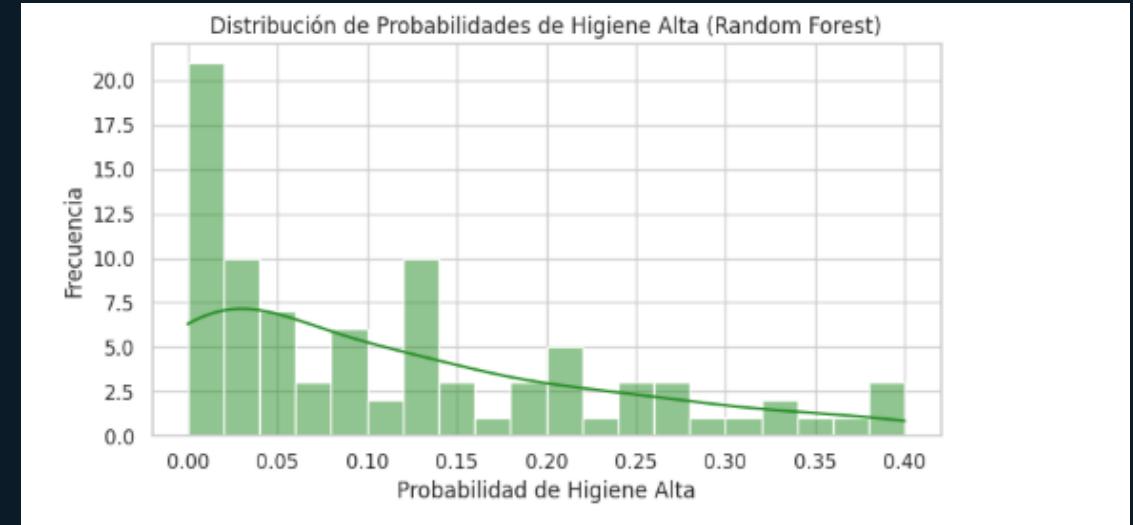


## Insight:

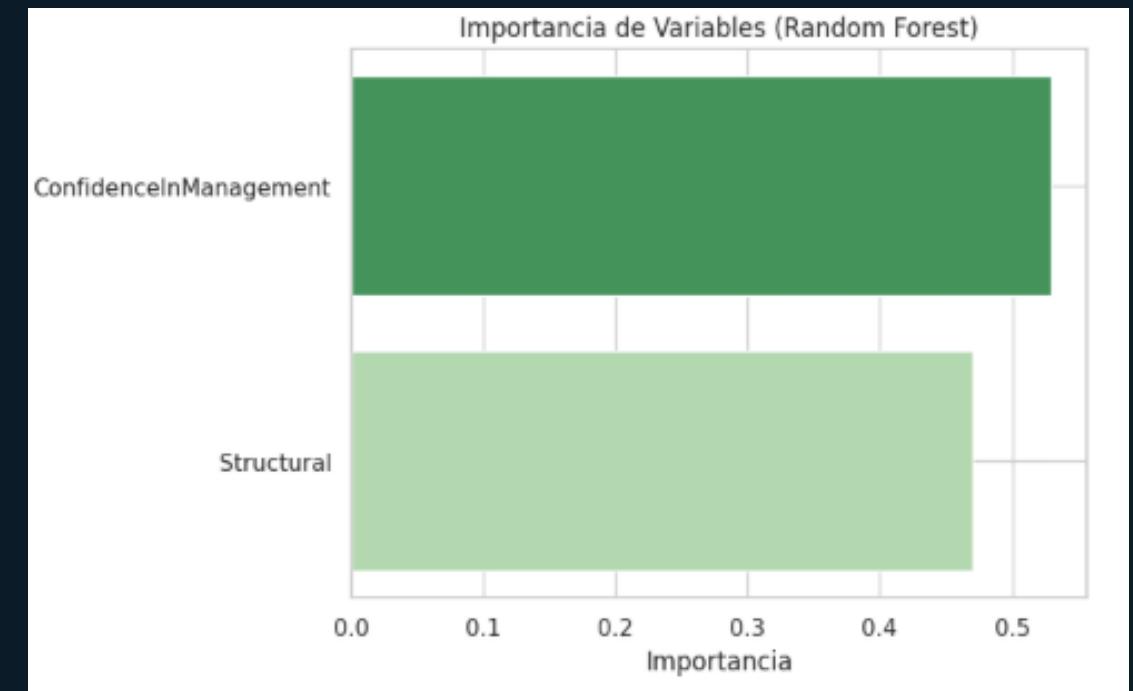
Este modelo revela que la variable `ConfidenceInManagement` (Confianza en la gestión) tiene un peso ligeramente mayor (52.97%) frente a `Structural` (47.03%) en la predicción de buena higiene. Esto sugiere que, aunque las condiciones estructurales son importantes, la percepción de una gestión confiable tiene un rol más decisivo en asegurar estándares de higiene elevados.

La mayoría de las probabilidades predichas para “Higiene Alta” son bajas, lo cual refuerza la idea de que la higiene elevada es menos común en el conjunto de datos. A su vez, indica que el modelo es conservador al momento de predecir buenas condiciones de higiene.

 Random Forest Classifier  
- Probabilidades de Higiene Alta  
Media: 0.11  
Mediana: 0.09  
Mínima: 0.00  
Máxima: 0.40



Importancia de Variables:  
`ConfidenceInManagement` 0.529663  
`Structural` 0.470337



# Modelo 2 : Regresion Logistica

Regresión Logística - Coeficientes

ConfidencelnManagement

0.457046

Structural

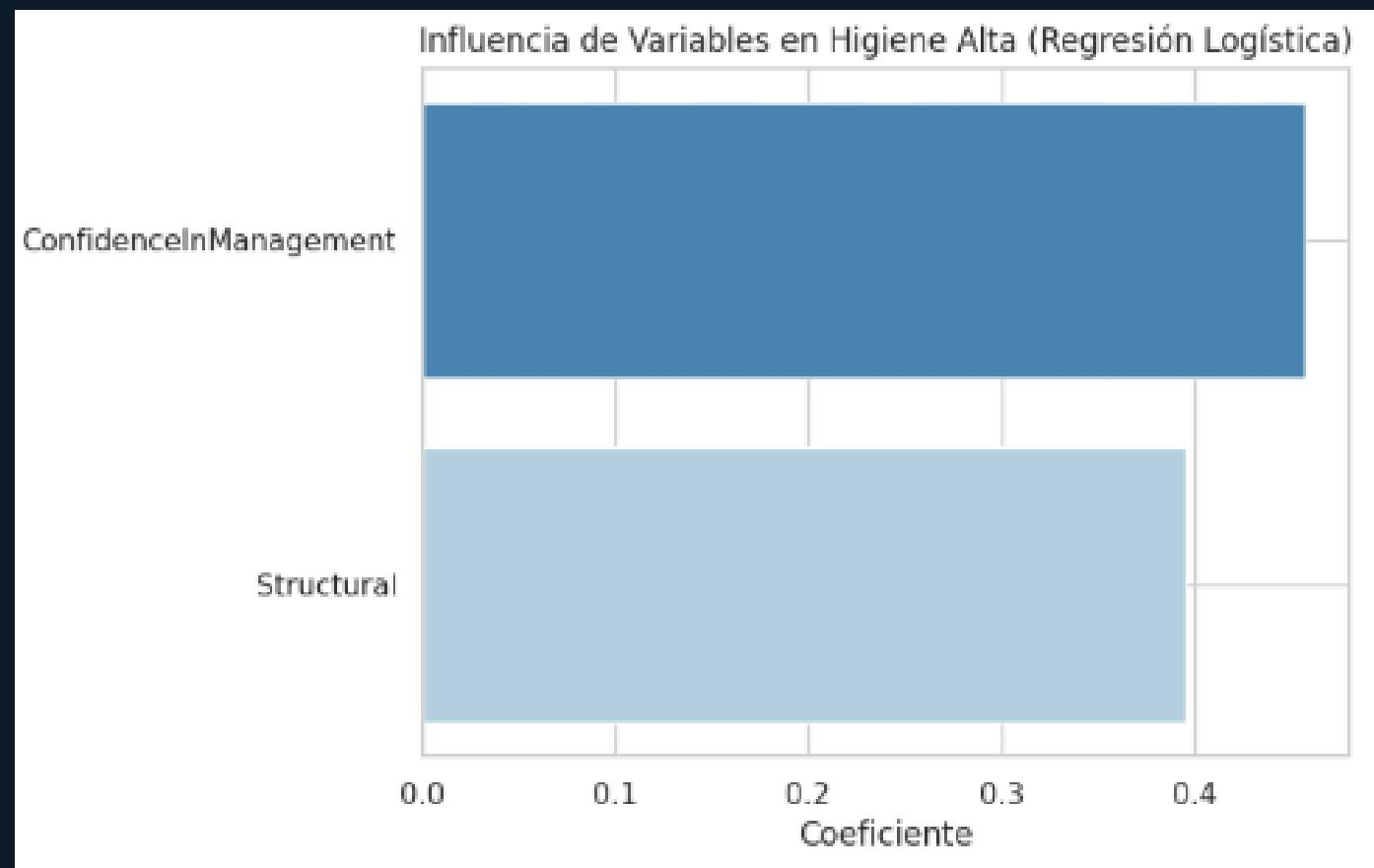
0.395479

Insight:

Ambas variables presentan coeficientes positivos, lo cual implica que mayores puntuaciones en confianza y estructura se asocian con mayor probabilidad de buena higiene. Pero nuevamente, ConfidencelnManagement (coef: 0.46) supera a Structural (coef: 0.39).

Además, la curva sigmoide muestra que al incrementar la confianza en la gestión, la probabilidad estimada de buena higiene crece de forma acelerada, alcanzando un valor máximo de 0.82.

Este comportamiento sugiere que la gestión impacta más fuertemente cuando está en niveles bajos a intermedios, lo que es útil para diseñar intervenciones enfocadas.



Máxima probabilidad estimada según confianza:  
0.82

# Modelo 3: Support Vector Machine (SVM)

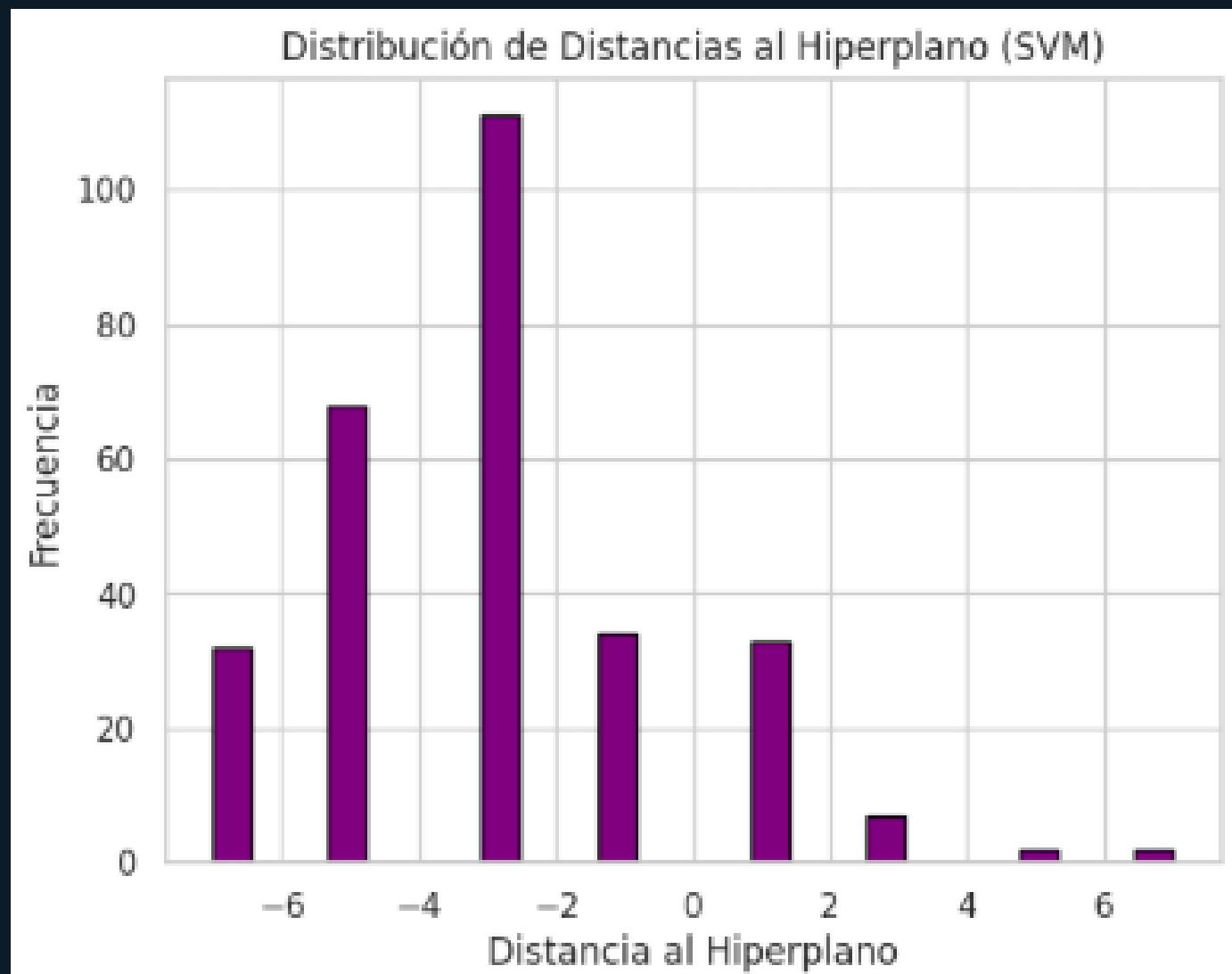
 SVM - Distancias al Hiperplano Rango: -7.0 a 7.0

Insight: El modelo SVM proyecta las muestras en función de su distancia al hiperplano de separación, el cual divide las clases (buena vs. mala higiene).

Rango de distancias: -7.0 a 7.0

Mayor concentración: entre -3 y -1

Esto significa que la mayoría de las predicciones están lejos del umbral de ambigüedad, lo que denota alta seguridad por parte del modelo. La mayoría de los establecimientos son claramente clasificados como “Higiene Baja”, lo cual concuerda con el patrón observado en los otros modelos.



# Modelo Neighbors (KNN)

## 4: K-Nearest

Insight: La matriz de confusión muestra que el modelo tuvo 10 falsos negativos y 13 falsos positivos, pero predijo correctamente 266 de 289 casos.

Higiene Baja: 235 correctos, 13 incorrectos

Higiene Alta: 31 correctos, 10 incorrectos

Además, las distancias a los vecinos más cercanos revelan:

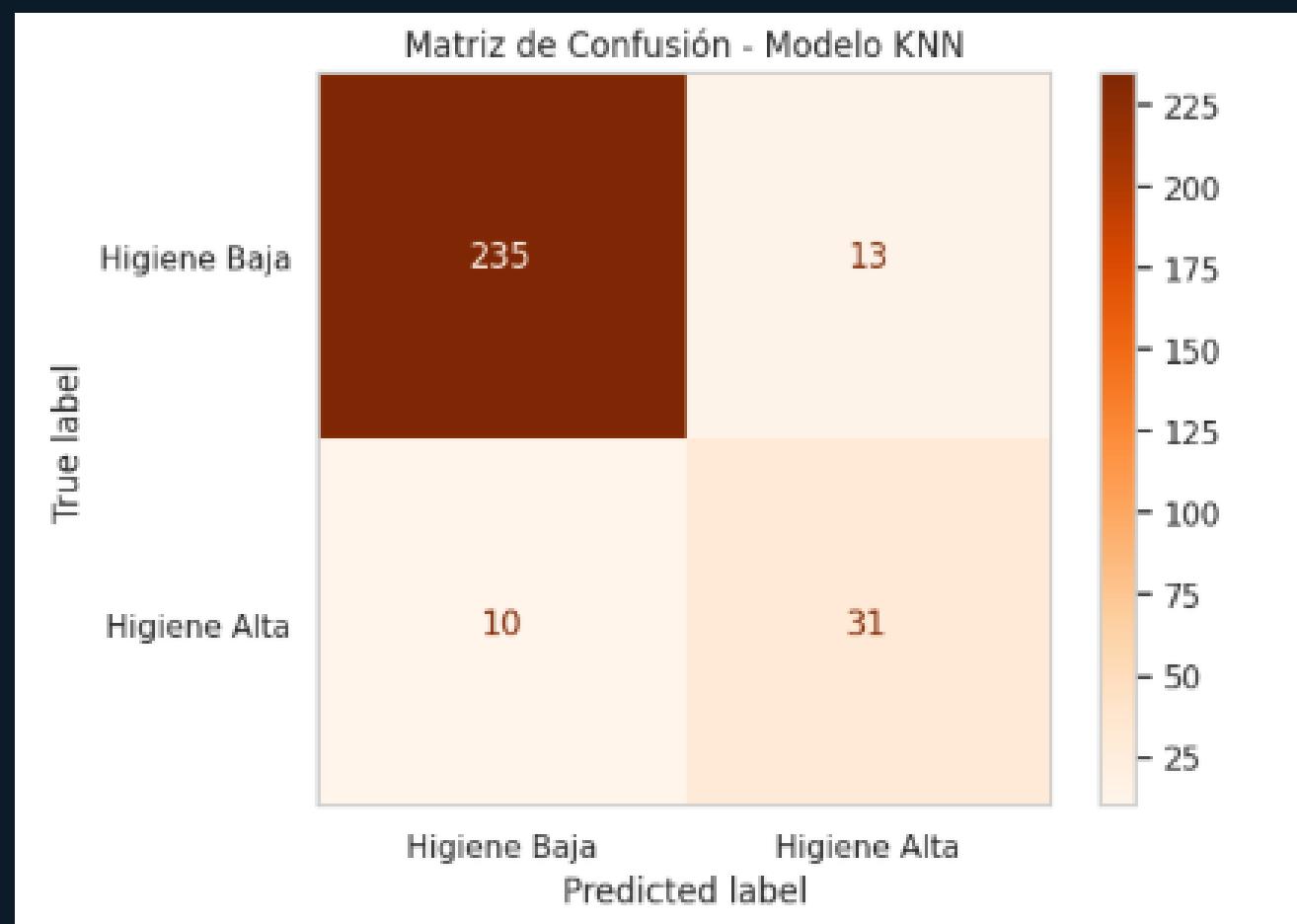
Promedio: 0.83

Máxima: 5.00

Moda: 0.03

Estas distancias nos indican que la mayoría de predicciones se basaron en vecinos cercanos, lo que sugiere que los casos están bien agrupados en el espacio de variables. Esto refuerza que hay patrones claros que permiten distinguir entre condiciones de higiene buenas y malas, sobre todo gracias a la variable de confianza.

- KNN - Matriz de Confusión
  - Higiene Baja
    - Correcto: 235 / Incorrecto: 13
  - Higiene Alta
    - Correcto: 31 / Incorrecto: 10



Distancias a vecinos más cercanos  
Promedio: 0.03  
Máxima: 5.00  
Moda: 0.00

# Evaluaciones

---

## Evaluacion de las modelos:

Todos los modelos obtuvieron un accuracy del 92%, lo que indica que 9 de cada 10 predicciones fueron correctas al clasificar entre establecimientos con higiene alta ( $>5$ ) y baja ( $\leq 5$ ).

La precisión para la clase 1 (higiene alta) fue 0.705, mientras que el recall fue 0.756, reflejando un buen equilibrio entre falsos positivos y falsos negativos.

El F1-score, que combina precisión y recall, fue constante en 0.729 para todos los modelos.

## Evaluacion de las predicciones:

Cada modelo predijo exactamente 245 casos como clase 0 (higiene baja) y 44 como clase 1 (higiene alta).

La cantidad de predicciones correctas fue de 266/289, mientras que solo 23 predicciones fueron incorrectas.

Esto demuestra que los modelos no solo tienen buen desempeño en términos de métricas globales, sino también una distribución estable entre clases.

No se observan sesgos fuertes hacia una clase, y los modelos generalizan bien sobre datos no vistos.

### Métricas Promedio de Todos los Modelos:

Accuracy	Precision (Alto)	Recall (Alto)	F1-score (Alto)
0.92	0.705	0.756	0.729

### Resumen Promedio de Predicciones:

Predicciones 0	Predicciones 1	Correctas	Incorrectas	Precisión global
245	44	266.0	23	0.92

# ANALISIS PCA

El análisis de componentes principales (PCA) permite visualizar los datos en un espacio 2D, facilitando el análisis exploratorio.

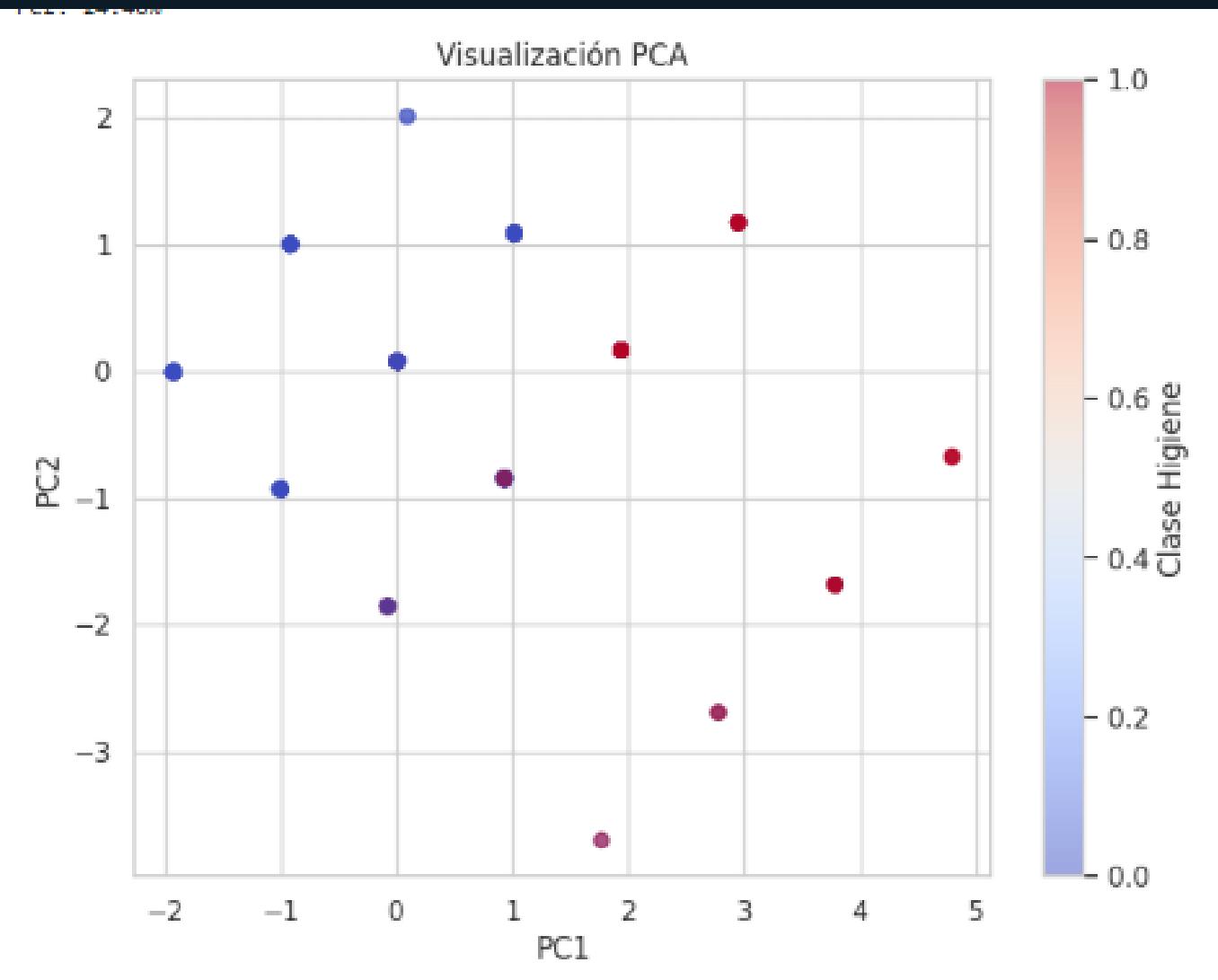
PC1 explica el 75.6% de la varianza, y PC2 el 24.4%, sumando un 100% de la varianza explicada, lo que valida la reducción.

Ambas variables originales (ConfidenceInManagement, Structural) contribuyen por igual a cada componente.

En la visualización, las clases (0 y 1) se separan claramente, lo que valida visualmente que estas características son informativas.

El espacio transformado por PCA preserva la estructura de los datos, confirmando que las variables seleccionadas son adecuadas para la clasificación

	Cargas de los componentes principales:	
ConfidenceInManagement	Structural	
PC1	0.707	0.707
PC2	-0.707	0.707



Varianza explicada:  
PC1: 75.60%  
PC2: 24.40%

# Optimización con GridSearchCV

---

Tras probar 12 combinaciones distintas de hiperparámetros con validación cruzada (`cv=5`), se obtuvo la siguiente configuración óptima para `RandomForestClassifier`:

```
{'max_depth': None, 'min_samples_split': 5, 'n_estimators': 100}
```

Esta combinación logró un accuracy de validación cruzada de 0.941, lo que representa una mejora respecto al 0.92 original. Además, la desviación estándar fue baja ( $\pm 0.017$ ), indicando resultados estables entre los pliegues.

El modelo optimizado supera al modelo base, siendo el candidato ideal para producción o implementación final.

 Mejor configuración de Random Forest: `{'max_depth': None, 'min_samples_split': 5, 'n_estimators': 100}`

 Accuracy CV mejor modelo: 0.941

 Accuracy promedio (CV): 0.935  $\pm$  0.017

# Clustering

- Los tres algoritmos de clustering muestran comportamientos distintos:

KMeans y Jerárquico obligan a 3 clusters, distribuyendo uniformemente las observaciones.

HDBSCAN, más flexible, encontró 2 grupos principales y varios puntos como ruido (sin asignar a ningún grupo).

Aunque no se usan etiquetas en el clustering, los patrones descubiertos apoyan los hallazgos supervisados, mostrando que los datos tienden a agruparse según las mismas características que afectan la higiene.

	PC1	PC2	KMeans	Jerárquico	HDBSCAN
	1.767.603	-3.697.973		2	1
	-1.924.731	-5.639	0	2	9
	1.937.740	164.498	2	1	2
	3.783.907	-1.681.669	2	1	0
	6.504	79.430	1	0	6

# Conclusión

---

El análisis con algoritmos de clasificación (Random Forest, Regresión Logística, SVM y KNN) permite identificar qué variables explican mejor los niveles de higiene en establecimientos.

## 1. La confianza en la gestión es el factor más influyente

Modelos como Random Forest y Regresión Logística mostraron que ConfidenceInManagement tiene más peso predictivo que Structural.

Insight: Invertir en buenas prácticas de gestión tiene mayor impacto que solo mejorar la infraestructura. La supervisión, liderazgo y cultura organizacional son claves.

## 2. La “Higiene Alta” es poco común

La mayoría de las predicciones se orientan a “Higiene Baja” en todos los modelos. Por ejemplo, en Random Forest, solo el 11% de las predicciones apuntan a higiene alta.

Insight: La buena higiene es una excepción. Esto sugiere posibles fallas sistémicas o estándares exigentes.

## 3. Pequeños cambios en la gestión, gran impacto en higiene

Modelos como la regresión logística muestran que mejoras mínimas en la confianza en la gestión aumentan significativamente la probabilidad de tener buena higiene.

Insight: Intervenciones focalizadas pueden tener alto retorno en términos de resultados sanitarios.

## 4. Predicciones claras y consistentes

SVM y KNN mostraron que la mayoría de los casos están bien definidos y separados, lo que indica estructuras sólidas en los datos.

### Recomendaciones

Fortalecer procesos de gestión y liderazgo.

Medir la percepción de confianza organizacional.

Replicar este análisis en otros distritos o rubros para validar hallazgos.