1. Classes = 3.

Records = 150.

Class Distribution = 50 records from each class (Setosa, Versicolor, Virginica).

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Min | Max | Std Dev |
| Sepal length | 4.3 | 7.9 | 0.83 |
| Sepal width | 2.0 | 4.4 | 0.43 |
| Petal length | 1.0 | 6.9 | 1.76 |
| Petal width | 0.1 | 2.5 | 0.76 |

No missing data. No errors except there is some overlapping values especially in sepal width and length for the three types.

Petal length and width seem to be the most important attribute in classification. Testing petal length vs width as a pair of attributes would give best results and performance as it is most clearly distinguishing the types.

Conclusion: Data set is balanced as there is even distribution of each type. Most important attribute in general seems to be petal length due to the least overlap. Minority class issue would not occur which is good for performance. F1 will have to evaluated for maximum accuracy. From observation of the plots it seems that Sertosa is easily identifiable unlike in some cases of Versicolor and Virginica there are similar attributes and can overlap causing some errors.

1. Table of data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Percentage | F1 Macro | F1 Micro | Accuracy |
| Default Test Set | 15% | 0.9470899470899471 | 0.9565217391304348 | 0.9565217391304348 |
| Small | 5% | 1.0 | 1.0 | 1.0 |
| Medium | 40% | 0.9674603174603175 | 0.9666666666666667 | 0.9666666666666667 |
| Large | 80% | 0.9088688411549476 | 0.9083333333333333 | 0.9083333333333333 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Percentage | Depth | First Split | Second Split |
| Default Test Set | 15% | 5 | Petal width <= 0.8 | Petal width <= 1.65 |
| Small | 5% | 5 | Petal width <= 0.8 | Petal width <= 1.25 |
| Medium | 40% | 4 | Petal length <= 2.6 | Petal width <= 1.65 |
| Large | 80% | 2 | Petal width <= 0.6 | Petal length <= 5.0 |

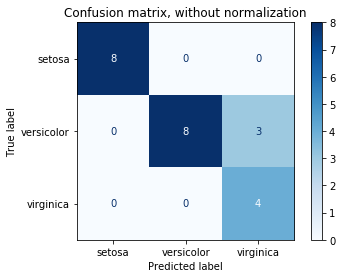
It is interesting to see that at medium test set the first split changes to petal length instead of petal width.

Cross validation: With the default test\_size set to 0.15, there is 10 fold cross validation. Which means that it uses 1/10th of data test for testing purposes and each time it iterates it starts from scratch, picking a new set of data to test/train.

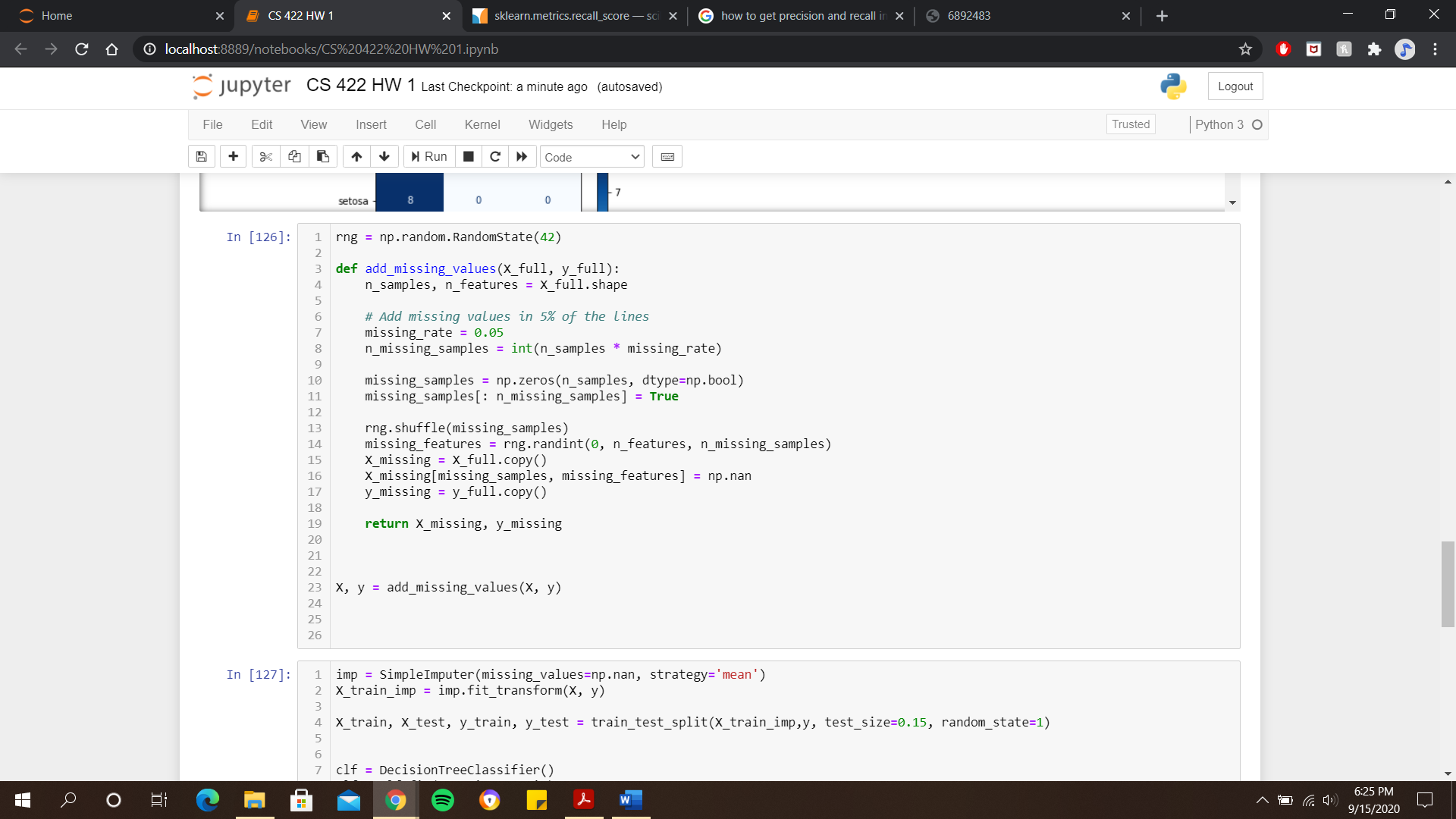
Test recall macro with Cross-validation: [0.96666667 0.96666667 0.9 0.96666667 1. ]

Test precition macro with Cross-validation: [0.96969697 0.96969697 0.9023569 0.96969697 1. ]

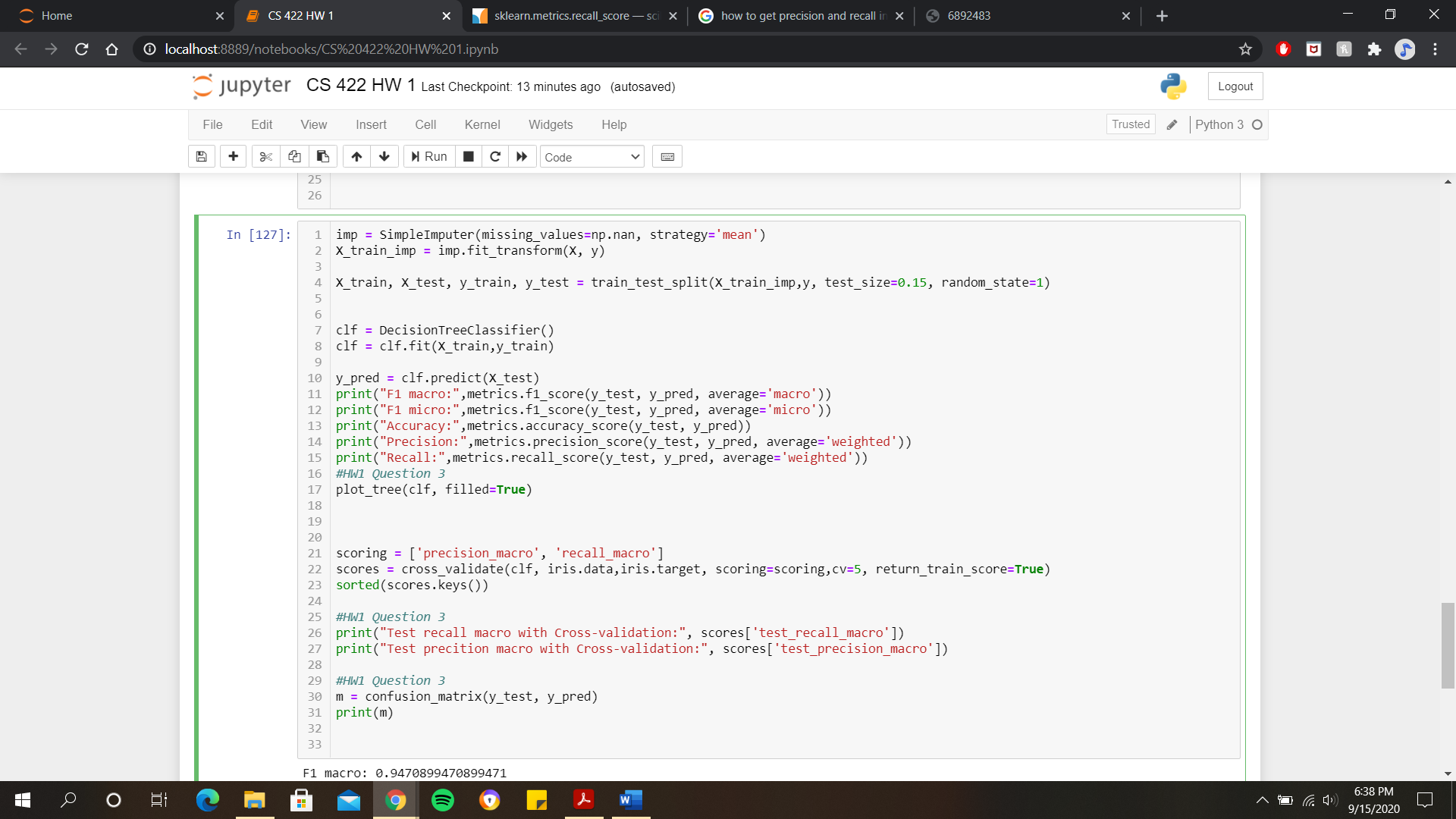
Confusion matrix conclusion:

This is what the confusion matrix comes out to be. The left label is True, and bottom label is Predicted. The top left to bottom right diagonal represents correctly predicted values.

Performance improves with greater number of cross validation folds



On the left is a snippet of the code used to add missing values to the dataset. The missing rate was changed to 5%, 10%, and 50%.

Then the dataset was filled with the mean for missing values and relearnt. Here’s the code that does that:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Missing values | Precision | Recall | F1 Macro | Accuracy | Training Time (s) |
| Default – 0% | 0.9652173913 | 0.95652173913 | 0.947089947 | 0.956521739 | 0.0000 |
| 5% | 0.9652173913 | 0.95652173913 | 0.947089947 | 0.956521739 | 0.0040011405 |
| 10% | 0.9420289855 | 0.91304347826 | 0.900000000 | 0.913043478 | 0.0019757747 |
| 50% | 0.9254658385 | 0.86956521739 | 0.853535353 | 0.869565217 | 0.0019648075 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Noise | Precision | Recall | Accuracy | F1 | Time (s) |
| 5% | 0.945 | 0.945 | 0.940 | 0.940 | 0.00 |
| 10% | 0.913 | 0.913 | 0.896 | 0.896 | 0.00 |
| 50% | 0.800 | 0.800 | 0.699 | 0.699 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute Values (petal width/sepal length \*1000) | Precision | Recall | Accuracy | F1 | Time (s) |
| 5% (Noise) | 0.945 | 0.945 | 0.940 | 0.940 | 0.00 |
| 10% | 0.913 | 0.913 | 0.896 | 0.896 | 0.00 |
| 50% | 0.800 | 0.800 | 0.699 | 0.699 | 0.00 |

It appears that even with changing attribute values (multiplying by 1000) there is no real effect on accuracy. Accuracy values remain same to when attributes were not modified by multiplying by 1000. This was tested on petal width and sepal length with same results. The conclusion is that both noise and missing values adversely affect the accuracy and F1 of the data. Time delay is more prominent in missing.