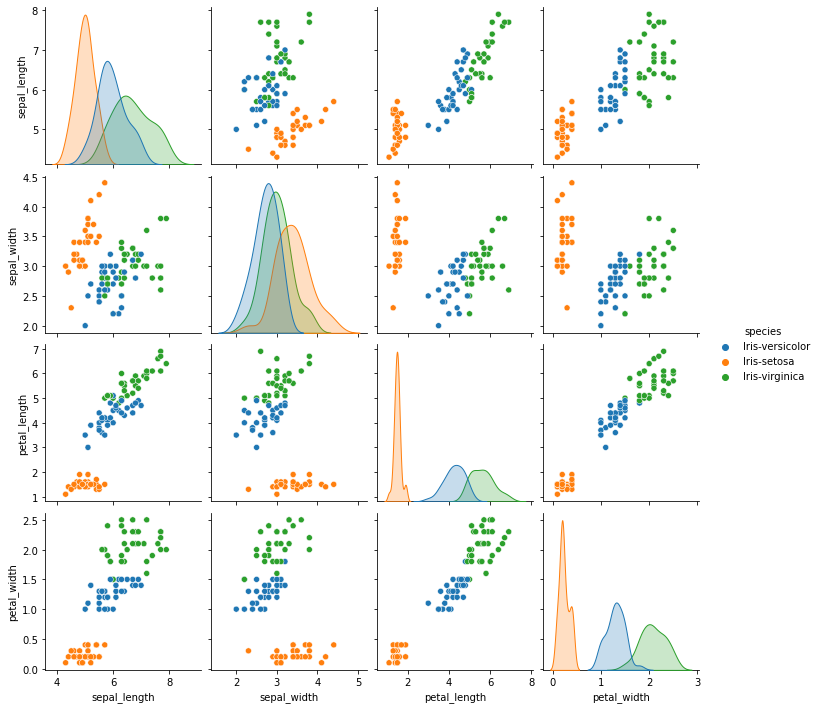
**Q 1.1**

**Iris Dataset**



**Comments about Iris pair plot**

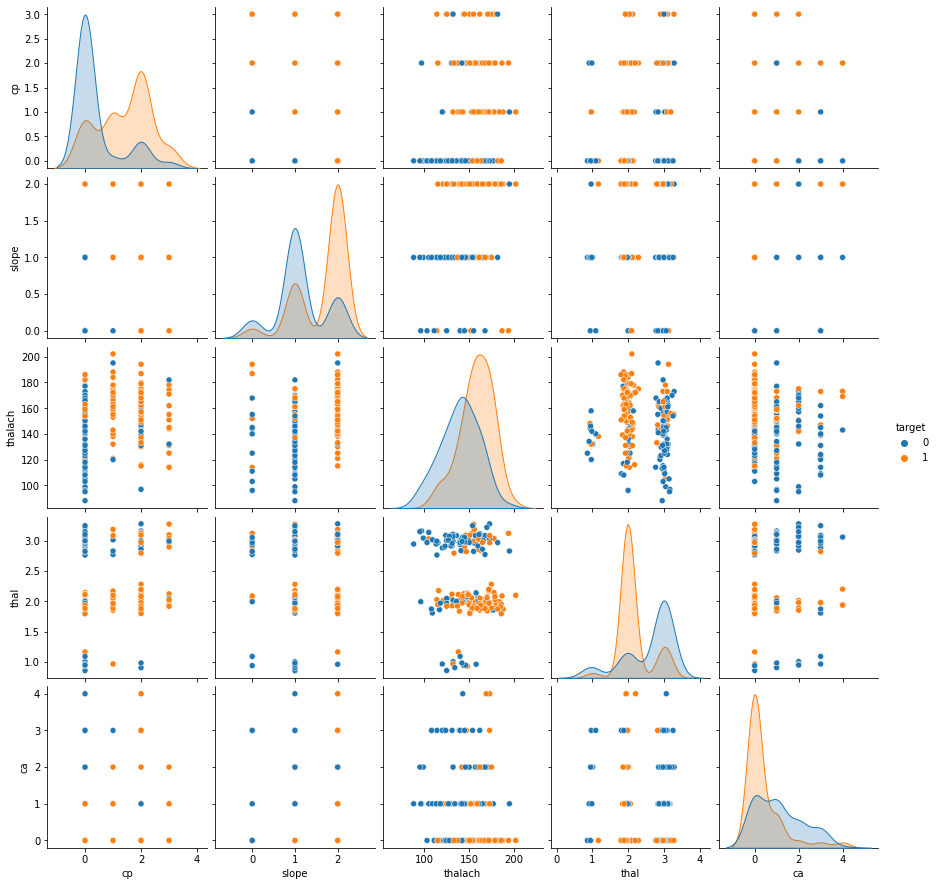
From the pair plot;

* There is a high correlation between petal width and petal length.
* Petal width and sepal width have a low correlation.
* There is a high correlation between petal length and sepal length.
* Sepal width has a low correlation with all other features

Looking at the distribution plots in the leading diagonal;

* Petal width seems to have the highest correlation with the species followed by petal length as a good line of separation can be drawn to separate the distribution curves for each specie in these features.
* Sepal width has the worst correlation with species as the distribution curves for the different species in this feature are almost completely overlapping.

**Heart Disease Dataset**



**Comments about Heart Disease Pair plot**

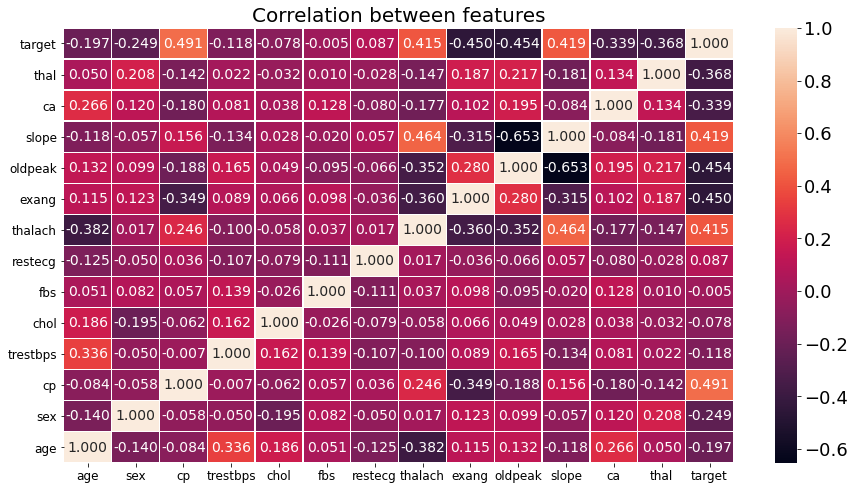
From the distribution plots in the leading diagonal;

* Most people without heart disease have a ca value of 0
* Most people with heart disease have a flat slope of peak exercise (slope = 2) and most without heart disease have slope = 1.
* The peak of the thal distribution for people with heart disease occurs at thal = 2
* The peak of the ca distribution for people with heart disease occurs at thal = 0

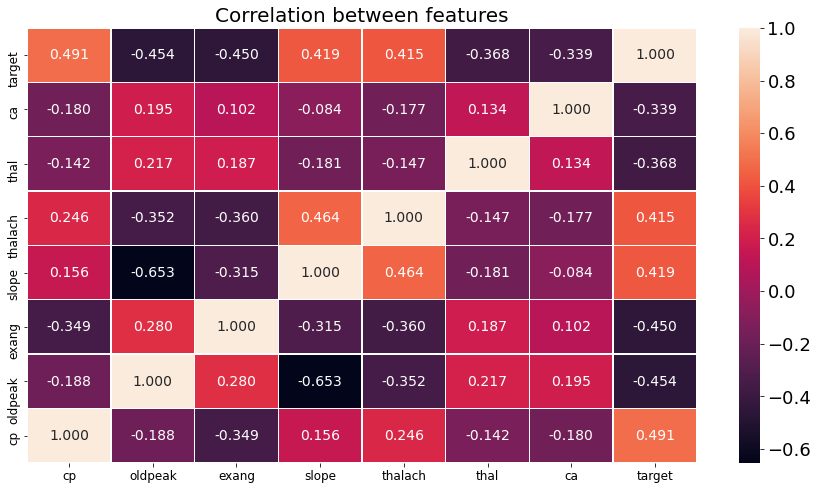
**Q 1.2**

**Why we chose our features**

We plotted the correlation heatmap of all features and studied correlation with the target to observe the inter-feature correlation and the correlation between each feature and the target.



Of all the features the following seven **- cp, oldpeak, exang, slope, thalach, thal, and ca** had the highest correlation with the target.



We narrowed our selection to **cp, slope, thalach, thal, and ca** for the following reasons:

* cp has the highest correlation with the target [0.491] so it is selected
* Oldpeak and slope have a high inter-feature correlation [-0.653] so we dropped one.
* Oldpeak has a high number of missing values [accounts for about 75% of missing values in the data] so we selected slope ahead of oldpeak.
* thal and ca have the least relative correlation with the other features so they are selected
* thalach was selected as it is the next numerical variable with the highest correlation with the target.