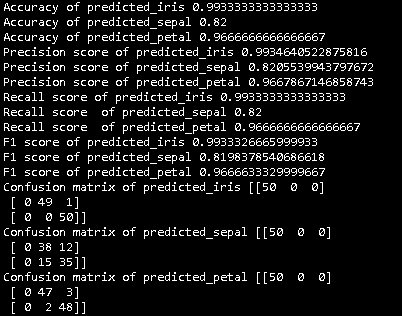
**Solutions and Screenshots**

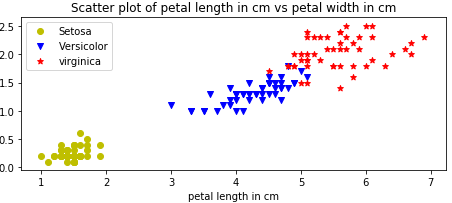
**Q1.1** What is the role of the fit and predict methods?

**A1.1** Fit is used to fit the model based on the training data X on target y and predict is for classifying the test data based on training data.

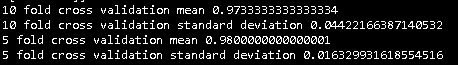
**Ans 1.2**



Because our training and texting data is same **predicted\_iris** has close to 100 percent accuracy. Next best model is **predicted\_petal** as feature set used is petal\_length and petal\_width and this feature set clearly distinguish all 3 classes of iris data as seen below -



**Ans 1.3**



To prevent overfitting, we use K fold cross validation where we divide data into k-1 folds and train our model using them. The remaining set of data is used to validate the model.

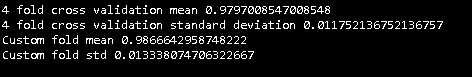
As our data set is small 5-fold cross validation performs better compared to 10 fold cross validation

**Ans 1.4**

Implemented function performs nearly as good as sklearn’s cross validation.



If instead of 5-fold cross validation we do a 4-fold cross validation, then implemented function works better compared to sklearn’s cross validation.



**Solution 2**

**Ans 2.1** categories specify which category to load if the list of categories is provided as parameter, load all categories if None is provided.

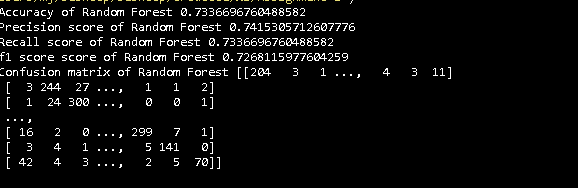
**Ans 2.2**

CountVectorizer - It converts text data into token counts matrix.

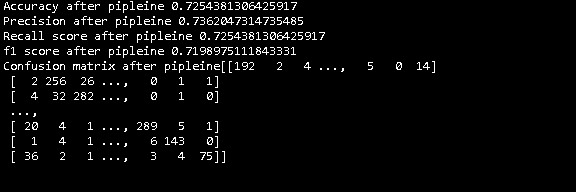
TfidfTransformer - The count matrix is transformed to a normalized tf or tf-idf

TfidfVectorizer - It converts text data TF-IDF feature matrix.

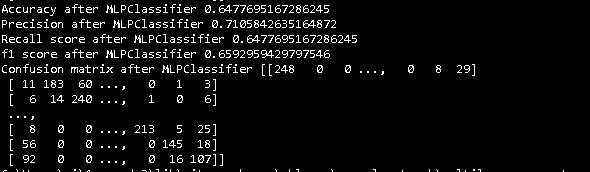
**Ans 2.3**



**Ans 2.4**



**Ans 2.5**



**Solution 3**

**Ans 3**

File name is **wine\_final\_classification.csv** that contains classification. Best result is produced by using polynomial kernel in support vector machines as the data points are not linearly separable. Also 2nd dimension of wine data that is ‘Ash’ is ignored as they are overlapping for different categories.

**References**

[1] "3.3. Model evaluation: quantifying the quality of predictions — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/model\_evaluation.html. [Accessed: 21- Sep- 2017].

[2] "3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/cross\_validation.html. [Accessed: 21- Sep- 2017].

[3] "5.6.2. The 20 newsgroups text dataset — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/datasets/twenty\_newsgroups.html. [Accessed: 21- Sep- 2017].

[4] "Sample pipeline for text feature extraction and evaluation — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/auto\_examples/model\_selection/grid\_search\_text\_feature\_extraction.html. [Accessed: 21- Sep- 2017].

[5] "Scikit Learn - Feature Extraction", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature\_extraction.text. [Accessed: 21- Sep- 2017].

[6] "1.17. Neural network models (supervised) — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/neural\_networks\_supervised.html. [Accessed: 21- Sep- 2017].