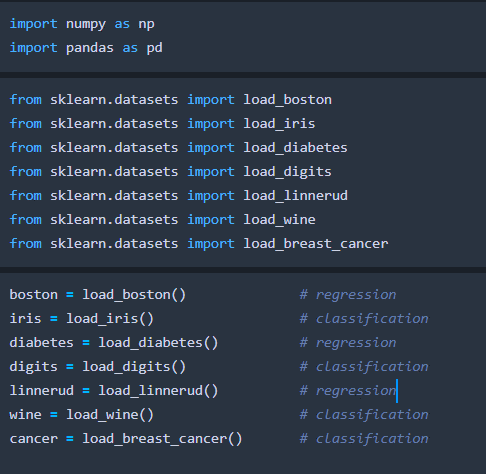
**Project Report**

**Sklearn + validation techniques**

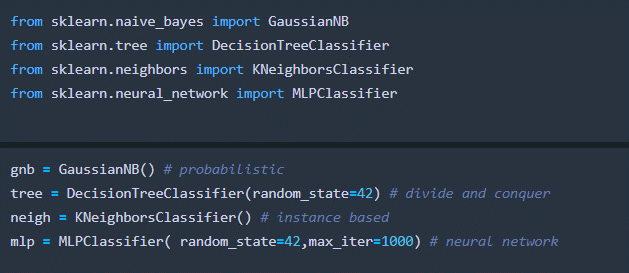
**Introduction:** In this project we have load the toy dataset that we load by sklearn. Then we have perform the classification and regression on toy data set. Here we have perform linear regression, random forest , knn, SVR, Decision tree classifiers , Gaussian naïve bias , MLP classifier (neural network) etc. Here we have load all dataset by sklearn. Here we have check the performance of model, accuracy of model, MSE (mean square value), by using t-test and p-value by using mlxtend package check the best algorithm which has highest performance.

**Load data set:**

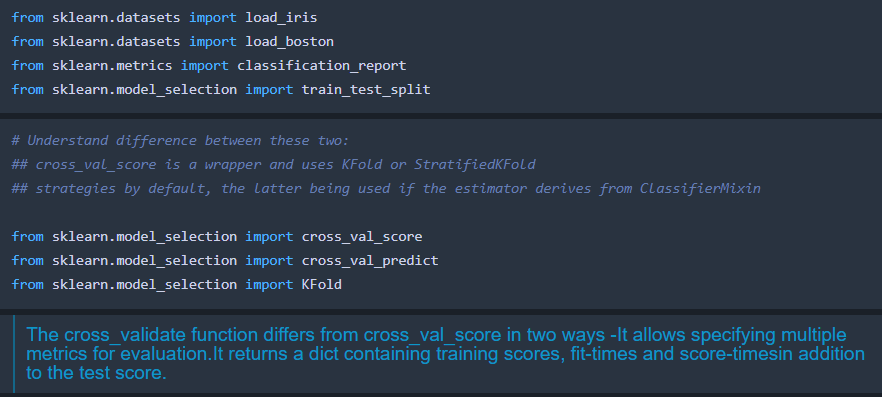


**Trained and run each classifier on each classification dataset separately:**

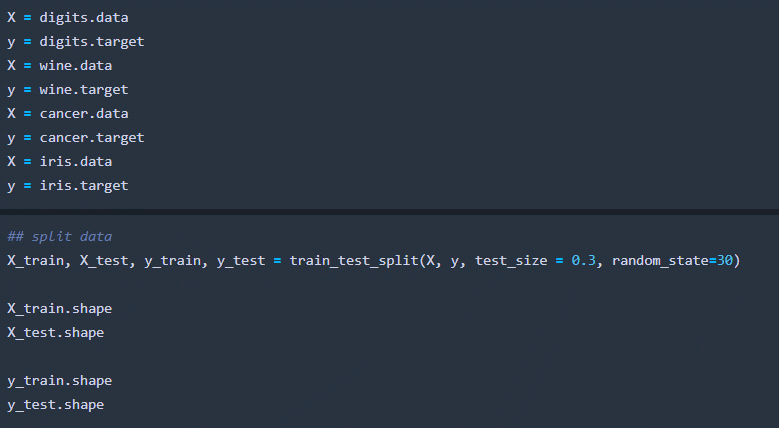
Import classifier packages and load them:



We have used cross validation on all classifier dataset:



**Feature selection and split the data:**

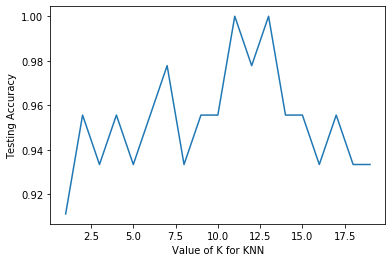


**Iris classification**: we have performed KNN classifiers here.

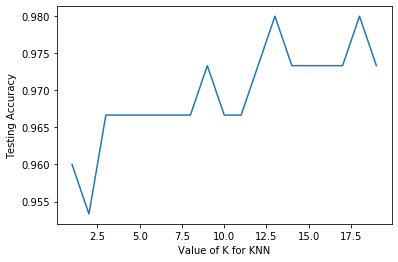


We get **97%** accuracy.

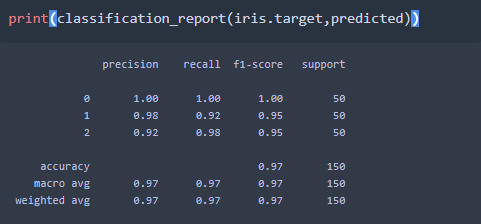
Evaluate with different K:



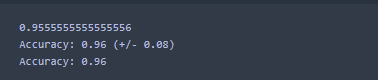
Evaluate with different K, using KFold split:



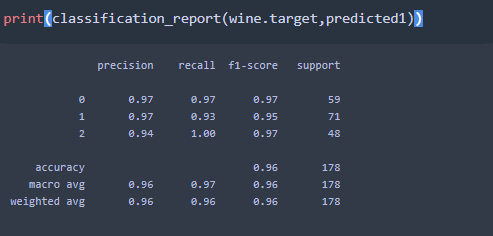
### Iris classification report



### Wine data classification



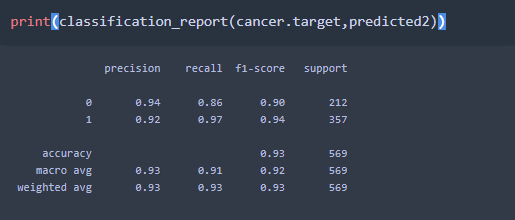
**Classification report:**



### Cancer Classification:

**We got** 93% accuracy.

**Cancer classification report:**



### Digits classification:

### accuracy score:

### 

### Classification report

### 

### 4 classifiers performed on each classification dataset (used given dataset, without test/train)

### 

#### **We have checked Classifier has highest mean accuracy across all datasets**

### 

### So Decision Tree classifier has highest mean accuracy across all dataset.

#### **We have checked dataset has the highest mean accuracy across all classifiers?**

### 

### 

### 3x2 matrix (dataset row, classifier column) MSE value:

### 

### Then we have displayed this matrix using imshow:

### 

### C:\Users\SPEED TAIL\Desktop\MIMP Project Regression and Classification on Sklearn Toy Dataset\imshow.png

### We have checked in which regression model has lower MSE (mean-squared error)

### 

### So here Support vector machine has highest mean-squared error.

### We have checked in which dataset has lowest MSE(mean squared-error) (across all regression methods which we have used ):

### 

### So Linnerud data set has lowest MSE values.

# **Q2: california housing prediction with validation**

### using gaussian naive bayes: for each instance output a probability that the house is worth over $300k (target variable is in units of $100,000's)

### fetch califirnia housing dataset:

### 

### We have got 100% accuracy.

### 

#### **B) perform k-fold (k=10) cross-validation (CV)**

### 

### Got 100% accuracy.

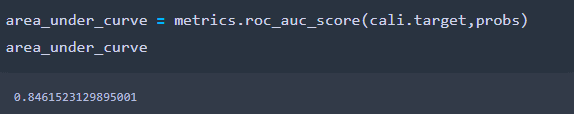
### Answer the following:

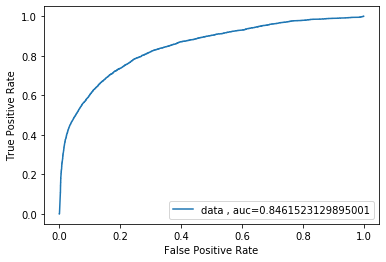
### the average error across all folds :

### 

# 2. C)Plot the ROC curve using the training data from 2a

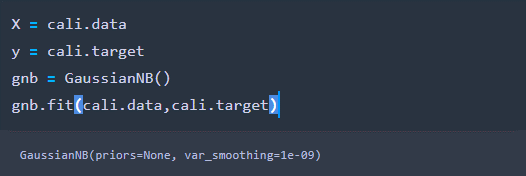
AUC value we got:



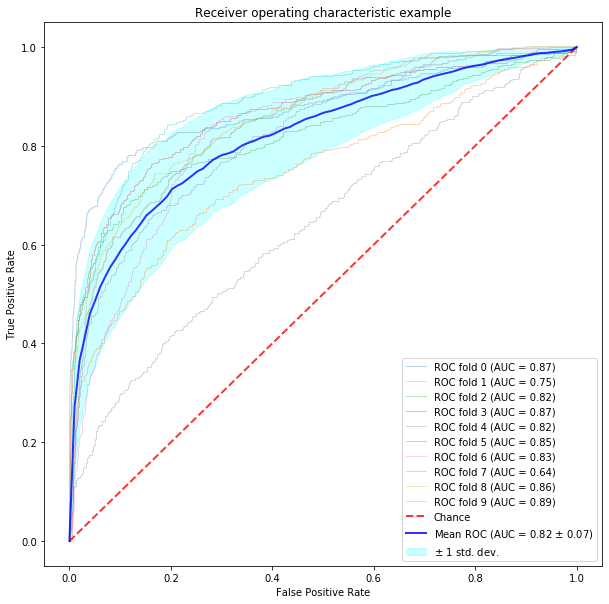


# Q3:

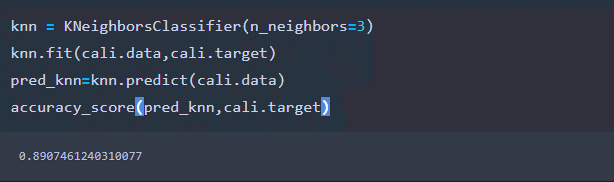
Gaussian naïve Bias:



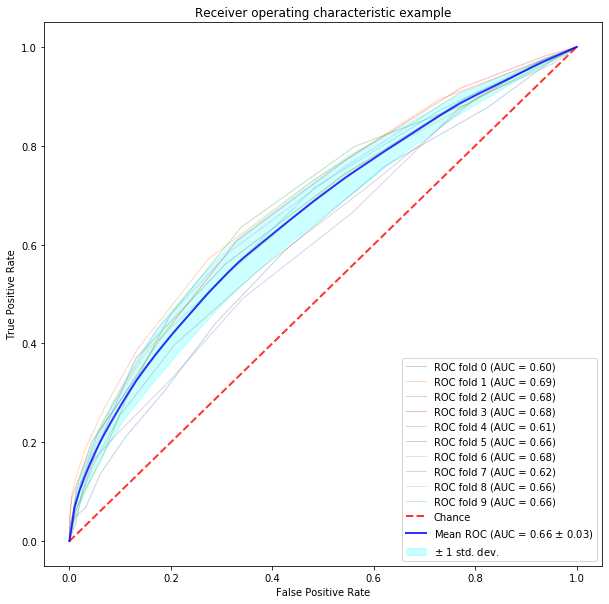
AUC curve for Gaussian naïve Bias:



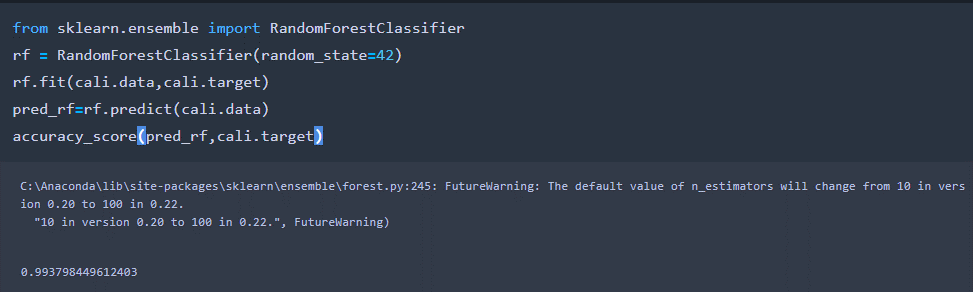
KNN classiferis:



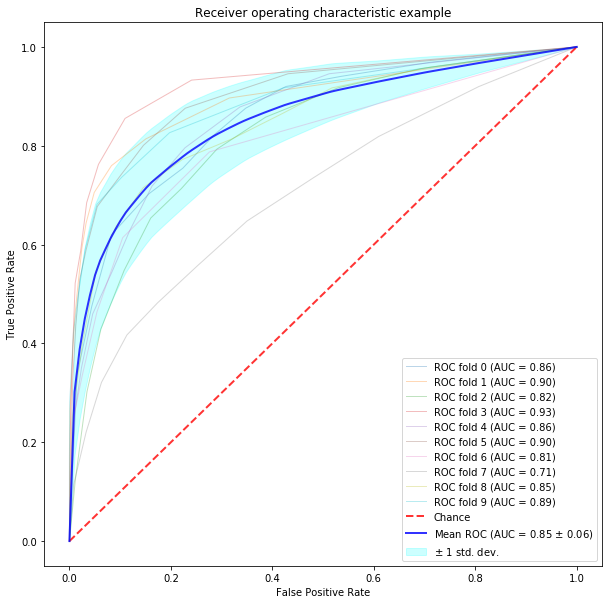
AUC curve for KNN:



Random Forest classifiers: got 99% accuracy



AUC curve for Random forest:



### We have compared the algorithm to check the highest area under average ROC curve from knn, Random forest and Gaussian naïve bias:

### By Gaussian naïve bias we get :

### 

### By KNN we get:

### 

### By Random forest:

### 

### So after comparing mean Roc values for Random forest algorithm gives the highest under average ROC curve.

### Now by using a t-test by using mlxtend package, we have compared the algorithm with the highest performance to the other two algorithm, to output a p-value and state if the p-value is significant (p<0.05):

### By using Gaussian Naïve Bias with KNN we got:

### 

### By using Gaussian naïve bias with Random forest algorithm we got:

### 

### By using Gaussian naïve bias with Random forest algorithm we got:

### 

### So we can say KNN classifier has highest performance than other classifiers on California data.

### Conclusion: In this project we have perform regression model and classification model on toy dataset and California dataset. Here we have learn all basic algorithm of machine learning with model evolution. We have check the model accuracy of all classifiers and regression models for all given data set. We have check the performance of model and we have check the best model by using t-test and p-values. We have seen random KNN classifier has highest mean AUC values and so we can say KNN classifier has highest performance than other classifiers on California data. “mlxtend” is important package to test the model , which has highest performance of model.

# **Reference:**

* Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd.
* Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd.
* Goksuluk, D., Korkmaz, S., Zararsiz, G., & Karaagaoglu, A. E. (2016). easyROC: an interactive web-tool for ROC curve analysis using R language environment. *RJ*, *8*(2), 213-230.
* Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1-9.
* Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine*, *44*(2), 368.
* David, A., Du, D., Larsen, K. G., Legay, A., & Mikučionis, M. (2013, May). Optimizing control strategy using statistical model checking. In *NASA Formal Methods Symposium* (pp. 352-367). Springer, Berlin, Heidelberg.
* <http://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_kfold_cv/>