Rajat Srivastava

Consider a girl and boy for basic classification .

What will be common features between them -

- Nose
- Eyes
- Ears
- Lips
- Hair

What is different features between them -

- Maybe Nose ring .
- Eye lashes length
- Eye brow setup
- Maybe an Earing
- Lipstick
- Long or short hair
- Facial Hair

Now understand how a human may differentiate between a boy and girl on the basis of features like facial hair , lipstick , eyebrow tuning .

Features like nose, ears , lips are common to both boys and girls and don't give us a justifiable output , so they will become un-required features for me to differentiate .

Features like Earing or nose ring or even lipstick can also be used to differentiate between maximum humans but these can change too based on background of the person , race and culture . But they still might contribute towards major required features .

Similarly in Machine Learning , a machine also needs to classify between important or non important features to differentiate between cat or a dog , boy or a girl or any other use case it needs to be defined with the feature it has to take to differentiate between both entity .
In some cases these problems become more complex based on what needs to be classified and if classification classes are more .
And this can easily be done using feature score based on maybe a PCA or RFE .
Most important feature which will contribute to classify the entities will have the highest score .

Hence for feature selection -

- Consider optimal features which help differentiate .
- Consider one with a higher RandomForest classification score .
- Omit the ones which might seem to give repeated information as the selected ones (pearson correlation matrix ).
- Omit those which are common between classes as it just adds training time and complexity and even lead to misclassification .