

**Carrera:**

**Ingeniería en Software**

**Asignatura:**

**Minería de Datos**

**Proyecto 1. Conociendo tus datos**

**Alumno:**

**Luis Gustavo Jaime Esquivel**

**Profesor:**

**Dr. Sergio Valadez Godínez**

**Fecha de entrega: 19/06/23**



### Instrucciones:

A partir de un caso práctico sobre minería de datos, elaborar un reporte que incluya:

- a) Descripción de la base de datos
- b) Tipos de atributos
- c) Medidas de tendencia central
- d) Medidas de dispersión
- e) Diagramas de caja y valores atípicos
- f) Regresión lineal
- g) Análisis de Componentes Principales (PCA)
- h) Reglas de separación de patrones
- i) Conclusiones
- j) Bibliografía

#### a) Descripción de la base de datos

Descripción de la base de datos a utilizar, así como la fuente de la misma.

De la Guía de Campo de la Sociedad Audobon; setas descritas en términos de características físicas; clasificación: venenosas o comestibles.

Este conjunto de datos incluye descripciones de muestras hipotéticas correspondientes a 23 especies de setas con branquias de las familias Agaricus y Lepiota (pp. 500-525). Cada especie se identifica como definitivamente comestible, definitivamente venenosa, o de comestibilidad desconocida y no recomendada. Esta última clase se ha combinado con la de venenosas. La Guía establece claramente que no existe una regla sencilla para determinar la comestibilidad de una seta; ninguna regla como "hojas tres, que sea" para el roble y la hiedra venenosos.

Fuente de información:

- Unknown. (1981). Mushroom [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5959T>
- UCI Machine Learning Repository. (s. f.). <https://archive.ics.uci.edu/dataset/73/mushroom>

### b) Tipos de atributos

Atributos de la base de datos y el tipo de dato de cada uno de los atributos.

Attribute Information	Values
Class	edible=e, poisonous=p
*----- Features related to CAP	----- *
Cap-shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
Cap-surface	fibrous=f, grooves=g, scaly=y, smooth=s
Cap-color	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
*----- Features related to GILL	----- *
Gill-attachment	attached=a, descending=d, free=f, notched=n
Gill-spacing	close=c, crowded=w, distant=d
Gill-size	broad=b, narrow=n
Gill-color	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
*----- Features related to STALK	----- *
Stalk-shape	enlarging=e, tapering=t
Stalk-root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
Stalk-surface-above-ring	fibrous=f, scaly=y, silky=k, smooth=s
Stalk-surface-below-ring	fibrous=f, scaly=y, silky=k, smooth=s
Stalk-color-above-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y

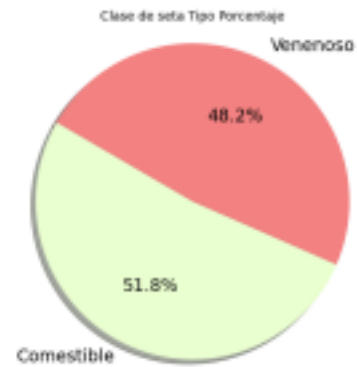
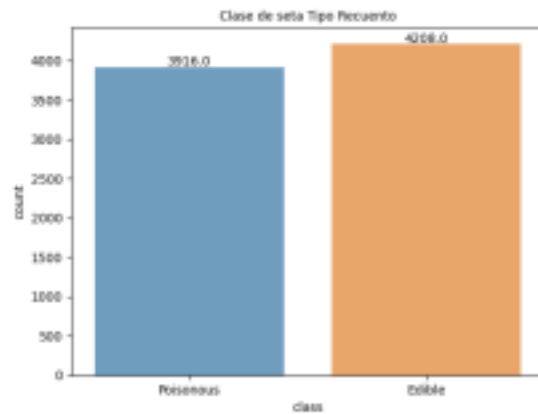
Stalk-color-below-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
*----- Features related to VEIL	----- *
Veil-type	partial=p, universal=u
Veil-color	brown=n, orange=o, white=w, yellow=y
Ring-number	none=n, one=o, two=t
Ring-type	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
*----- Features related to MISC	----- *
Bruises	bruises=t, no=f
Odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
Spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
Population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y

3

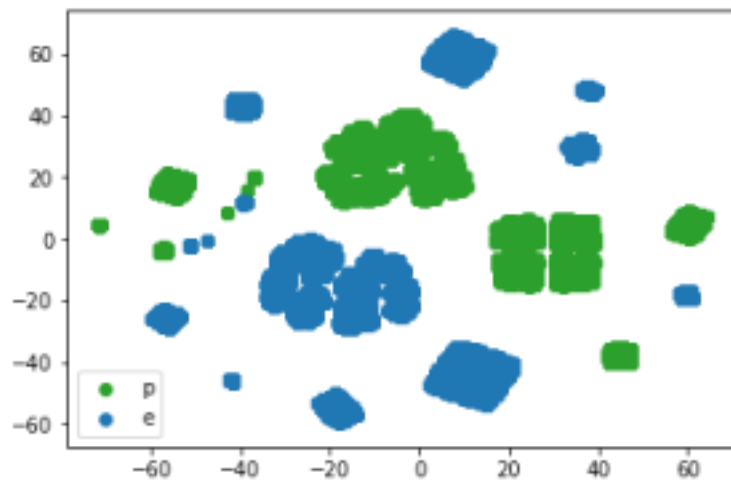


Habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d
---------	--

Distribución grafica en base si es venenosa o comestible la seta.  
El número de setas por clase - e=Comestibles, p=Venenosas.



Distribución grafica en base si es venenosa o comestible la seta.



Calcular la media, mediana y moda para cada uno de los atributos de la base de datos. Poner código fuente del cálculo y el resultado.

NOTA:

Se necesita convertir estos datos en datos numéricos. Se ha utilizado una gran biblioteca de Sklearn

LabelCoder para convertir todas las categorías en valores numéricos.

```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
for col in df.columns:
    df[col] = le.fit_transform(df[col])
df.head()
print(df)
```

Analizando el diagrama de caja y la matriz de correlación, los datos no se pueden codificar con etiquetas. Este método utiliza diferentes métodos como la media, la covarianza y otras técnicas matemáticas para encontrar la relación entre las características. Y nuestro conjunto de datos son datos categóricos para los que no es necesario calcular una media o una varianza.

Atributos.	MEDIA	MODA	MEDIANA
Class	0.48202855736090594	0	0.0
Cap-shape	3.348104382077794	5	3.0
Cap-surface	1.82767109798129	3	2.0
Cap-color	4.504677498769079	4	4.0
Bruises	4.0	0.4155588380108321	0
Odor	0	4.144756277695716	5
Gill-attachment	0.9741506646971935	1	
Gill-spacing	0.16149679960610536	0	
Gill-size	0.30920728705071393	0	
Gill-color	4.81068439192516	0	
Stalk-shape	0.5672082717872969	1	
Stalk-root	1.1097981290004924	1	
Stalk-surface-above-ring	1.5750861644510095	2	
Stalk-surface-below-ring	1.603643525356967	2	
Stalk-color-above-ring	5.816346627277204	7	
Stalk-color-below-ring	5.794682422451994	7	
Veil-type	0.0	0	
Veil-color	1.965534219596258	2	
Ring-number	1.069423929098966	1	

Ring-type	2.291974396848843	4	
Spore-print-color	3.5967503692762186	7	
Population	3.6440177252584935	4	

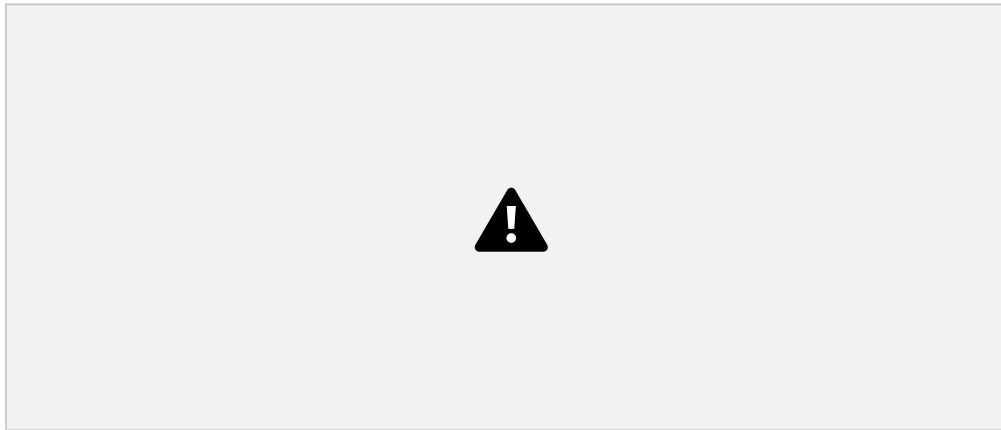
5



Habitat	1.5086164451009354	0	
---------	--------------------	---	--

CODIGO.

El siguiente código se aplicó para cada uno de los atributos de la base de datos:



#### d) Medidas de dispersión

Cálculo del rango, los cuartiles, la varianza, la desviación estándar y el rango intercuartílico para cada uno de los atributos de la base de datos.

Atributos.	RANGO	CUARTILES	VARIANZA	DESVIACION ESTANDAR	RANGO INTERCUARTILICO
Class	1	0.0, 0.0, 1.0	0.2496770272494697	0.499676922870638 2	1.0
Cap-shape	5	2.0, 3.0, 5.0	2.5735549984593793	1.604230344576295 3	3.0
Cap surface	3	0.0, 2.0, 3.0	1.5124002384507393	1.229796828118669	3.0
Cap-color	9	3.0, 4.0, 8.0	6.4804064814188544	2.545664251510567	5.0

Bruises	1	0.0, 0.0, 1.0	0.2428696901619190 6	0.49281811062695235	1.0
Odor	8	2.0, 5.0, 5.0	4.4251312921503985	2.103599603572504 4	3.0
Gill attachment	1	1.0, 1.0, 1.0	0.02518114716720957 7	0.15868568671184424	0.0
Gill-spacing	1	0.0, 0.0, 0.0	0.1354155833230908	0.36798856411998837	0.0
Gill-size	1	0.0, 0.0, 1.0	0.2135981406854513 8	0.46216678881703666	1.0
Gill-color	11	2.0, 5.0, 7.0	12.532598793055257	3.540141069654606	5.0

6



Stalk-shape	1	0.0, 1.0, 1.0	0.2454830482033649	0.495462458924351 7	1.0
Stalk-root	4	0.0, 1.0, 1.0	1.1258074924829593	1.061040759105398	1.0
Stalk surface above-ring	3	1.0, 2.0, 2.0	0.3861641358548602 6	0.621421061644083 8	1.0
Stalk surface below-ring	3	1.0, 2.0, 2.0	0.4568848044866364	0.675932544331634	1.0
Stalk-col or above-ring	8	6.0, 7.0, 7.0	3.6161975834438196	1.901630243618306 6	1.0
Stalk-col or below-ring	8	6.0, 7.0, 7.0	3.637312934595018	1.90717407034466	1.0
Veil-type	0	0.0, 0.0, 0.0	0.0	0.0	0.0
Veil-color	3	2.0, 2.0,	0.0588810415419685	0.24265416036402201	0.0



		2.0	1		
Ring number	2	1.0, 1.0, 1.0	0.07346687641410427	0.2710477382567585	0.0
Ring-type	4	0.0, 2.0, 4.0	3.245622443460678	1.8015611128853437	4.0
Spore print-color	8	2.0, 3.0, 7.0	5.67638530400651	2.382516590499741	5.0
Population	5	3.0, 4.0, 4.0	1.5675159110595227	1.252004756803872	1.0
Habitat	6	0.0, 1.0, 2.0	2.957951360024979	1.7198695764577554	2.0

7

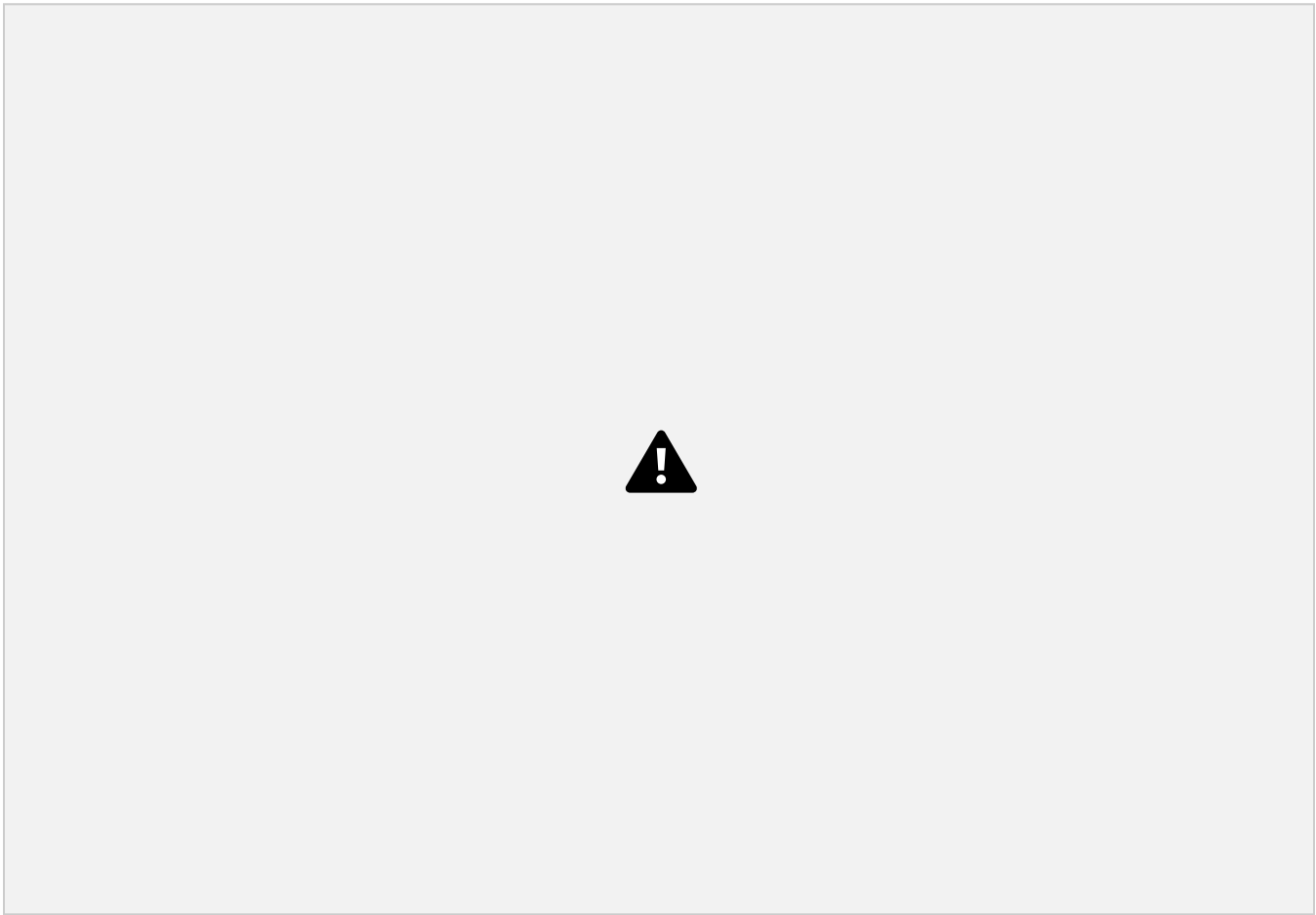


CODIGO.

El siguiente código se aplicó para cada uno de los atributos con respecto a el rango, los cuartiles, la varianza, la desviación estándar y el rango intercuartílico.



Matriz de correlación.



9



Matriz de covarianza.



10



CODIGO.

El siguiente código se aplicó para la creación de la matriz de correlación y covarianza.





Calcular los diagramas de caja y los valores atípicos para cada uno de los atributos de la base de datos.

CODIGO.

El siguiente código se utilizó para la visualización de valores atípicos si los hubiera.



Graficar los diagramas de caja para todos los atributos.

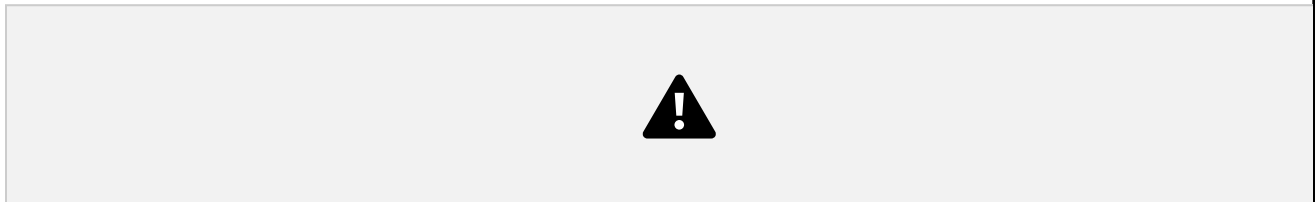
Diagrama de caja de valores atípicos para el conjunto de datos. En todos los atributos y sus valores atípicos.



#### f) Regresión lineal

Cálculo de los valores de la regresión lineal (coeficientes y cruces de las líneas, coeficiente de determinación y correlación) para todos los atributos de la base de datos.

Class:



Cap-shape:



Cap-surface:



Cap-color:



Bruises:



Odor:



Gill-attachment:



Gill-spacing:





Gill-size:



Gill-color:



Stalk-shape:



Stalk-root:



Stalk-surface-above-ring:

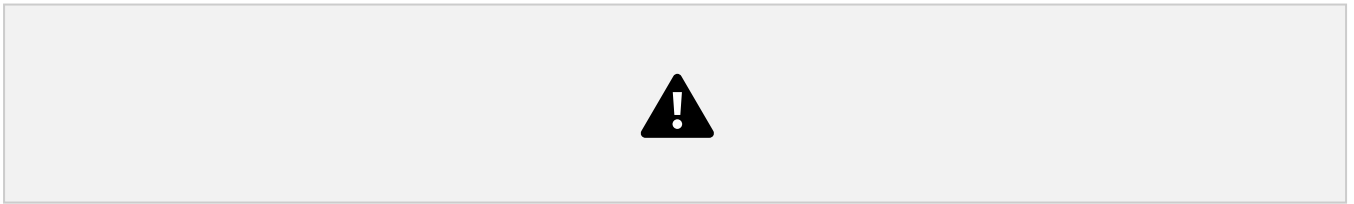


Stalk-surface-below-ring:





Stalk-color-above-ring:



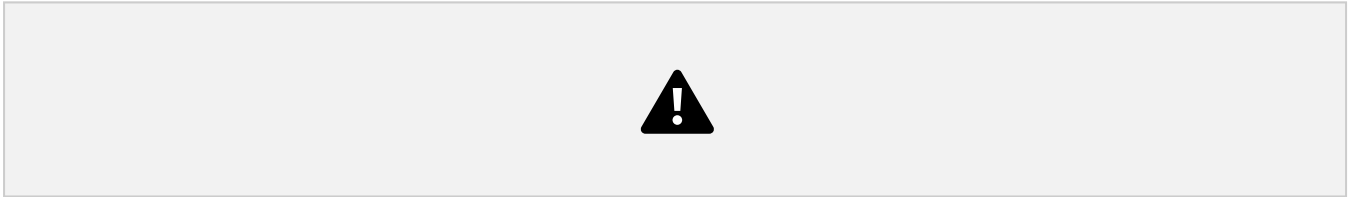
Stalk-color-below-ring:



Veil-type:



Veil-color:

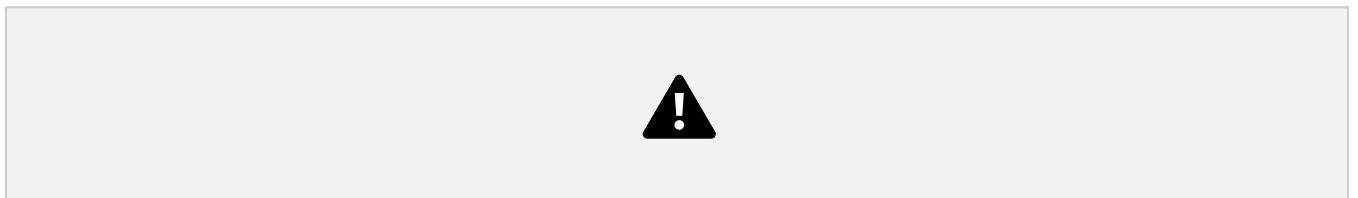


Ring-number:

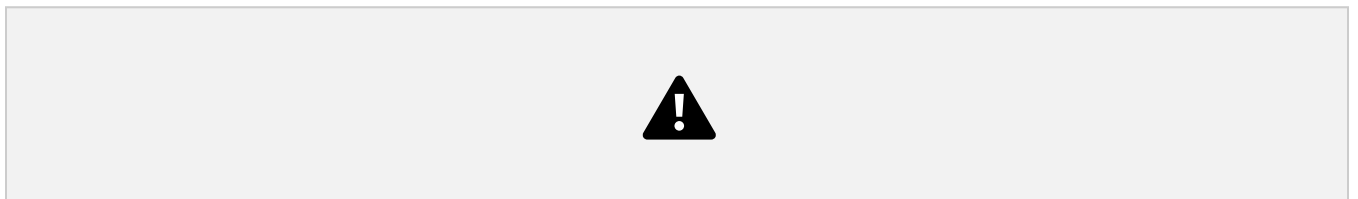


Ring-type:


15



Spore-print-color:



Population:



Habitat:

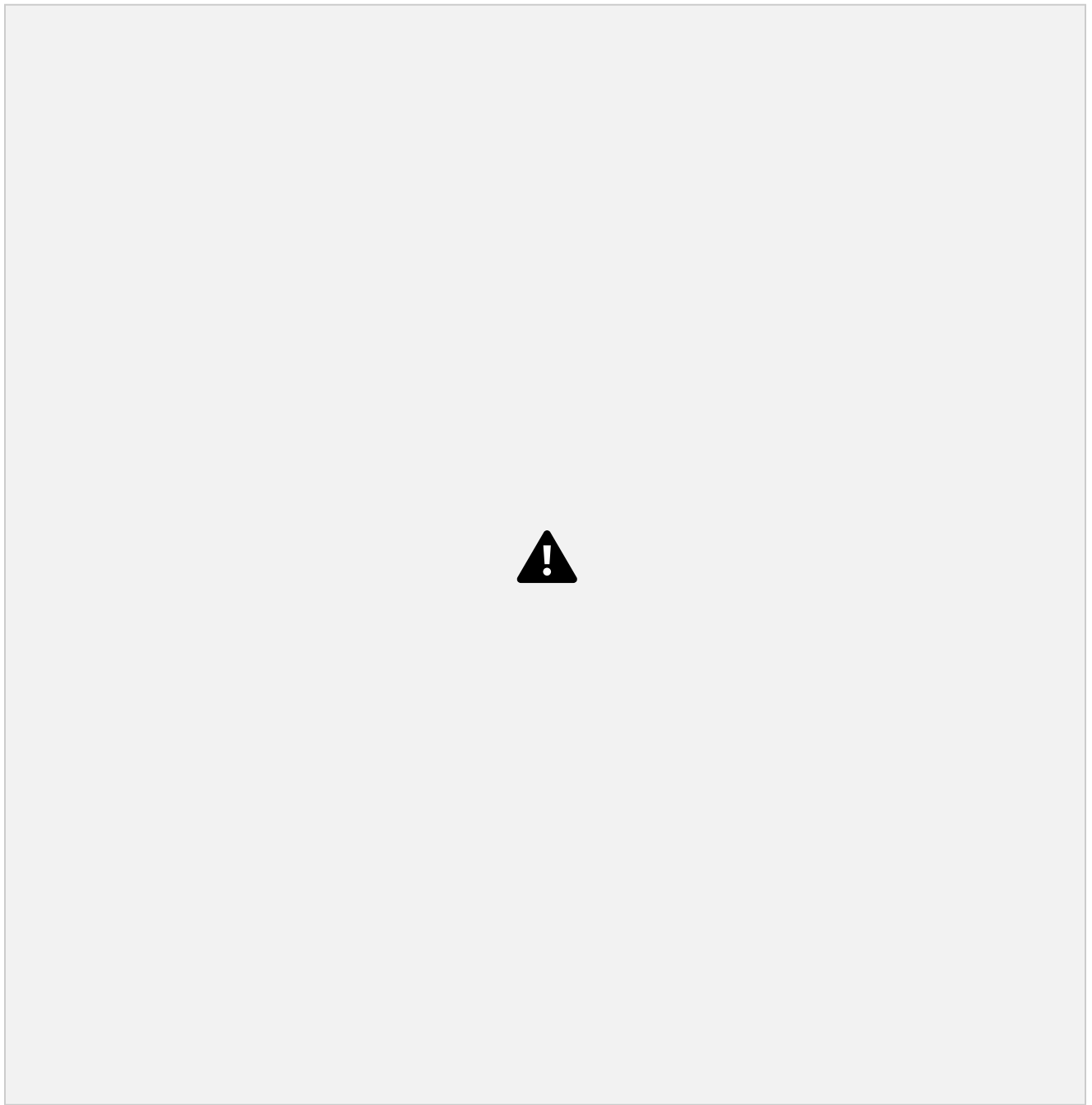


CODIGO:



16

Este código pertenece para el cálculo de regresión lineal (coeficientes y cruces de las líneas, coeficiente de determinación y correlación).



17



Gráfico de la distribución de los datos con su respectiva línea de regresión.



Resultados de la regresión lineal:



CODIGO:

Este código pertenece para el grafico de distribución de la línea de regresión.



18



g) Análisis de Componentes Principales (PCA)

Calcular los componentes principales de la base de datos y su nivel de varianza.

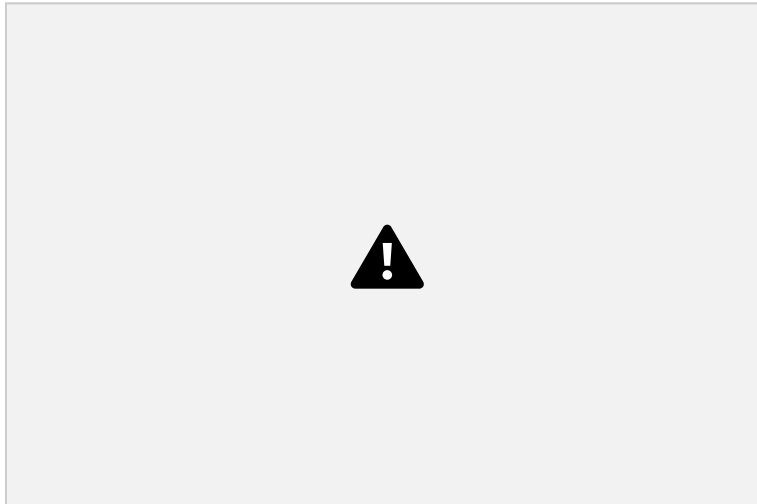
Pasando por el análisis de componentes principales (PCA) y los resultados se pueden ver a continuación. Aquí podemos ver que la varianza máxima ha sido capturada por 45 componentes.



Dado que sólo necesitamos los dos primeros componentes principales.



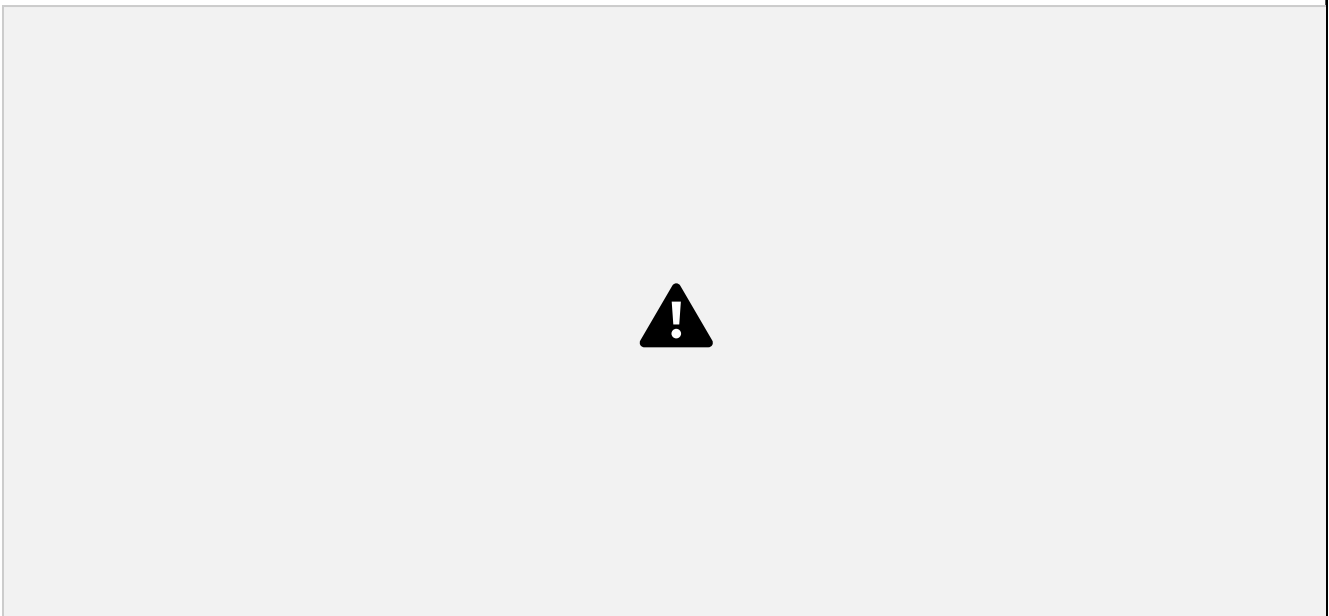
Graficar la distribución de las nuevas dimensiones de la base de datos.



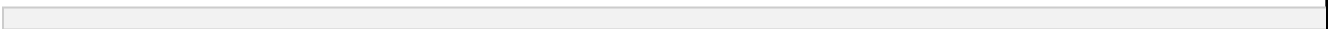
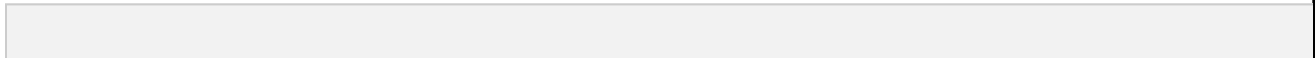
Poner código fuente del cálculo y el resultado.

CODIGO:

Este código pertenece al cálculo de los componentes principales del PCA y los dos componentes necesarios a utilizar.



Este código es el responsable de arrojar el porcentaje de la varianza de los datos que es igual a: 0.981



h) Reglas de separación de patrones



Crear un algoritmo con reglas IF-THEN-ELSE que permita separar lo mejor posible las clases de los patrones según el análisis estadístico realizado.





Resultado del rendimiento del clasificador:

### i) Conclusiones

En conclusión, este estudio de caso sobre minería de datos ha mostrado cómo pueden utilizarse técnicas y medidas estadísticas para analizar y descubrir patrones en una base de datos específica. Se describieron los distintos atributos de la base de datos y se identificó su relevancia en el análisis. Se utilizaron medidas de tendencia central y dispersión para comprender la ubicación central y la variabilidad de los datos en cada atributo. Se utilizaron diagramas de caja para identificar posibles valores atípicos y regresión lineal para establecer relaciones entre variables. El análisis de componentes principales (PCA) se utilizó para reducir la dimensionalidad de los datos y visualizar patrones y estructuras subyacentes. Por último, se utilizaron reglas de





separación de patrones para identificar patrones y relaciones específicos en los datos, lo que permitió clasificarlos y tomar decisiones.

Este caso práctico ha demostrado el poder de la minería de datos para extraer información valiosa de grandes conjuntos de datos. Las técnicas y medidas utilizadas en este estudio de caso pueden aplicarse a otras bases de datos para conocer mejor los datos y tomar mejores decisiones.

#### j) Bibliografía

- Unknown. (1981). Mushroom [Data set]. UCI Machine Learning Repository.  
<https://doi.org/10.24432/C5959T>
- UCI Machine Learning Repository. (s. f.). <https://archive.ics.uci.edu/dataset/73/mushroom>
- Layton, R. (2015). Learning Data Mining with Python.
- McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter. O'Reilly Media.