Practica de laboratorio 1: Pruebas de hipótesis

Machine learning

Lauro Reyes Cocoletzi

Castro Elvira Diego

UPIIT: Unidad Profesional Interdisciplinaria en Ingeniería Campus Tlaxcala Instituto Politécnico Nacional, Tlaxcala, Tlaxcala, México 9000

Ingeniera en Inteligencia Artificial

16 de septiembre de 2023

Practica de laboratorio 1: Pruebas de hipótesis

A. Marco teórico

Definición 1.1. Una hipótesis es una afirmación acerca de un parámetro poblacional.

El problema de prueba de hipótesis es un procedimiento basado en una muestra poblacional sobre la cual se realizan dos afirmaciones (o hipótesis) en las que se debe de decidir cuál es verdadera. En general estas dos afirmaciones son excluyentes.

Definición 1.2. Las dos hipótesis en un problema de prueba de hipótesis son llamadas hipótesis nula e hipótesis alternativa, las cuales son denotadas como H0 y H1 respectivamente.

Si θ denota un parámetro poblacional, la forma general de la hip´otesis nula y alternativa es el siguiente:

$$H_0 : \theta \in \Theta_0 \quad \text{y} \quad H_1 : \theta \in \Theta_0^c,$$

donde Θ_0 es algún subconjunto del espacio parametral Θ y Θ^c_0 es el complemento. En un problema de prueba de hipótesis, después de observar la muestra el experimentador debe decidir no rechazar H_0 y aceptarla como verdadera, o rechazar H_0 y aceptar H_1 como verdadera.

Definición 1.3. Un procedimiento de prueba de hipótesis (o prueba de hipótesis) es una regla que especifica

i) Para qué valores de la muestra se toma la decisión de aceptar H0 como verdadera. ii) Para qué valores de la muestra H0 es rechazada y H1 es aceptada como verdadera.

El subconjunto del espacio muestral para el cual H0 debe ser rechazada es llamado región de rechazo o región crítica. El complemento de la región de rechazo es llamado región de aceptación.

En ocasiones nos preocupamos en la distinción entre rechazar H0 y aceptar H1. En el primer caso, no hay nada implícito sobre que declaramos como aceptado, solo que la afirmación dada por H0 está siendo rechazada. De manera similar se puede hacer la distinción entre aceptar H0 y no rechazar H0. En la primera frase el experimentador está dispuesto a aceptar la afirmación especificada en H0, mientras que la segunda frase implica que realmente no creemos en H0, pero no tenemos evidencia para rechazarla.

No nos preocuparemos por estas cuestiones, en lo que sigue veremos el problema de pruebas de hipótesis como un problema en el cual una de las dos hipótesis es tomada como verdadera.

Definición 1.4. Consideramos el juego de hipótesis H0: θ E Θ0 vs H1: θ E Θc0. Si θ E Θ0 pero la prueba de hipótesis decide incorrectamente rechazar H0, se dice que se ha cometido un error de tipo I. Si, por otro lado θ E Θc0 pero se decide incorrectamente aceptar la hipótesis H0, se dice que se ha cometido un error de tipo II.

El proceso de una prueba de significancia generalmente sigue estos pasos:

1. Formulación de hipótesis.
2. Selección de un nivel de significancia (α): Este valor representa el umbral de probabilidad utilizado para determinar si se rechaza la hipótesis nula.
   Un valor comúnmente utilizado es α = 0.05, lo que significa que estamos dispuestos a aceptar un 5% de probabilidad de cometer un error de tipo I al rechazar la hipótesis nula cuando es verdadera.
3. Recopilación de datos: Se recopilan datos relevantes a partir de una muestra o población correspondiente al problema o información a analizar.
4. Cálculo estadístico de prueba: Se basa en los datos recopilados, la elección de la estadística depende del tipo de prueba y la naturaleza de los datos.
5. Determinación de un valor p: El valor p representa la probabilidad de obtener los resultados observados (o resultados más extremos) si la hipótesis nula fuera cierta. Un valor p bajo (generalmente menor que α) sugiere evidencia en contra de la hipótesis nula.

6. Toma de decisión: Si el valor p es menor que el nivel de significancia α, se rechaza la hipótesis nula en favor de la hipótesis alternativa. Si el valor p es mayor que α, no se rechaza la hipótesis nula.
7. Interpretación: Se interpreta la decisión en el contexto del problema en estudio y se llega a una conclusión.

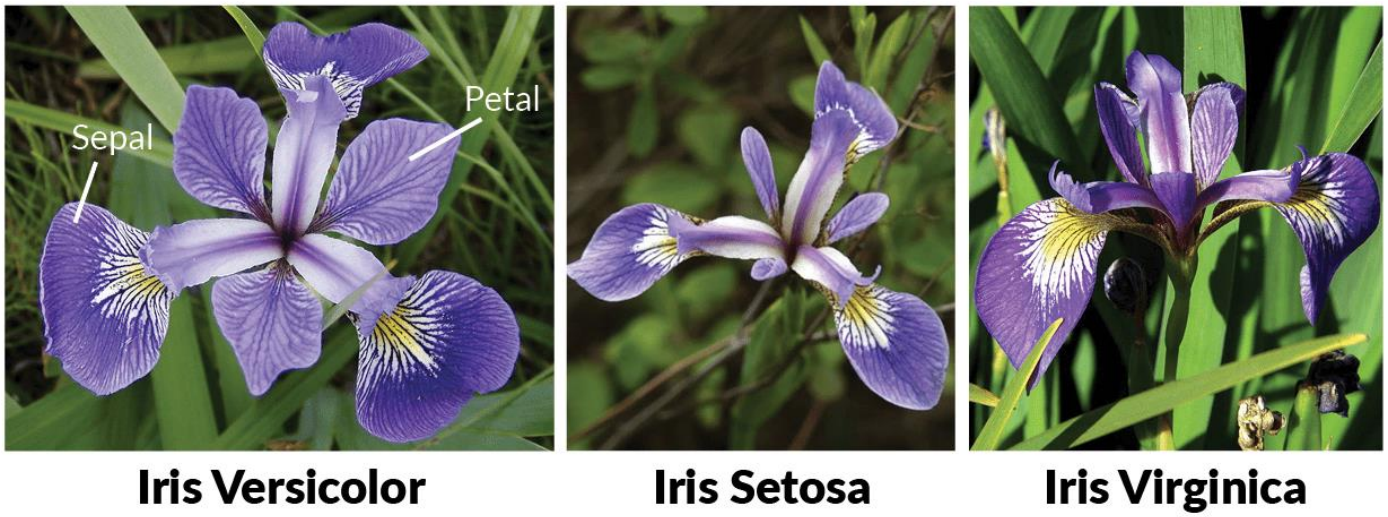
B. Conjunto de pruebas (muestra)

El conjunto de datos flor Iris o conjunto de datos iris de Fisher es un conjunto de datos multivariante introducido por Ronald Fisher en su artículo de 1936 es un ejemplo de análisis discriminante lineal.

Consiste en una colección de datos para cuantificar la variación morfológica de la flor Iris de tres especies relacionadas, dos de las tres especies se recolectaron en la Península de la Gaspesia , todos son de la misma pastura, y recolectado el mismo día y medidos al mismo tiempo por la misma persona con el mismo aparato.

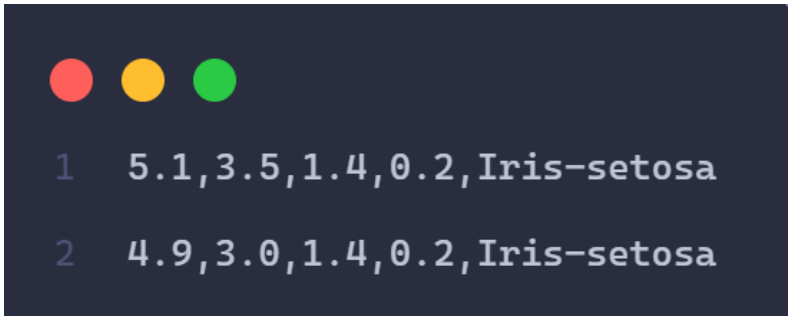
El conjunto de datos contiene 50 muestras de cada una de tres especies de Iris (Iris setosa, Iris virginica e Iris versicolor). Se midió cuatro rasgos de cada muestra: el largo y ancho del sépalo y pétalo, en centímetros.

Basado en la combinación de estos cuatro rasgos, Fisher desarrolló un modelo discriminante lineal para distinguir entre una especie y otra.



C. Procedimiento:

1. Obtener el conjunto de pruebas (datos) de iris del repositorio oficial en la liga siguiente: https://archive.ics.uci.edu/dataset/53/iris
   Descargamos los archivos que nos proporciona el dataset, sin embargo el que nos sera de interés es el de iris.data, el cual nos da la siguiente forma:



Importante destacar que los datos están divididos por comas, el primer valor sera la longitud del sepal, el segundo el ancho del sepal, el tercero la longitud del petalo, el cuarto la anchura del petalo y el quinto la clase o tipo que pertenece, en este caso tenemos 3 tipos que serán: Iris setosa, Iris virginica e Iris versicolor

2. Desplegar la información de forma gráfica para plantear posibles hipótesis con respecto a los datos observados. Comenzamos leyendo el archivo y etiquetando la información que contiene

```
archivo_iris = "iris.data"
nombres_columnas = ["sepal_length", "sepal_width", "petal_length", "petal_width", "class"]
iris_data = pd.read_csv(archivo_iris, header=None, names=nombres_columnas)
iris_data.head(15)
```

Posteriormente realizamos diferentes gráficos y datos para analizarlos desde diferentes puntos de vista

```
sns.set(style="dark")
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(20, 13))
features = ["sepal_length", "sepal_width", "petal_length", "petal_width"]

# Iterar a través de las características
for i, feature in enumerate(features):
    row, col = i // 2, i % 2
    ax = axes[row][col]

    # Gráfico de caja
    sns.boxplot(x="class", y=feature, data=iris_data, ax=ax)
    ax.set_title(feature)
    ax.set_xlabel("Clase")
    ax.set_ylabel(feature)

    # Estadísticas resumidas
    summary_stats = iris_data.groupby("class")[feature].agg(["mean", "var", "std"])
    print(f"\nEstadísticas para {feature}:")
    print(summary_stats)

# histogramas de densidad
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(13, 8))
for i, feature in enumerate(features):
    row, col = i // 2, i % 2
    ax = axes[row][col]
    sns.histplot(data=iris_data, x=feature, hue="class", element="step", common_norm=False, kde=True, ax=ax)
    ax.set_title(f"Densidad de {feature}")
    ax.set_xlabel(feature)
    ax.set_ylabel("Densidad")

plt.tight_layout()
plt.show()
```

3. Calcular de forma inicial, media, varianza, desviación estándar, observar la dispersión de los datos.

Primero creamos las funciones para calcular la media, varianza y desviación estándar

```python
# Función para calcular la media
def custom_mean(data):
    mean = sum(data) / len(data)
    return mean

# Función para calcular la varianza
def custom_var(data):
    n = len(data)
    mean = custom_mean(data)
    squared_diff = [(x - mean) ** 2 for x in data]
    variance = sum(squared_diff) / n
    return variance

# Función para calcular la desviación estándar
def custom_std(data):
    return np.sqrt(custom_var(data))
```

Posteriormente la utilizamos en cada rasgo que se nos proporciona

```python
s_l_mean = custom_mean(iris_data["sepal_length"])
s_l_var = custom_var(iris_data["sepal_length"])
s_l_std = custom_std(iris_data["sepal_length"])

s_w_mean = custom_mean(iris_data["sepal_width"])
s_w_var = custom_var(iris_data["sepal_width"])
s_w_std = custom_std(iris_data["sepal_width"])

p_l_mean = custom_mean(iris_data["petal_length"])
p_l_var = custom_var(iris_data["petal_length"])
p_l_std = custom_std(iris_data["petal_length"])

p_w_mean = custom_mean(iris_data["petal_width"])
p_w_var = custom_var(iris_data["petal_width"])
p_w_std = custom_std(iris_data["petal_width"])
```

Por último, mostramos una tabla y un gráfico de barras para su posterior análisis

```python
# diccionario
data = {
    "Feature": ["sepal_length", "sepal_width", "petal_length", "petal_width"],
    "Mean": [s_l_mean, s_w_mean, p_l_mean, p_w_mean],
    "Var": [s_l_var, s_w_var, p_l_var, p_w_var],
    "Std": [s_l_std, s_w_std, p_l_std, p_w_std]
}

stats_df = pd.DataFrame(data)
stats_df.set_index("Feature", inplace=True)
print(stats_df)
# Crear un gráfico de barras
fig, ax = plt.subplots(figsize=(10, 6))
stats_df.plot(kind="bar", ax=ax, rot=0)
ax.set_ylabel("Valor")
ax.set_title("Estadísticas para las características")

plt.tight_layout()
plt.show()
```

4. Implementar las herramientas correlación de pearson para realizar una prueba de hipótesis simple.
   Para realizar la correalacion de Pearson usamos el código visto en clase, donde le agregamos en una función y modificamos para que lea los datos

```python
import numpy as np
def pearson(x, y):
    iris_x = np.array(iris_data[x])
    iris_y = np.array( iris_data[y])

    # Calcular la media de cada variable
    media_iris_x = np.mean(iris_x)
    media_iris_y = np.mean(iris_y)

    # Calcular las desviaciones de cada variable con respecto a su media
    desviaciones_iris_x = iris_x - media_iris_x
    desviaciones_iris_y = iris_y - media_iris_y

    # Calcular el coeficiente de correlación de Pearson
    correlacion = np.sum(desviaciones_iris_x * desviaciones_iris_y) / \
        np.sqrt(np.sum(desviaciones_iris_x**2) * np.sum(desviaciones_iris_y**2))

    # Imprimir el coeficiente de correlación
    print(f"\nCoeficiente de correlación de Pearson entre {x} y {y} es: {correlacion:.2f}")
        # Interpretar la correlación
    if correlacion > 0:
        print(f"Hay una correlación positiva: A medida que {x} aumenta, {y} aumenta.")
    elif correlacion < 0:
        print(f"Hay una correlación negativa: A medida que {x} aumenta, {y} disminuye")
    else:
        print(f"No hay una correlación lineal significativa entre {x} y {y}")

```

Posteriormente lo pondremos a prueba con todas las combinaciones posibles de pruebas que este pueda tener

```python
# Lista de características
features = ["sepal_length", "sepal_width", "petal_length", "petal_width"]

for i in range(len(features)):
    for j in range(i+1, len(features)):
        feature1 = features[i]
        feature2 = features[j]
        correlation = pearson(feature1, feature2)
```

5. Implementar la herramienta de ANOVA para realizar una prueba de hipótesis múltiple.
   En el caso de ANOVA después de una búsqueda, el código inicial podemos decir que está diseñado para comparar las medias de diferentes grupos utilizando datos unidimensionales Sin embargo, el conjunto de datos de iris es un conjunto de datos multivariado en el que tienes múltiples características, por lo que se tuvo que adaptar el código para que separe por clases y lo realice de manera segmentada

```python
import scipy.stats as stats

grupos = iris_data["class"].unique()
caracteristicas = ["sepal_length", "sepal_width", "petal_length", "petal_width"]

for caracteristica in caracteristicas:
    print(f"ANALISIS DE ANOVA PARA: {caracteristica}\n")

    grupo_data = [iris_data[iris_data["class"] == grupo][caracteristica].values for grupo in grupos]

    # media total
    mean_total = np.mean(iris_data[caracteristica])

    #  suma de cuadrados entre grupos
    ss_between = sum(len(grupo) * (np.mean(grupo) - mean_total)**2 for grupo in grupo_data)

    #  suma de cuadrados dentro de grupos
    ss_within = sum(sum((x - np.mean(grupo))**2 for x in grupo) for grupo in grupo_data)

    # estadístico F
    df_between = len(grupos) - 1
    df_within = len(iris_data) - len(grupos)
    f_statistic = (ss_between / df_between) / (ss_within / df_within)

    # valor p utilizando la distribución F
    p_value = 1 - stats.f.cdf(f_statistic, df_between, df_within)

    print(f"Estadístico F: {f_statistic}")
    print(f"Valor p: {p_value}")

    if p_value < 0.05:
        print("Rechazamos la hipótesis nula: Hay diferencias significativas entre los grupos.\n")
    else:
        print("No rechazamos la hipótesis nula: No hay diferencias significativas entre los grupos.\n")

```

D. Resultados

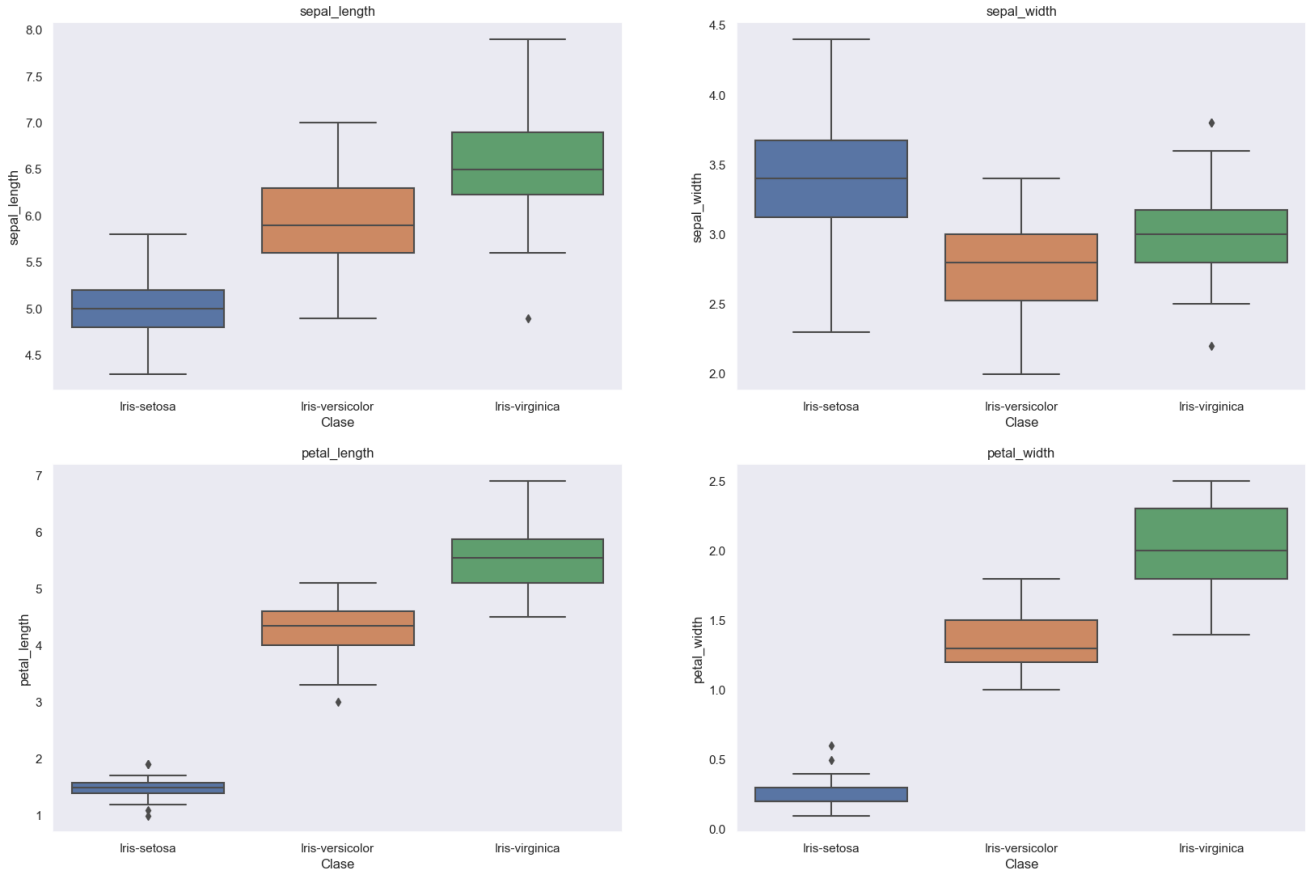
Primero checamos que los valores se están extrayendo de forma correcta:

	sepal_length	sepal_width	petal_length	petal_width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa

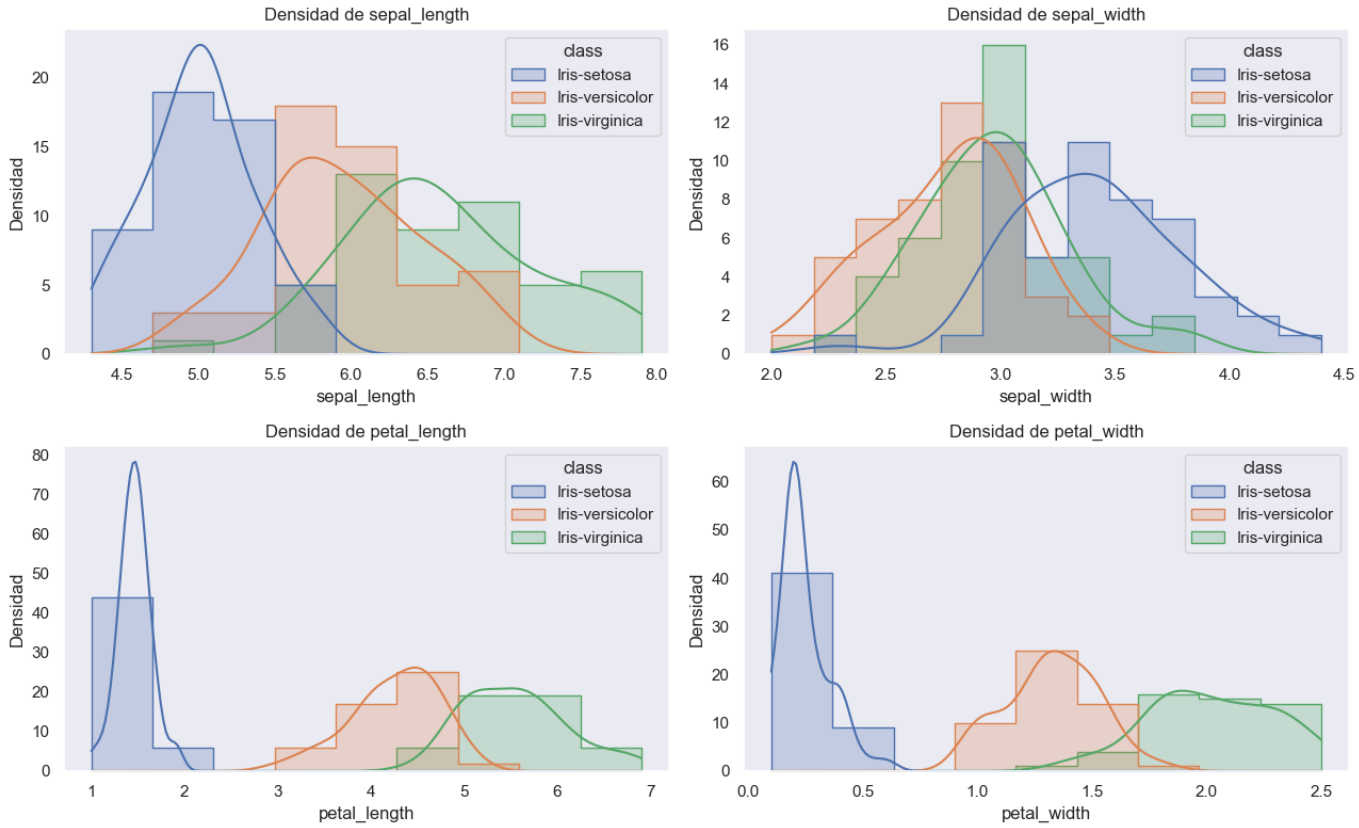
Para realizar un análisis mas profundo, podemos revisar los valores agrupados para tener un panorama mas concreto de los datos

```
Estadísticas para sepal_length:
                 mean       var       std
class
Iris-setosa      5.006  0.124249  0.352490
Iris-versicolor  5.936  0.266433  0.516171
Iris-virginica   6.588  0.404343  0.635880

Estadísticas para sepal_width:
                 mean       var       std
class
Iris-setosa      3.418  0.145180  0.381024
Iris-versicolor  2.770  0.098469  0.313798
Iris-virginica   2.974  0.104004  0.322497

Estadísticas para petal_length:
                 mean       var       std
class
Iris-setosa      1.464  0.030106  0.173511
Iris-versicolor  4.260  0.220816  0.469911
Iris-virginica   5.552  0.304588  0.551895

Estadísticas para petal_width:
                 mean       var       std
class
Iris-setosa      0.244  0.011494  0.107210
Iris-versicolor  1.326  0.039106  0.197753
Iris-virginica   2.026  0.075433  0.274650
```

Para el análisis de los datos se opto por usar diagramas de cajas para conocer la dispersión de los datos, así como los valores atípicos que tiene



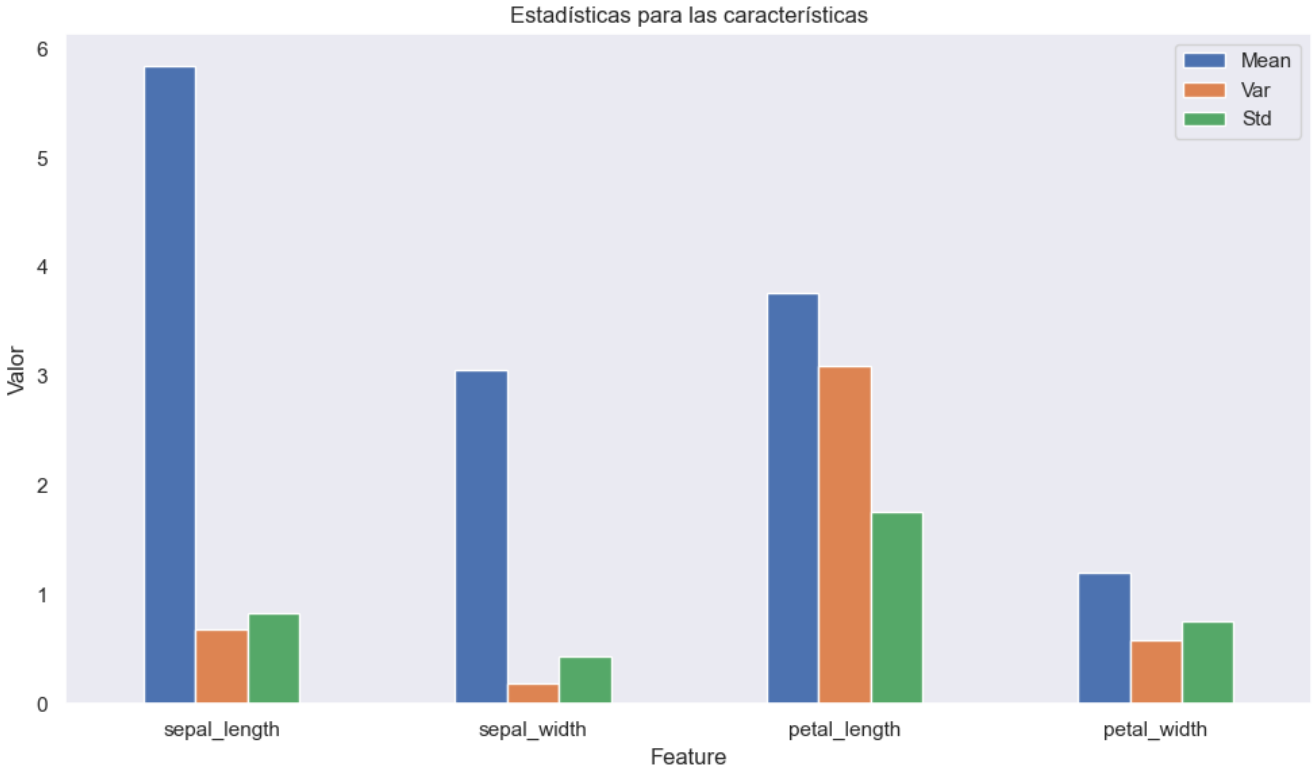
A su vez analizamos la densidad con un histograma de cada rasgo del iris



De la parte del análisis mas global de los rasgos, obtenemos una tabla de la media, varianza y desviación estándar

Feature	Mean	Var	Std
sepal_length	5.843333	0.681122	0.825301
sepal_width	3.054000	0.186751	0.432147
petal_length	3.758667	3.092425	1.758529
petal_width	1.198667	0.578532	0.760613

La cual se logra analizar mejor con un gráfico de barras:



Adentrándonos en la correlación de Pearson, obtenemos los siguientes valores:

```
Coeficiente de correlación de Pearson entre sepal_length y sepal_width es: -0.11
Hay una correlación negativa: A medida que sepal_length aumenta, sepal_width disminuye

Coeficiente de correlación de Pearson entre sepal_length y petal_length es: 0.87
Hay una correlación positiva: A medida que sepal_length aumenta, petal_length aumenta.

Coeficiente de correlación de Pearson entre sepal_length y petal_width es: 0.82
Hay una correlación positiva: A medida que sepal_length aumenta, petal_width aumenta.

Coeficiente de correlación de Pearson entre sepal_width y petal_length es: -0.42
Hay una correlación negativa: A medida que sepal_width aumenta, petal_length disminuye

Coeficiente de correlación de Pearson entre sepal_width y petal_width es: -0.36
Hay una correlación negativa: A medida que sepal_width aumenta, petal_width disminuye

Coeficiente de correlación de Pearson entre petal_length y petal_width es: 0.96
Hay una correlación positiva: A medida que petal_length aumenta, petal_width aumenta.
```

Por otra parte, obtenemos los siguientes resultados con el algoritmo anova

```
ANALISIS DE ANOVA PARA: sepal_length

Estadístico F: 119.26450218450438
Valor p: 1.1102230246251565e-16
Rechazamos la hipótesis nula: Hay diferencias significativas entre los grupos.

ANALISIS DE ANOVA PARA: sepal_width

Estadístico F: 47.36446140299379
Valor p: 1.1102230246251565e-16
Rechazamos la hipótesis nula: Hay diferencias significativas entre los grupos.

ANALISIS DE ANOVA PARA: petal_length

Estadístico F: 1179.0343277002205
Valor p: 1.1102230246251565e-16
Rechazamos la hipótesis nula: Hay diferencias significativas entre los grupos.

ANALISIS DE ANOVA PARA: petal_width

Estadístico F: 959.32440572576
Valor p: 1.1102230246251565e-16
Rechazamos la hipótesis nula: Hay diferencias significativas entre los grupos.
```

E. Análisis de Resultados

Gráfico de cajas:

- Analizando el grafico de cajas, en la primera sección de la longitud del sépalo, observamos que tenemos que para la clase setosa es mas pequeña a diferencia de las demás, donde se tiene un amplio rango de valores que van desde 4 hasta 8, aproximad ante, solo tenemos un valor atípico negativo en virginica, pero al juntar todos los valores no afecta de manera considerable
- A deferencia de la longitud del sépalo, en la anchura principalmente tenemos un amplio rango en setosa, teniendo el promedio mas alto a diferencia de los demás, el rango se encuentra de 2 a 4.5 aproximadamente, teniendo dos valores atípicos de virginica, sin embargo, estos valores no son muy considerables.
- Entrando en el análisis del pétalo, en la longitud vemos una tendencia de datos más concentrados, es decir a diferencia del sépalo, los promedios de los valores contiene mayor cantidad de datos, además de que los vigores de las cajas no son tan amplios como en el anterior, en este caso tenemos un rango de 1 a 7 aproximadamente, sin embargo, tenemos una gran diferencia entre setosa y virginica, siendo la más dominante esta segunda
- Por último, tenemos la anchura del pétalo, el cual, de manera similar a la longitud, existen una clara concentración de datos, además de repetirse la enorme diferencia entre setosa y virginica. En todos los casos la clase versicolor se encuentra en la mitad aproximadamente, en este caso los bigotes en general se aprecian más cortos a diferencia del sépalo

Histograma de densidad:

- Comenzando del mismo modo con la longitud de sépalo, podemos ver que se tiene una densidad marcada en setosa, a pesar de tener los valores mas chicos, se tiene una concentración de estos entre 4.5 y 5.5, pero la longitud de este es más limitada, a diferencia de virginica que tienen valores más altos, pero se encuentran mas dispersos, de forma similar pasa con versicolor
- En el análisis de la anchura del sépalo, observamos que en todas las clases abarca una gran cantidad de datos a lo largo del rango, sin embargo, tenemos valores altos en virginica a diferencia de las damas que son mas similares

- En la longitud del pétalo, afirmamos lo que se observo con las graficas de cajas, donde observamos una dispersión de datos mayor, es decir las clases están muy separadas entre sí, podemos ver como setosa se separa de versicolor y virginica, las cuales son las que mas datos relacionados tienen, en este caso vemos una mayor concentración de datos en un punto especifico y no se observa que abarquen mucho
- Para la anchura del pétalo, del mimo modo que el anterior, tenemos mayores concentraciones de datos en puntos específicos y no tanta dispersión de estos, se observa nuevamente una separación entre setosa y versicolor con virginica, sin embargo, setosa tiene una gran concentración de datos, a diferencia del otro grupo que están mas dispersos a lo largo del rango

Grafico de barras:

- En la media, podemos ver que entre los rasgos existen diferencias grandes, por ejemplo, entre la longitud del sépalo y la anchura de sépalo, es grande la separación, sin embargo, entre la anchura del sépalo y la longitud del pétalo es más cercana
- En la varianza, en la mayoría de los casos tenemos valores pequeños, sin embargo, en la longitud del pétalo si existe un rango de los resultados
- Del mismo mod, al estar relacionada la varianza con la desviación estándar, la longitud del pétalo es elevada comparada con las demás

Correlación de Pearson

Recordemos que la correlación de Pearson nos indican la fuerza y la dirección de la relación lineal entre las diferentes variables.

1. Coeficiente de correlación de Pearson entre sepal_length y sepal_width (-0.11):
    - Hay una correlación negativa muy débil entre la longitud del sépalo (sepal_length) y el ancho del sépalo (sepal_width).
    - Esto sugiere que, en general, a medida que la longitud del sépalo aumenta, el ancho del sépalo tiende a disminuir, pero la relación es muy débil y apenas significativa.

2. Coeficiente de correlación de Pearson entre sepal_length y petal_length (0.87):
    - Hay una fuerte correlación positiva entre la longitud del sépalo (sepal_length) y la longitud del pétalo (petal_length).
    - Esto indica que, en general, a medida que la longitud del sépalo aumenta, la longitud del pétalo tiende a aumentar de manera significativa.

3. Coeficiente de correlación de Pearson entre sepal_length y petal_width (0.82):
    - Existe una correlación positiva fuerte entre la longitud del sépalo (sepal_length) y el ancho del pétalo (petal_width).
    - Esto implica que, en general, a medida que la longitud del sépalo aumenta, el ancho del pétalo tiende a aumentar de manera significativa.

4. Coeficiente de correlación de Pearson entre sepal_width y petal_length (-0.42):
    - Hay una correlación negativa moderada entre el ancho del sépalo (sepal_width) y la longitud del pétalo (petal_length).
    - Esto sugiere que, en general, a medida que el ancho del sépalo aumenta, la longitud del pétalo tiende a disminuir, aunque la relación no es muy fuerte.

5. Coeficiente de correlación de Pearson entre sepal_width y petal_width (-0.36):
    - Existe una correlación negativa moderada entre el ancho del sépalo (sepal_width) y el ancho del pétalo (petal_width).

- Esto implica que, en general, a medida que el ancho del sépalo aumenta, el ancho del pétalo tiende a disminuir, pero la relación no es muy fuerte.

6. Coeficiente de correlación de Pearson entre petal_length y petal_width (0.96):
    - Hay una correlación positiva muy fuerte entre la longitud del pétalo (petal_length) y el ancho del pétalo (petal_width).
    - Esto indica que, en general, a medida que la longitud del pétalo aumenta, el ancho del pétalo tiende a aumentar de manera significativa y altamente correlacionada.

Los resultados de la correlación de Pearson sugieren que la longitud y el ancho del sépalo están débilmente relacionados entre sí, mientras que tanto la longitud del sépalo como el ancho del sépalo están fuertemente relacionados con la longitud y el ancho del pétalo. Además, la relación entre la longitud y el ancho del pétalo es extremadamente fuerte.

Análisis de ANOVA

1. Para sepal_length:
    - El estadístico F es 119.26450218450438.
    - El valor p es 1.1102230246251565e-16, lo que es esencialmente cero.
    - Los resultados indican que hay diferencias significativas entre los grupos de datos en función de esta característica. En otras palabras, las longitudes de sépalo varían de manera significativa entre las diferentes categorías o especies de iris presentes en el conjunto de datos.

2. Para sepal_width:
    - El estadístico F es 47.36446140299379.
    - El valor p es 1.1102230246251565e-16, que es esencialmente cero.
    - De manera similar al caso anterior, los resultados sugieren que existen diferencias significativas en el ancho del sépalo entre las diferentes categorías de iris. Esto significa que el ancho del sépalo varía considerablemente según la especie de iris.

3. Para petal_length:
    - El estadístico F es 1179.0343277002205.
    - El valor p es 1.1102230246251565e-16, que es esencialmente cero.
    - Los resultados muestran que hay diferencias altamente significativas en la longitud del pétalo entre las diferentes categorías de iris. Esto indica que la longitud del pétalo es una característica que varía sustancialmente según la especie de iris.

4. Para petal_width:
    - El estadístico F es 959.32440572576.
    - El valor p es 1.1102230246251565e-16, que es esencialmente cero.
    - Al igual que las otras características, los resultados nos revelan que existen diferencias altamente significativas en el ancho del pétalo entre las diferentes categorías de iris. Esto significa que el ancho del pétalo varía de manera importante según la especie de iris.

Con los resultados obtenidos , en todos los casos, el valor p extremadamente bajo (prácticamente cero) indica que rechazamos la hipótesis nula, lo que significa que hay diferencias significativas entre los grupos (especies de iris) en función de las diferentes características analizadas. Esto sugiere que las características de longitud y ancho de sépalo y pétalo son buenas para distinguir entre las diferentes especies de iris en este conjunto de datos.

F. Conclusiones

Después de analizar los resultados de los coeficientes de correlación de Pearson y los análisis de ANOVA en el conjunto de datos de iris, podemos llegar a las siguientes conclusiones:

1. Coeficientes de correlación de Pearson:
    - Los coeficientes de correlación de Pearson nos proporcionaron información sobre la relación lineal entre pares de variables en el conjunto de datos.
    - Encontramos relaciones significativas entre algunas características, como una fuerte correlación positiva entre sepal_length y petal_length (0.87) y sepal_length y petal_width (0.82), así como una fuerte correlación positiva entre petal_length y petal_width (0.96).
    - Estos resultados nos indican que estas características están relacionadas de manera lineal y pueden influirse mutuamente.

2. Análisis de ANOVA:
    - Los análisis de ANOVA se centraron en determinar si había diferencias significativas entre los grupos de datos en función de cada característica (sepal_length, sepal_width, petal_length y petal_width).
    - Todos los análisis de ANOVA arrojaron valores p muy bajos (prácticamente cero), lo que indica que hay diferencias altamente significativas entre los grupos (especies de iris) en función de cada característica.
    - Esto significa que las características de longitud y ancho de sépalo y pétalo son muy efectivas para distinguir entre las diferentes especies de iris en este conjunto de datos.

3. Comparación y Conclusiones Generales:
    - Ambos enfoques estadísticos proporcionaron resultados consistentes y respaldaron la idea de que las características de sépalo y pétalo son importantes para diferenciar las especies de iris.
    - Los coeficientes de correlación de Pearson cuantificaron la fuerza y la dirección de las relaciones lineales entre las variables, mientras que el ANOVA se centró en identificar las diferencias entre grupos.
    - En resumen, los coeficientes de correlación de Pearson ayudaron a entender cómo las características están relacionadas entre sí, mientras que el ANOVA confirmó que estas diferencias eran estadísticamente significativas entre las especies de iris.
    - Estos análisis conjuntos permiten concluir que las características de sépalo y pétalo son fundamentales en la clasificación y diferenciación de las especies de iris en este conjunto de datos.

G. Referencias bibliográficas

[1] https://revistas.pucsp.br/emp/article/view/37072
[2] https://cienciadedatos.net/documentos/9_homogeneidad_de_varianza_homocedasticidad
[3] https://derek-corcoran-barrios.github.io/AyduantiaStats/_book/Explorando.html
[4] http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP/MATERIALESMASTER/Mat_14_Iris%20data.pdf
[5] https://jcoliver.github.io/learn-r/002-intro-stats.html
[6] https://felipebravom.com/teaching/explora.pdf
[7] https://rpubs.com/rajesh1/563805
[8] https://rpubs.com/Karolina_G/848706