

## COMP9321 – Assignment3

### Team FreshCoke

#### - Machine Learning:

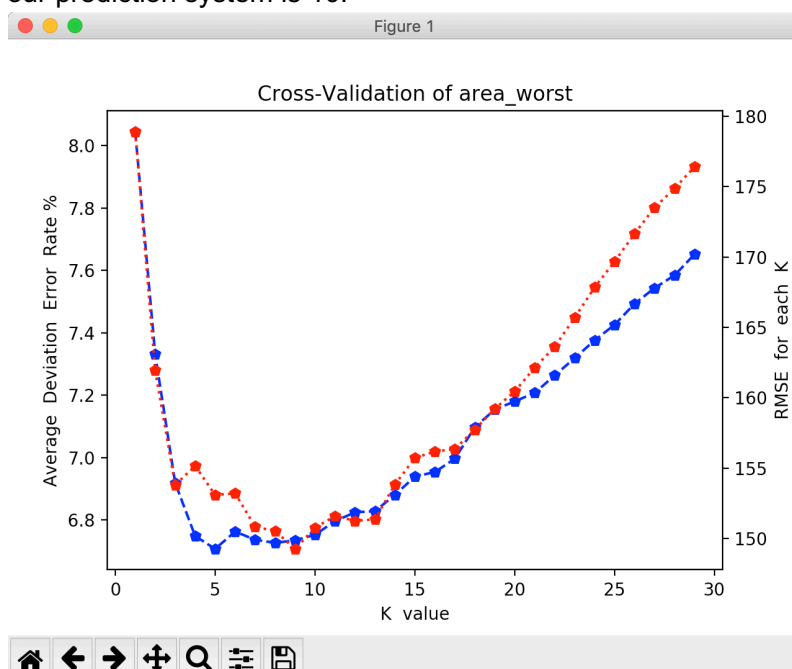
In the classification part, we test different ML models and finally choose SVM (support vector machine) as the model to train the dataset. The SVM cross-one-validation result can reach a high accuracy at **95.43%**.

We have also tried KNN to fit this dataset, but it shows lower accuracy than SVM. Random Forest can achieve a decent as well, but we are more familiar with SVM than this.

In the prediction part, we used K-NN as the algorithm to predict values of the features. Due to the nature of SVM, it is more suitable for classification but not the prediction. It is important to choose the best K number for the KNN to fit datasets, and that is why we use cross-validation again.

```
mr.zhao@zhaohaoxiangdeMacBook-Pro:~/Desktop/9321/ass3$ python cross_validation.py
feature "area_worst"
cross-validation: k=1, error=8.044150672721994, RMSE=178.89476460032583
cross-validation: k=2, error=7.330609229794198, RMSE=161.9456261933638
cross-validation: k=3, error=6.9185879035596045, RMSE=153.8038903287686
cross-validation: k=4, error=6.748459003416059, RMSE=155.1636248638854
cross-validation: k=5, error=6.707362670484963, RMSE=153.09332845297928
cross-validation: k=6, error=6.762698706789777, RMSE=153.21622482437436
cross-validation: k=7, error=6.736660821069288, RMSE=150.8345534757127
cross-validation: k=8, error=6.726643965928064, RMSE=150.5314777916586
cross-validation: k=9, error=6.734799177506511, RMSE=149.26451085686332
cross-validation: k=10, error=6.752679395334056, RMSE=150.7736575418193
cross-validation: k=11, error=6.7961037093872925, RMSE=151.59254405342554
cross-validation: k=12, error=6.824705862605011, RMSE=151.25785971633402
cross-validation: k=13, error=6.828106399403551, RMSE=151.36370649722653
cross-validation: k=14, error=6.880613631109106, RMSE=153.837127380887
cross-validation: k=15, error=6.939007865438975, RMSE=155.7279057485719
cross-validation: k=16, error=6.95367628068367, RMSE=156.17785916567058
cross-validation: k=17, error=6.99645944627709, RMSE=156.35997489613692
cross-validation: k=18, error=7.097427499305356, RMSE=157.73698358582902
cross-validation: k=19, error=7.154011795027131, RMSE=159.22423349938708
cross-validation: k=20, error=7.180147450475778, RMSE=160.4563693595947
```

After calculating error and RMSE (Root Mean Square Error) of 30 different Ks, the best K for feature “Area-Mean” is 9 (lower error and RMSE). Same as “Area-Mean”, we can finally find the best K for our prediction system is 10.



## - Table of Tasks:

Done	To do
Acquire dataset	Adjust webpage detail (more user friendly)
Data cleaning	UI polish
Machine Learning (classification and prediction)	
RESTful API server	
Back-end frame work	
Web/UI design (structure and layout)	
User Interaction Design	

## - Webpage review:



[Classification](#) [Prediction](#)

Fill in 5 features to classify diagnosis

Area Mean

143.5 ~ 2501.0

⌵

Area SE

6.802 ~ 542.2

⌵

Area Worst

185.2 ~ 4254.0

⌵

Perimeter Mean

43.79 ~ 188.5

⌵

Perimeter Worst

50.41 ~ 251.2

⌵

Submit

Clear

Classification results

Area Mean: None

Area SE: None

Area Worst: None

Perimeter Mean: None

Perimeter Worst: None

Description of all features

**idID:** number

**Diagnosis:** The diagnosis of breast tissues (M = malignant, B = benign)

**Radius Mean:** mean of distances from center to points on the perimeter

**Texture Mean:** standard deviation of gray-scale values

**Perimeter Mean:** mean size of the core tumor

Dataset [Show / Hide](#)