**MACHINE LEARNING FROM DATA**

**Report: Lab Session 0 – Exploratory data analysis**

**Names: Abdullah Qureshi, S M Rakib Hasan and Niina Hietamäki**

**Group:**

**Instructions**

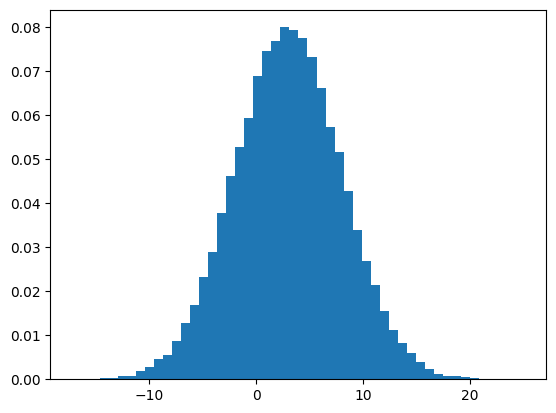
* Answer the questions in a document **Lab0\_report\_team\_surnames.docx**
* Write the new code in a Colab Notebook **Mlearn\_lab0\_3\_team\_surnames.ipynb**.

**Questions**

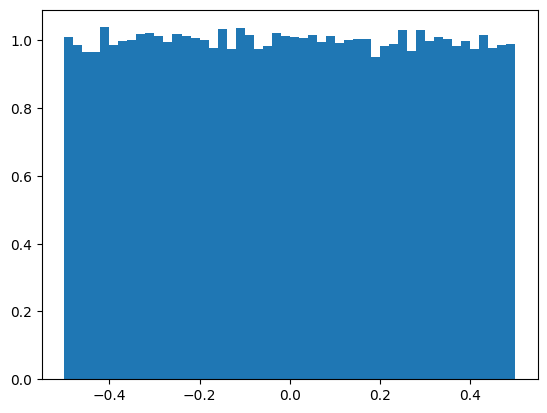
Q1. Briefly describe the conclusions of your analysis (you can insert plots)

Answer: The distributions show different properties in reference to their skewness, kurtosis quantile values and CDF.

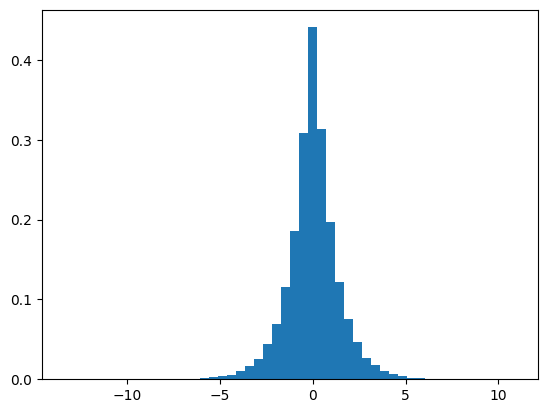
1. Normal Distribution: The shape of the histogram resembles a bell curve and most of the data is centred around 3 and the frequency decreases as we move away from the mean toward the tail. The kurtosis is 3 meaning the points are slightly distributed towards the tail. The skewness is 0. So the data points are symmetrical on both sides of the centre point. Quantile-Quantile plot against itself forms a straight diagonal line and CDF is an S-shaped curve.



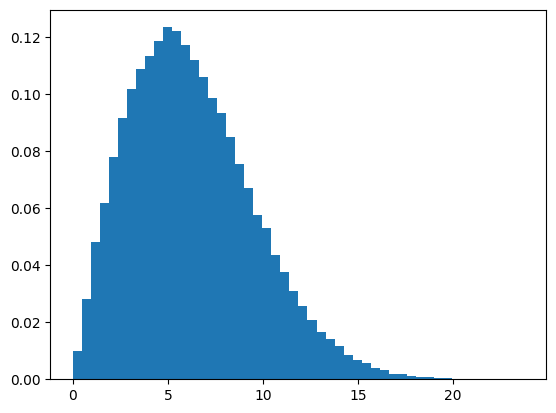
1. Uniform distribution: The shape of histogram for this distribution is flat and evenly spread which means all values are equally likely. The kurtosis value is -1 with respect to the normal distribution which means it has a flat shape with no tails. The skewness is 0 so it is symmetric about the mean. The Quantile-Quantile plot shows a curved pattern



1. Laplacian distribution: It has sharper peaks and fatter tails. The data points are more concentrated near the mean. It has a kurtosis of +3 compared to normal distribution means that it has a greater probability of producing outliers. It has a skewness of 0 indicating that it is symmetric just like the normal distribution



1. Rayleigh distribution The peak occurs to the left of the distribution and it gradually declines towards the right, with a tail extending to larger values. Most of the data is concentrated near the lower end of the range. In this example we have a positive skewness of around 0.6 indicating it is right-skewed.



|  | Normal | Uniform | Laplace | Rayleigh |
| --- | --- | --- | --- | --- |
| Shape | peaked at center | Flat, rectangular | Sharp peak at the center | Not uniform |
| Skewness | symmetric | symmetric | symmetric | Right-skewed |
| Kurtosis | Thin tails | No tails | greater probability of producing outliers | Almost same as normal distribution |
| Data | Most values around the mean | Uniformly spread | Most values near the mean but with more outliers | Concentrated near one peak |

Q2. For each class and each feature, analyze histograms, cdfs and normal plots. Can we assume a Gaussian distribution for any of the features?

Based on the analysis of histograms, cdfs and normal plots,

* For Iris Setosa, sepal length, sepal width are roughly normally distributed but petal width and petal length are not.
* For Iris Versicolor, the sepal length, sepal width, and petal length follow a normal distribution while petal width shows a bit of skew
* For Iris Virginica, sepal length, sepal width and petal length are approximately normal but petal width being slightly skewed

|  | Iris-setosa | Iris-Versicolor | Iris-virginica |
| --- | --- | --- | --- |
| sepal\_length | Gaussian | Gaussian | Gaussian |
| sepal\_width | Gaussian | Gaussian | Gaussian |
| petal\_length | Not | Gaussian | Gaussian |
| petal\_width | Not | Not | Not |

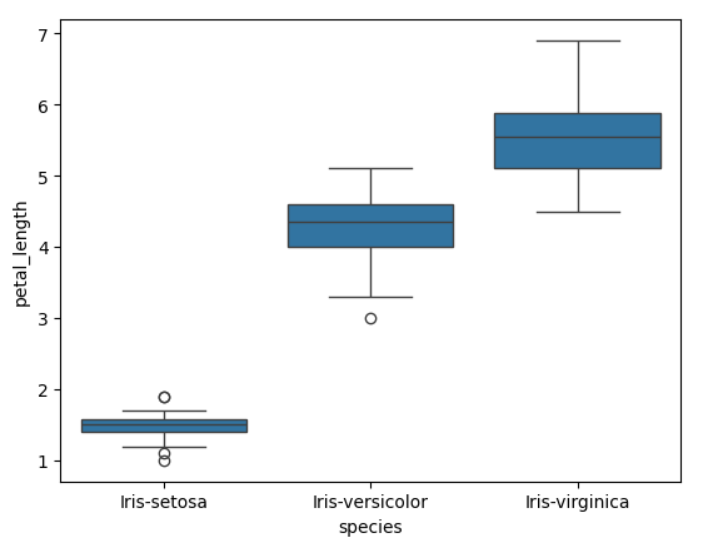
Q3. Analyze kurtosis and skewness values for each feature and class.

|  | Iris-setosa | Iris-Versicolor | Iris-virginica |
| --- | --- | --- | --- |
| sepal\_length | Skewness: 0.12, Kurtosis: -0.25 - Slight right skew, close to normal | Skewness: 0.11, Kurtosis: -0.53 - Slight right skew, light tails | Skewness: 0.12, Kurtosis: 0.03 - Slight right skew, close to norma |
| sepal\_width | Skewness: 0.11, Kurtosis: 0.89 - Slight right skew, slightly heavy tails | Skewness: -0.36, Kurtosis: -0.37 - Slight left skew, light tails | Skewness: 0.37, Kurtosis: 0.71 - Slight right skew, slightly heavy tails |
| petal\_length | Skewness: 0.07, Kurtosis: 1.03 - Nearly symmetric, heavy tails | Skewness: -0.61, Kurtosis: 0.05 - Moderate left skew, close to normal tails | Skewness: 0.55, Kurtosis: -0.15 - Slight right skew, close to normal |
| petal\_width | Skewness: 1.20, Kurtosis: 1.57 - Strong right skew, heavy tails (non-Gaussian) | Skewness: -0.03, Kurtosis: -0.41 - Almost symmetric, light tails   |  | | --- | | Skewness: -0.13, Kurtosis: -0.60 - Nearly symmetric, light tails |

Q4. Analyze boxplots by feature. Are there ‘significant’ differences between the classes?

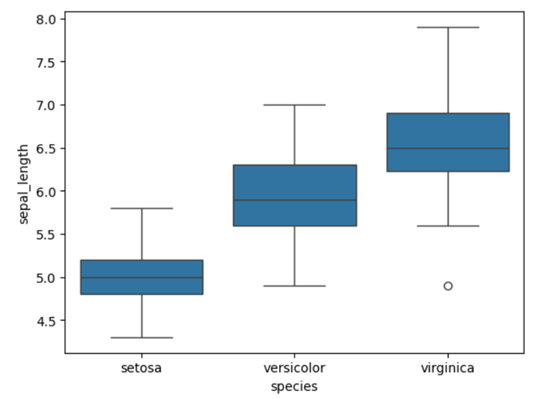
Petal Length:

The petal length for Setosa is significantly smaller, ranging between 1 and 2, with a few outliers, indicating consistently shorter petals. Versicolor’s petal length is more spread out, ranging from 3 to 5, showing moderate variation. Virginica exhibits the largest petal length, ranging from 4.5 to 7, with a wider spread than the other species, indicating a greater variety in petal length.



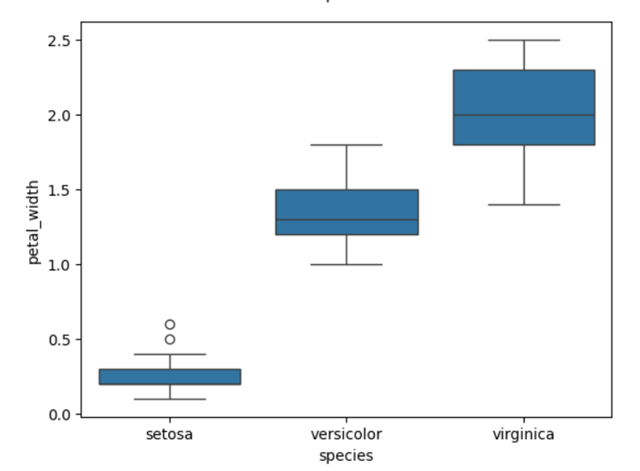
Sepal Length:

The sepal length for Setosa ranges between 4.5 and 5.5, and the values are tightly packed. This indicates that Setosa has consistently shorter sepal lengths. The sepal length for Versicolor is more spread out, ranging from 5.5. to 7, showing moderate variation. Virginica has the largest sepal length, ranging from 6 to 8, with one outlier. Virginica's sepal lengths are spread out more than Setosa’s and Versicolor’s.



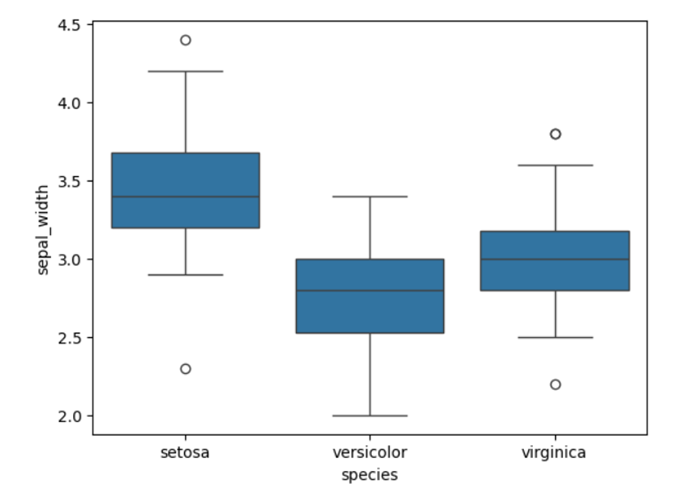
Petal Width:

The petal width for Setosa is quite small with values clustered around 0.2 to 0.4 and there are a few outliers. This shows that Setosa has consistently narrow petals. Versicolor's petal width is wider, ranging from 1 to 1.8. The variation is moderate.Virginica shows the widest petal width, ranging from 1.8 to 2.5. The data is spread out more compared to the other species.



Sepal width

The sepal width for Setosa is the largest, with values ranging from 3.4 to 4.2, though there are some outliers. This suggests that Setosa typically has wider sepals compared to other species. Versicolor’s sepal width is narrower, ranging from 2.7 to 3.4, with moderate variation. Virginica's sepal width is similar to Versicolor’s, ranging from 2.5 to 3.5, with some outliers, indicating a slightly more spread-out distribution.



Conclusion:

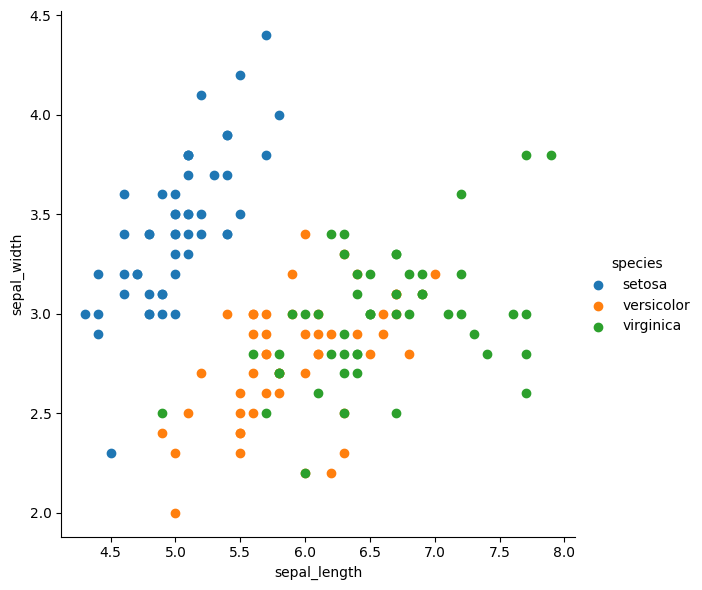
Overall, these differences highlight clear separations between the species, with Setosa being the most distinct, and Versicolor and Virginica showing some overlap in certain features.

Q5. Analyze the scatter plot. Are features related in any way? What can you say about the separability of the classes?

In this scatter plot, the relationship between sepal length (x-axis) and sepal width (y -axis) is visualised. From the plot with color-coding for the different species: Setosa (blue), Versicolor (orange), and Virginica (green). Sepal length varies from around 4.5 to 8, while sepal width ranges from 2.0 to 4.5.

The scatter plot shows that as sepal length increases, sepal width does not increase in a predictable way. This suggests that these two features are not strongly correlated. However, with the color-coding, we can now clearly see the separability of the species. Setosa (blue points) forms a distinct cluster with shorter sepal lengths and wider sepal widths, making it easily separable from the other species. Versicolor (orange points) and Virginica (green points), however, overlap in several areas, particularly around sepal lengths of 5.5 to 7 and sepal widths of 2.5 to 3.5, making them harder to distinguish based solely on these two features.

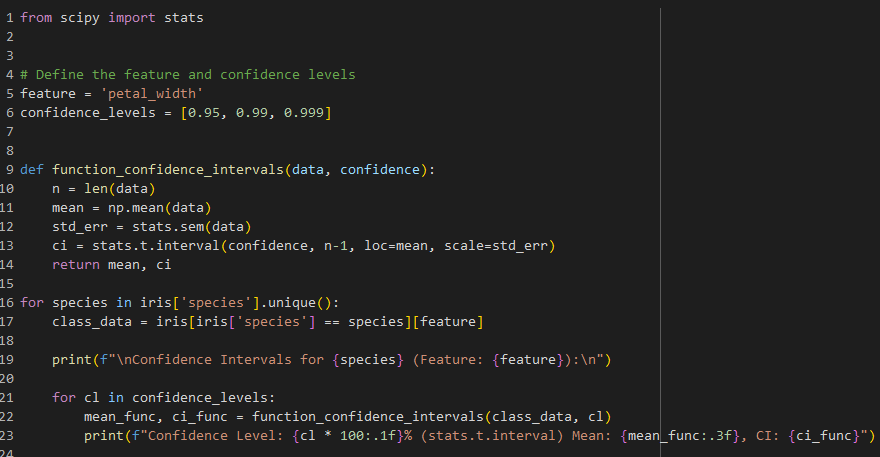
The spread in sepal width is greater for shorter sepal lengths, particularly for Setosa. For larger sepal lengths, especially for Virginica, the sepal width is more consistent, showing less variation.



Create a new Colab Notebook: **Mlearn\_lab0\_3\_team\_surnames.ipynb**.

Q6. Choose one feature (among the four available), write the code to compute the feature mean and confidence intervals at confidence levels 95%, 99% and 99.9% for the three classes.

Link to the Colab Notebook: [Here](https://colab.research.google.com/drive/1Vl7MfX9z9-8up-zwnS8jjUBSxdvKVzAH?usp=sharing)



Feature selected: **Petal Width**

|  | Mean | CI at 95% | CI at 99% | CI at 99,9% |
| --- | --- | --- | --- | --- |
| Class 1 | 0.244 | (0.21353139634492724, 0.27446860365507275) | (0.20336733644476213, 0.28463266355523786) | (0.19092728836237272, 0.29707271163762783) |
| Class 2 | 1.326 | (1.2697993101768552, 1.3822006898231445) | (1.2510512512155796, 1.40094874878442) | (1.228105031179435,1.4238949688205655) |
| Class 3 | 2.026 | (1.9479453178263322, 2.1040546821736683) | (1.9219069580090313, 2.1300930419909694) | (1.890037955730865,2.1619620442691367) |

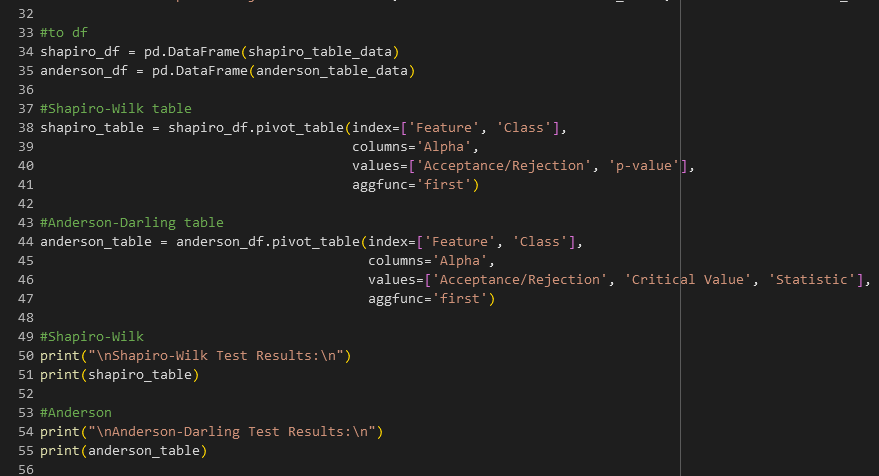
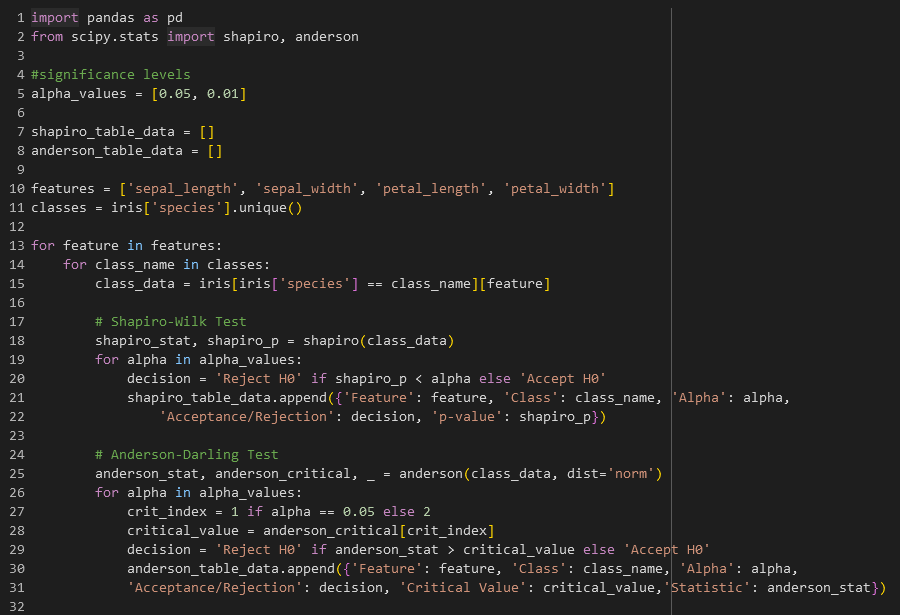
Q7. Write the code to conduct the following hypothesis tests, using the Shapiro-Wilk test and the Anderson Darling test, for **all** the features K and classes J.

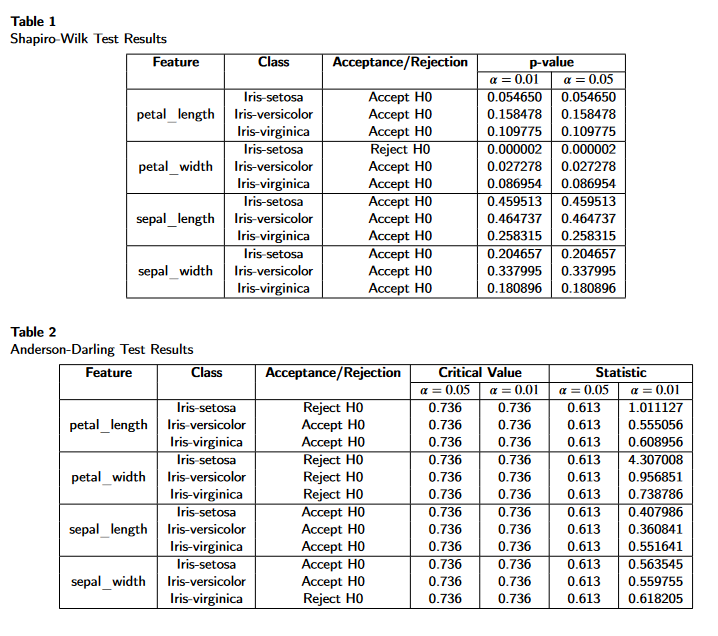
* Null hypothesis : Feature K from class J comes from a Gaussian distribution at the significance level *α*



For each test complete the corresponding table with the decisions (acceptance/rejection) for the null hypothesis H0 (feature Gaussianity), and the p-value or the critical and statistic values, respectively, for *α* = 0,05 and *α* = 0,01

Explain the meaning of the p-value / critical value and interpret the results accordingly.





**Shapiro-Wilk Test**:

* If the p-value is less than the specified alpha (0.05 or 0.01), we reject the null hypothesis, indicating the feature data does not follow a normal distribution.

**Anderson-Darling Test**:

* If the test statistic exceeds the critical value for a given alpha, we reject the null hypothesis, indicating the feature data does not follow a normal distribution.