**MACHINE LEARNING FROM DATA**

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

**Names: Miquel Gómez I Esteve / Ricard Abril Ferreres**

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

A group of blue lines

Description automatically generated

The previous plots, which are histograms, provide information about how the data is distributed. It provides the amount of each of the random variables being in each of the bins. We can clearly see 4 kinds of distribution: Normal (Gauisian), Uniform, Laplace and Rayleigh.

A blue graph with numbers

Description automatically generated

We can also analyse the probability density function (shown above for a Reilegh distribution), which is used because we are working with continuous random variables. A specific point does not represent the probability of a random variable to be exactly at that point, but instead it gives us the likelihood of being in that point related to the rest of the points in the curve. The area under the curve must always be 1.

A group of blue graphs

Description automatically generated

Above we can see the histograms for the four distributions we have been talking about. Each plot shows the number of values found in each of the bins. The plot also shows a line, which is a histogram smoothing, which produces a smooth curve that tries to represent the probability density of the data.

A group of graphs showing different types of distribution

Description automatically generated

Regarding the cumulative distribution (shown above) we can see the probability of a random number being equal or less than a defined threshold. The blue line shows us the cdf for our original data. We can compare it to the red line, which shows us what a theoretical cdf would be for the four distributions. In all cases we see that the data mostly follows the red line, however not exactly. So, in general, we can assume that our data follows the defined distributions, even if it is not perfect.

A screenshot of a graph

Description automatically generated

The skewness gives us information about how asymmetric a distribution is, based on the mean. The lower the skewness (using its absolute value) the more symmetric the distribution will be around the mean value and the mean and the median will be more similar.

If it is positive, the mean will be higher than median, and there will be a higher density of values on the left side. If it is negative, the mean will be lower than the median and there will be a higher density of values on the right side. This Rayleigh distribution is a clear example of a right skewed (positive) distribution.

On the other hand, kurtosis is used to analyse the tailedness of how prone a distribution is to have outliers.

* In the case of our normal distribution, the kurtosis is 0, which means that it follows a normal bell curve.
* In the case of the uniform distribution, the kurtosis is less than 0, which means that It has less tails and is flatter than the normal.
* Regarding the laplace the kurtosis is very positive. Compared to the normal distribution, the laplace has a taller peak and longer tails.
* Finally, regarding Rayleigh, the kurtosis is very similar to the normal. With that in mind, this distribution has a longer tail, compared to the ones in the normal distribution.

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

A group of blue and white graphs

Description automatically generatedA screenshot of a graph

Description automatically generated

A chart of different values

Description automatically generated with medium confidence

To see if any feature follows a normal gaussian distribution, we must check the histograms, the cumulative distribution function as well as the quantile-quantile plot shown above for each feature-class combination.

It is important to note that the amount of data is very small, so it is hard to get conclusions from it.

Analysing the histograms:

* The only histograms that mildly resemble a normal distribution are (iris-setosa, petal-length) and are (iris-setosa, sepal-length).
* The other histograms show a clear skewness or an irregular distribution in general.

Analysing the cdf plots:

* The combinations (iris-setosa, petal-width) shows a clear deviation compared to the other plots.
* The features that are closer to a normal regarding the cdf are (iris-setosa, petal-length), (iris-versicolor, petal-length) and (iris-versicolor, petal-width).

Analysing the quantile-quantile plots:

* (iris-setosa, petal-width) shows a clear deviation from the normal, but that could also be caused by the discretization of the data (nbin).
* (iris-setosa, petal-length) and (iris-versicolor, petal-length) are the plots that could be the most like the normal distribution.

In conclusion, after the analisys, we have three features which are candidates to have a normal distribution. These are (iris-setosa, sepal-length), (iris-setosa, petal-length) and (iris-versiocolor, sepal-width). However, ass mentioned above, with the dataset being so small we cannot assume a normal distribution.  
With the rest of the pairs (class, feature) we could assume that they do not follow a normal distribution.

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

A screenshot of a graph

Description automatically generated

Regarding Kurtosis:

* The kurtosis for the sepal length negative with iris setosa and versicolor, that means that the tails will be smaller, and they will have less outliers, when compared to the same sepal length with the class iris virginica. The sepal length for iris virginica is almost 0 so regarding tailedness, it is very similar to a normal distribution.
* The sepal width kurtosis is very big in both iris setosa and iris virginica, that means that the distribution has bigger tails and more outliers than the iris versicolor compared to a default normal distribution. The sepal width for the iris versicolor has less outliers and smaller tails in general when compared to a normal distribution, as the value is negative.
* The petal length kurtosis is very big for iris setosa that means that it has bigger and longer tails than the other classes and the normal distribution. The petal length for the iris versicolor class is very close to 0 so regarding tailednees, it will be very similar to a normal distribution. The petal length for the iris virginica is less than 0 so it will have less outliers and smaller tails than the normal distribution.
* Finally, regarding the petail width, for iris virginica and iris versicolor, the values for kurtosis are lower than zero, so when compared to a normal distribution, they will have less outliers and smaller tails. For the iris setosa class kurtosis is the biggest from all the features and classes so I will have more outliers and longer tails than any other class-feature pair.

A screenshot of a graph

Description automatically generated

Regarding Skewnees:

* For sepal length, all the values are very similar, and a bit bigger than 0, that means that the data is slightly moved to the left compared to a normal.
* Regarding sepal width for iris setosa, the data is a bit moved to the left, for iris virginica as well, but it is a bit more asimetric. For iris versicolor, the value is negative, so it is less symmetric and is moved to the right.
* For petal length and iris setosa, the value is almost 0, so it is very symmetric. For iris versicolor and iris virginica, the distribution is not that symmetric. For iris versicolor the data is more concentrated to the right, and the opposite applies to iris virginica.
* Finally, for petal width, the iris setoda distribution is very asymmetric and is hightly concentrated to the left. For iris versicolor the distribution is very close to the normal regarding symmetry because the skewness is very close to 0. For iris virginica the data is slightly concentrated in the right, but still is very close to the normal.

In conclusion:

* By checking the Kurtosis, we could assume that iris virginica-sepal length and iris versicolor-petal length would be very good candidates to follow a gauissian distribution
* By checking the Skewness, we could assume that iris setosa-petal length and iris versicolor-petal width would be very good candidates to follow a gauissian distribution.
* Knowing that most of the combinations did not appear in the other questions as candidates or that either the skewness value or the kurtosis value is not close to the normal. We cannot assume that any pair of class-feature follows a normal distribution.

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

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

In this box plot we can see that the data dispersion for iris setosa is smallest than for the other classes, the median, is almost at the midel of the box so we can say that the data is almost simetric at the intercuadratic segment.

We can see that the whiskers are really short iris setosa, so the data dispersion there is small, but we have the bigger number of outlayer too, for the rest the classes we can see that the whiskers longer so we can say that we have more data dispersion outside the intercuadratic range. We can see that longitude of the whiskers are different so we can say that the data is biased.

Finally we can conclude that te iris setosa have the minimum deviation of the 3 classes.

This data it would be useful to differentiate one class with the otter, because we dont have overlapping between the q-plots.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

In this case we can see that for iris-setosa, the distribution is biased to the left, we have some outliers at the positive whisker, so the distribution is not symmetric. Inside of the intercuadratic segment the data have less dispersion than outside the intercuadratic range.

For Iris versicolors the data is biased to the left, we can see this taking a look to the median, the positive whisker is bigger than the negative one so at the end this data is going to be more symmetric than for the rest of the cases. In the intercuadratic range we can see that we have more data dispersion.

For iris virginica, we have even more data dispersion inside the intercuadratic range, the negative whisker is bigger indicating that the distribution is biased to the left, also the median is indicating us the same so the dispersion is not symmetric.

This data would be useful to differentiate one class with the otter, because we don't have overlapping between the q-plots.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

For this feature we can see that for the three classes the distribution is symmetric or almost symmetric. The longitude of the whiskers is pretty similar and the median is situated at the middle of the box. Iris virginica is the one with more deviation and more difference between the whiskers.

This feature is not the best to differentiate between the classes, because we have overlapping values.

Gráfico, Gráfico de cajas y bigotes

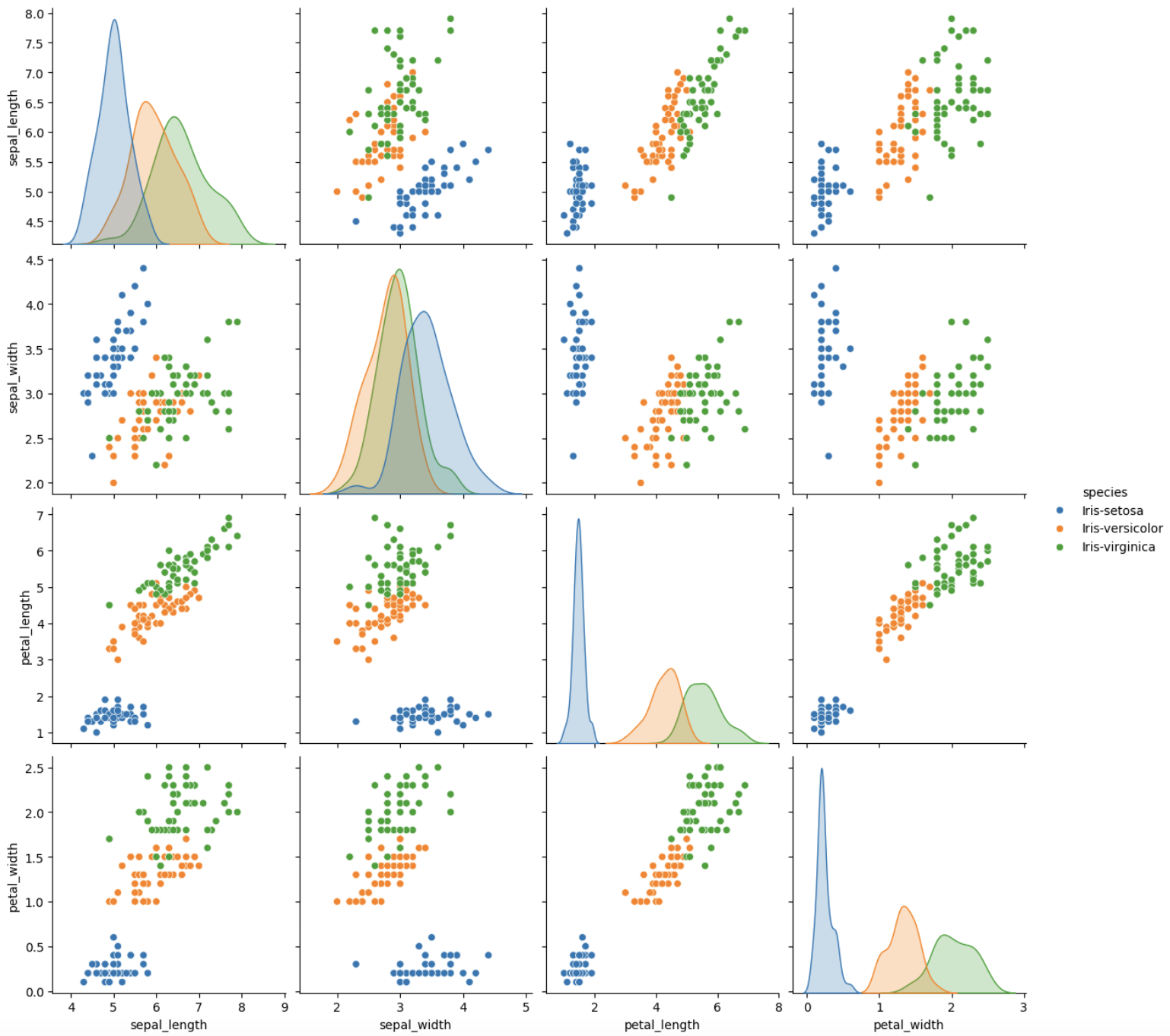
Descripción generada automáticamente

For irsi setosa we can see that the data is almost symmetric, and the median is at the middle of the box

For the rest of the classes we see that the data is biased to the right and we have outliers for iris virginica

Again the data is overlapping so its not the best to diferenciate the classes.

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



We can assume that petal length and petal width, independently of the class, are related. It means that the larger the length the larger the width as well. We can clearly see these features have a linear relationship. For the other pairs of features does not seem to be a relation with these plots.

Regarding separability of the classes, we clearly see that in all pairs of features iris setosa (blue) is clearly differentiated from the other two classes. The other two classes (iris versicolor and iris virginica) are more close to each other, however, they are not completely overlapping and we could classify data by finding a good decision boundary.

**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.**

The chosen feature is petal\_length. The computed confidence intervals are the following:

| **petal\_length** | **Mean** | **CI at 95%** | **CI at 99%** | **CI at 99,9%** |
| --- | --- | --- | --- | --- |
| **Iris-setosa** | 1.464 | (1.4147, 1.5133) | (1.3982, 1.5298) | (1.3781, 1.5499) |
| **Iris-versicolor** | 4.26 | (4.1265, 4.3935) | (4.0819, 4.4381) | (4.0274, 4.4926) |
| **Iris-virginica** | 5.5520 | (5.3952, 5.7088) | (5.3428, 5.7612) | (5.2788, 5.8252) |

**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 *H*0: Feature K from class J comes from a Gaussian distribution at the significance level a

For the Spahiro Wlik test the computed decisions for each feature, class and alphas are the following:

Test Feature Class Alpha Statistic P-value Decision

Shapiro-Wilk sepal\_length Iris-setosa 0.05 0.977699 0.459513 **Accept**

Shapiro-Wilk sepal\_length Iris-setosa 0.01 0.977699 0.459513 **Accept**

Shapiro-Wilk sepal\_length Iris-versicolor 0.05 0.977836 0.464737 **Accept**

Shapiro-Wilk sepal\_length Iris-versicolor 0.01 0.977836 0.464737 **Accept**

Shapiro-Wilk sepal\_length Iris-virginica 0.05 0.971179 0.258315 **Accept**

Shapiro-Wilk sepal\_length Iris-virginica 0.01 0.971179 0.258315 **Accept**

Shapiro-Wilk sepal\_width Iris-setosa 0.05 0.968692 0.204657 **Accept**

Shapiro-Wilk sepal\_width Iris-setosa 0.01 0.968692 0.204657 **Accept**

Shapiro-Wilk sepal\_width Iris-versicolor 0.05 0.974133 0.337995 **Accept**

Shapiro-Wilk sepal\_width Iris-versicolor 0.01 0.974133 0.337995 **Accept**

Shapiro-Wilk sepal\_width Iris-virginica 0.05 0.967391 0.180896 **Accept**

Shapiro-Wilk sepal\_width Iris-virginica 0.01 0.967391 0.180896 **Accept**

Shapiro-Wilk petal\_length Iris-setosa 0.05 0.954946 0.054650 **Accept**

Shapiro-Wilk petal\_length Iris-setosa 0.01 0.954946 0.054650 **Accept**

Shapiro-Wilk petal\_length Iris-versicolor 0.05 0.966004 0.158478 **Accept**

Shapiro-Wilk petal\_length Iris-versicolor 0.01 0.966004 0.158478 **Accept**

Shapiro-Wilk petal\_length Iris-virginica 0.05 0.962186 0.109775 **Accept**

Shapiro-Wilk petal\_length Iris-virginica 0.01 0.962186 0.109775 **Accept**

Shapiro-Wilk petal\_width Iris-setosa 0.05 0.813817 0.000002 **Reject**

Shapiro-Wilk petal\_width Iris-setosa 0.01 0.813817 0.000002 **Reject**

Shapiro-Wilk petal\_width Iris-versicolor 0.05 0.947626 0.027278 **Reject**

Shapiro-Wilk petal\_width Iris-versicolor 0.01 0.947626 0.027278 **Accept**

Shapiro-Wilk petal\_width Iris-virginica 0.05 0.959771 0.086954 **Accept**

Shapiro-Wilk petal\_width Iris-virginica 0.01 0.959771 0.086954 **Accept**

The p-value is a measure to analyze the strength of evidence against the null hypothesis. In the test itself, if the p-value is greater than alpha, we can accept the null hypothesis. Otherwise, we can reject it.

With that in mind, with the result we found, we can assume that the features sepal\_length, sepal\_width and petal\_length all follow a normal distribution across all classes of the sample.

The feature petal\_width follows a normal distribution for class Iris-virginica and iris-versicolor (at the 0.01 level), but it does not follow a normal distribution for Iris-setosa and Iris-versicolor at the 0.05 level.

For the Anderson-Darling the computed decisions for each of the feature, class and alphas are the following:

Test Feature Class Alpha Statistic Critical Value Decision

Anderson-Darling sepal\_length Iris-setosa 0.05 0.407986 0.538 **Accept**

Anderson-Darling sepal\_length Iris-setosa 0.01 0.407986 0.613 **Accept**

Anderson-Darling sepal\_length Iris-versicolor 0.05 0.360841 0.538 **Accept**

Anderson-Darling sepal\_length Iris-versicolor 0.01 0.360841 0.613 **Accept**

Anderson-Darling sepal\_length Iris-virginica 0.05 0.551641 0.538 **Reject**

Anderson-Darling sepal\_length Iris-virginica 0.01 0.551641 0.613 **Accept**

Anderson-Darling sepal\_width Iris-setosa 0.05 0.563545 0.538 **Reject**

Anderson-Darling sepal\_width Iris-setosa 0.01 0.563545 0.613 **Accept**

Anderson-Darling sepal\_width Iris-versicolor 0.05 0.559755 0.538 **Reject**

Anderson-Darling sepal\_width Iris-versicolor 0.01 0.559755 0.613 **Accept**

Anderson-Darling sepal\_width Iris-virginica 0.05 0.618205 0.538 **Reject**

Anderson-Darling sepal\_width Iris-virginica 0.01 0.618205 0.613 **Reject**

Anderson-Darling petal\_length Iris-setosa 0.05 1.011127 0.538 **Reject**

Anderson-Darling petal\_length Iris-setosa 0.01 1.011127 0.613 **Reject**

Anderson-Darling petal\_length Iris-versicolor 0.05 0.555056 0.538 **Reject**

Anderson-Darling petal\_length Iris-versicolor 0.01 0.555056 0.613 **Accept**

Anderson-Darling petal\_length Iris-virginica 0.05 0.608956 0.538 **Reject**

Anderson-Darling petal\_length Iris-virginica 0.01 0.608956 0.613 **Accept**

Anderson-Darling petal\_width Iris-setosa 0.05 4.307008 0.538 **Reject**

Anderson-Darling petal\_width Iris-setosa 0.01 4.307008 0.613 **Reject**

Anderson-Darling petal\_width Iris-versicolor 0.05 0.956851 0.538 **Reject**

Anderson-Darling petal\_width Iris-versicolor 0.01 0.956851 0.613 **Reject**

Anderson-Darling petal\_width Iris-virginica 0.05 0.738786 0.538 **Reject**

Anderson-Darling petal\_width Iris-virginica 0.01 0.738786 0.613 **Reject**

Regarding the Anderson-Darling test, the critical value is the threshold where we define that the sample follows a normal distribution or not. If the critical value is less than the statistic, we can reject the null hypothesis, which means that the data does not follow a normal distribution. Otherwise we can accept the null hypothesis, the data does follow a normal distribution.

With this explanation in mind, the conclusions are the following:

* The feature sepal\_length does follow the normal distribution for both Iris-setosa and Iris-versicolor. Also we accept the hypothesis for class Iris-virginica when alpha is 0.01, but not for alpha 0.05.
* Regarding the feature sepal\_width, we can only assure it follows a normal distribution for Iris-versicolor at alpha 0.01 and Iris-setosa at alpha 0.01. For the rest of the combinations we will reject the hypothesis.
* For the feature petal\_length we can accept the hypothesis for classes Iris-versicolor at alpha 0.01 and Iris-virginica at alpha 0.01. For the rest of the combinations we will reject the hypothesis.
* Finally, regarding the feature petal\_width does not follow a normal distribution for any class.