BREAST CANCER CLASSIFICATION

By:

Madhan Chaithanya,

MCA.

Problem Statement:

- Breast Cancer the most common cancer among women worldwide.
- Accounting for 25% of all cases.
- 2.1 million people were effected in 2015.
- The chances of survival has significantly increases by early diagnosis.
- Key challenge in cancer detection is to classify:
  - Malignant
  - Benign
- Using of Machine Learning techniques.
- It can dramatically improves the accuracy of diagnosis.

Motivation:

- Research indicates that most experience physicians:
  - Diagnose cancer with 79% accuracy.
- Using of ML techniques:
  - 91% accuracy is achieved by correct diagnosis.
- Our task is to classify tumours into benign or malignant using some features from several images.

Motivation(cont..):

- Using Classification Algorithms of ML:
  - Support Vector Machine
  - K-NN(K-Nearest Neighbours)
  - Decision Tree

- Using CNN (if possible)

Literature Survey: 1:

- Breast Cancer Detection with Reduced Feature Set
- **AhmetMert,[1] Niyazi KJlJç,[2] Erdem Bilgili,[1] and Aydin Akan**
- Hindawi Publishing Corporation Computational and Mathematical Methods in MedicineVolume 2015, Article ID 265138, 11 pages (http://dx.doi.org/10.1155/2015/265138)

Literature Survey: 2:

- Support vector machines combined with feature selection for breast cancer diagnosis.-Mehmet Fatih Akay
- Expert Systems with Applications (2009) 3240–3247
- [0957-4174/$ - see front matter 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.eswa.2008.01.009]

Literature Survey:
3:

- Breast Cancer Classification Using Machine Learning
- Meriem AMRANE[1] Saliha OUKID[2] Computer Science Department.
- IEEE Conference ©2018.

Literature Survey: 4:

- Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules
- [International Journal of Computer Applications (0975 - 8887)Volume 62 - No. 1, January 2013]

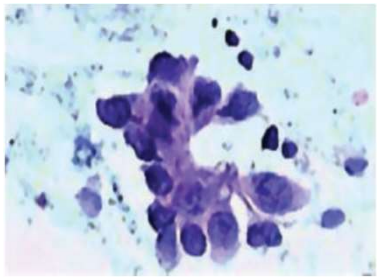
Requirements:

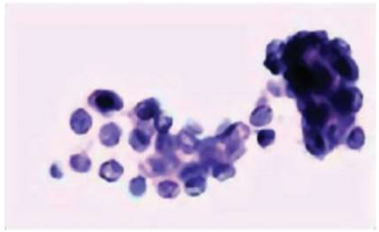
- OS : windows 10.
- Coding: Python-3.
- Platform : Anaconda Distribution
  - Jupyter (or)
  - Spyder
- Dataset:
  - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Cancer Diagnosis Procedure:

FINE NEEDLE ASPIRATE(FNA)

TUMOUR IMAGES



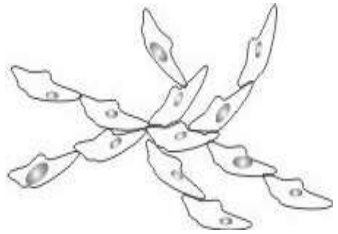


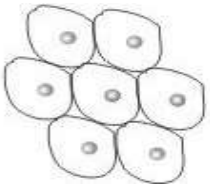
FEATURES

RADIUS
TEXTURE
PERIMETER
AREA
SMOOTHNESS

CLASSIFICATION

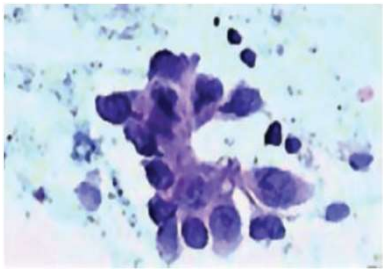
MALIGNANT

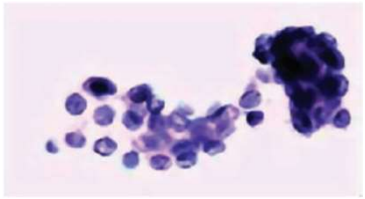




BENIGN

ML Terms(Dataset):





RADIUS
TEXTURE
PERIMETER
AREA
SMOOTHNESS
COMPACTNESS
CONCAVITY
CONCAVITY_POINTS
SYMMETRY
FRACTIONAL_DIMENSION

CLASSIFIER

MALIGNANT

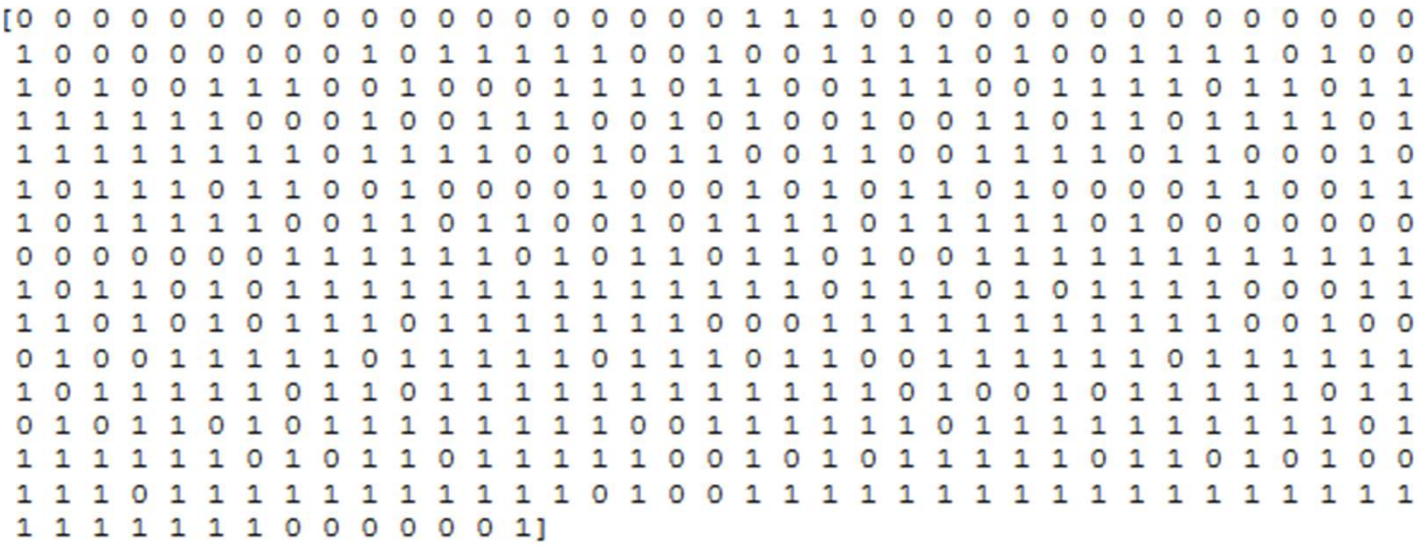
BENIGN

Dataset:

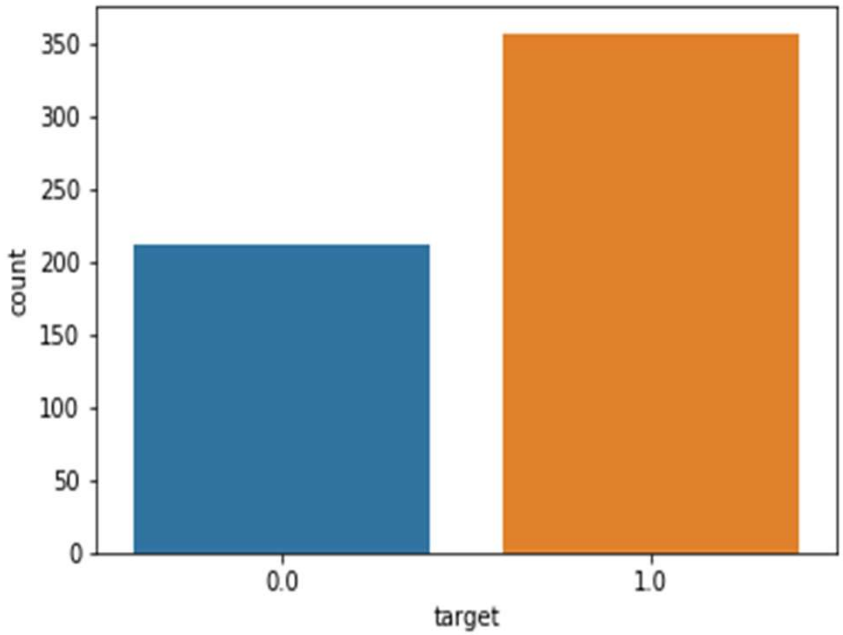
- Number of instances: 569
- Class distribution:
    - 212 Malignant
    - 357 Benign

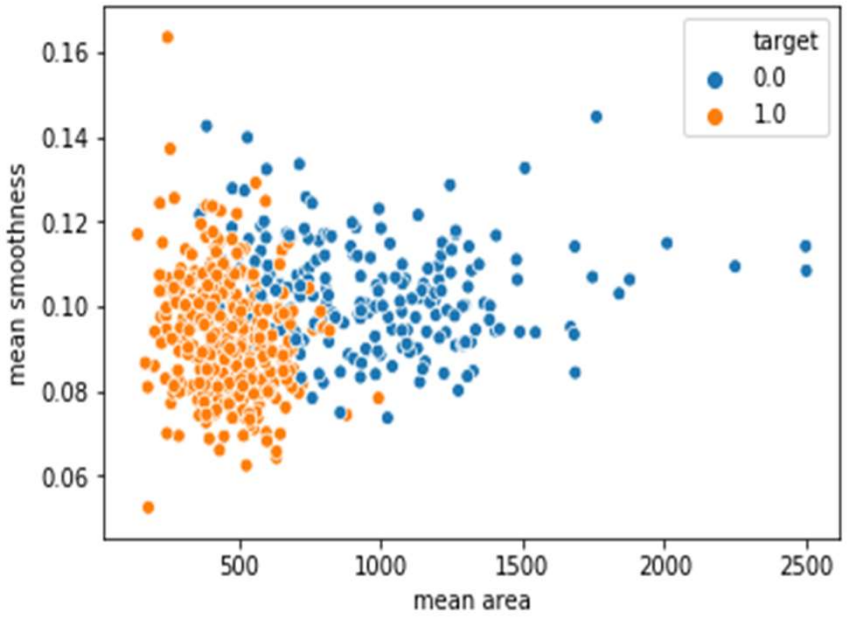
Expected Output:

- Based on this:
  - What we do
  - How we treat the classifier(training it)
  - When teach it – how we classify it
- Looking into 30 features from that we are teaching the target class.
  - 0 - Malignant
  - 1 - Benign
- It's a binary kind.

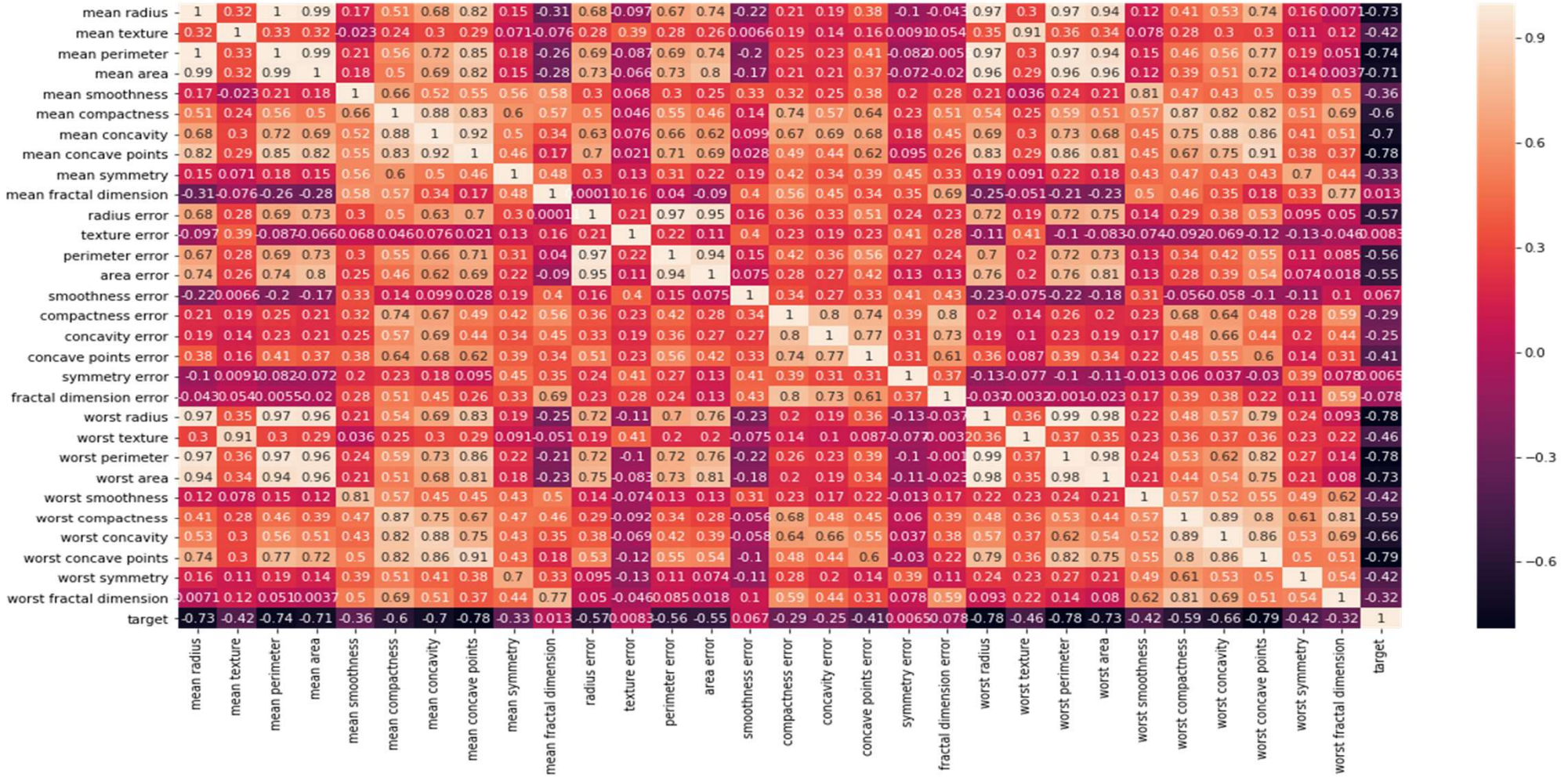


Visualizing the Data:

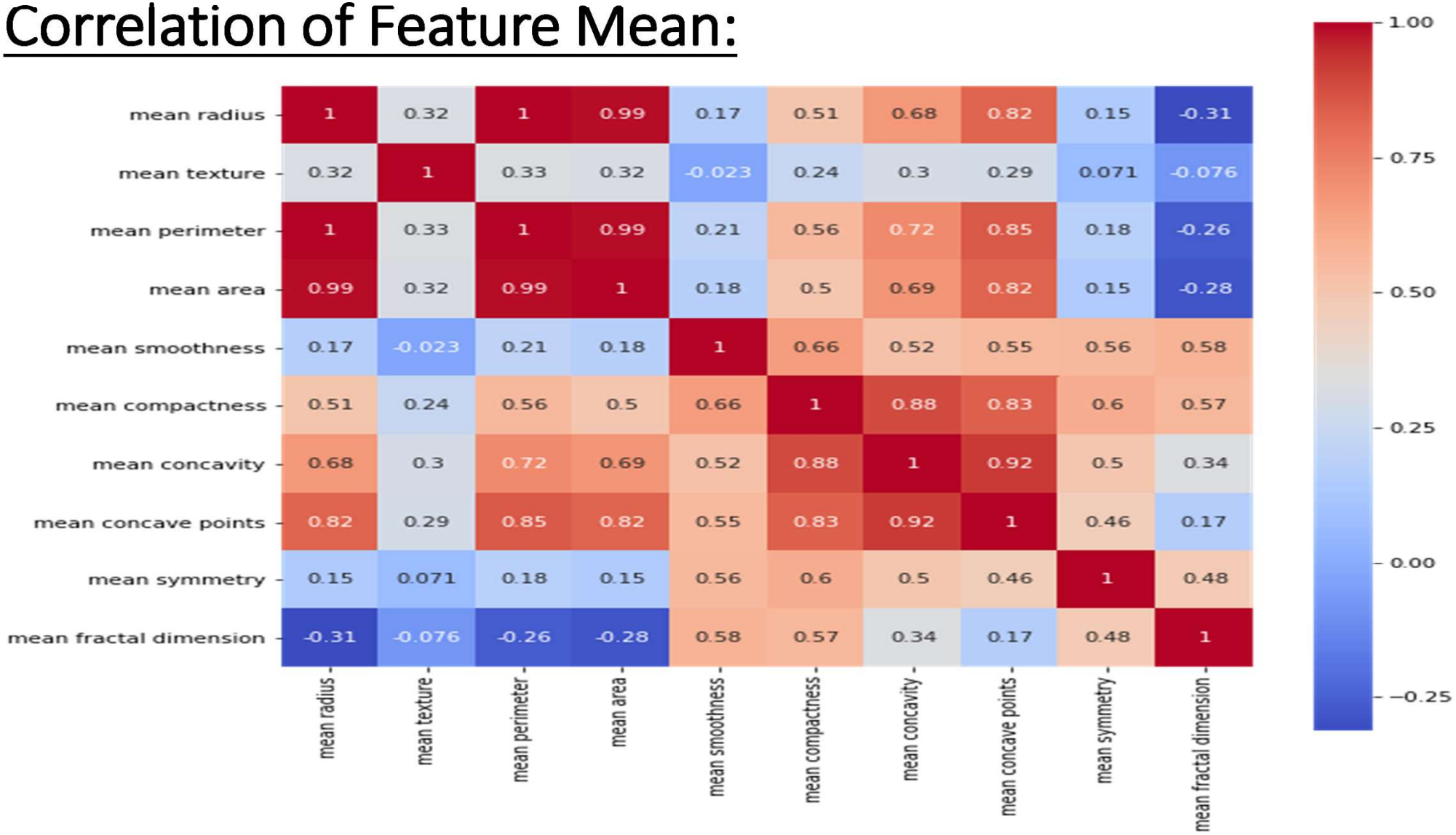




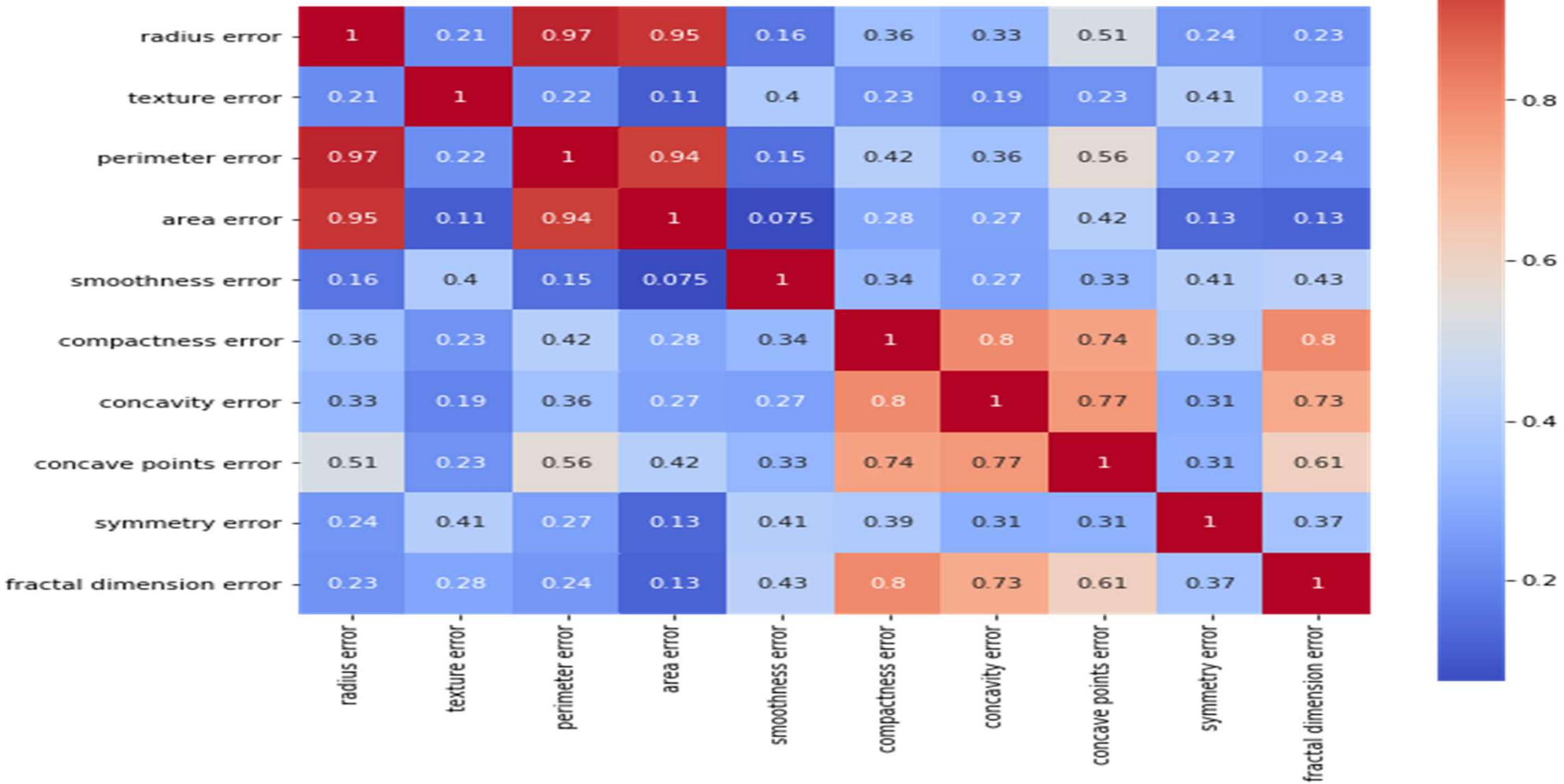
Correlation of Feature Names:



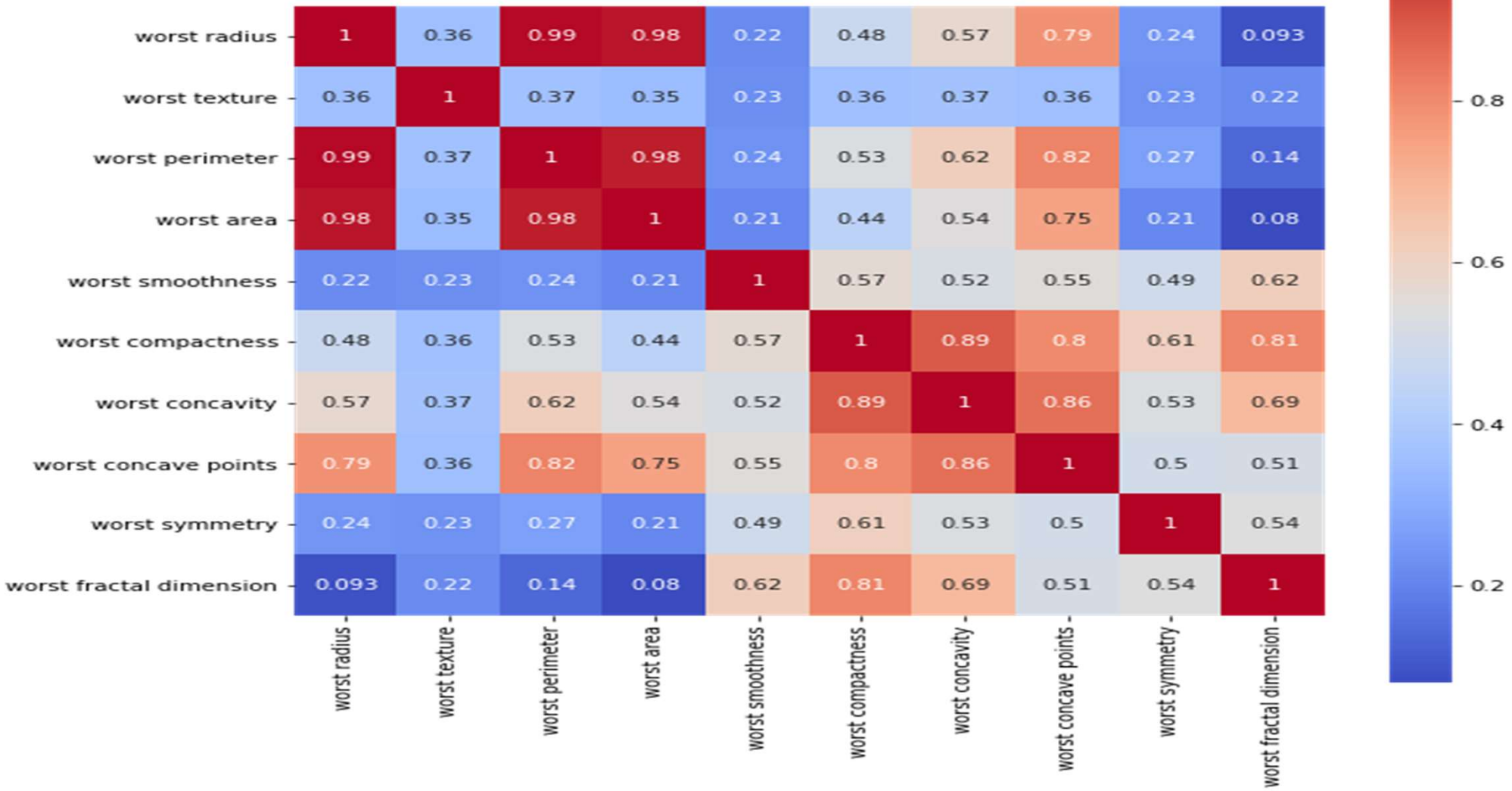
Correlation of Feature Mean:



Correlation of Feature Errors:



Correlation of Features Worst:



Model Training & Testing:

- X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 5)
- print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
- Output:

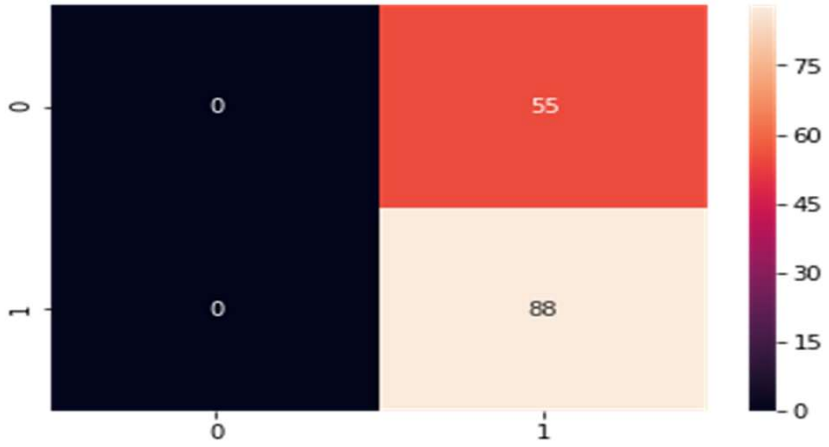
```
(426, 30) (143, 30) (426,) (143,)
```

SVM Results:

- Directly applying SVM

```
cm = confusion_matrix(y_test, y_predict)
cm
```

```
array([[ 0, 55],
       [ 0, 88]], dtype=int64)
```

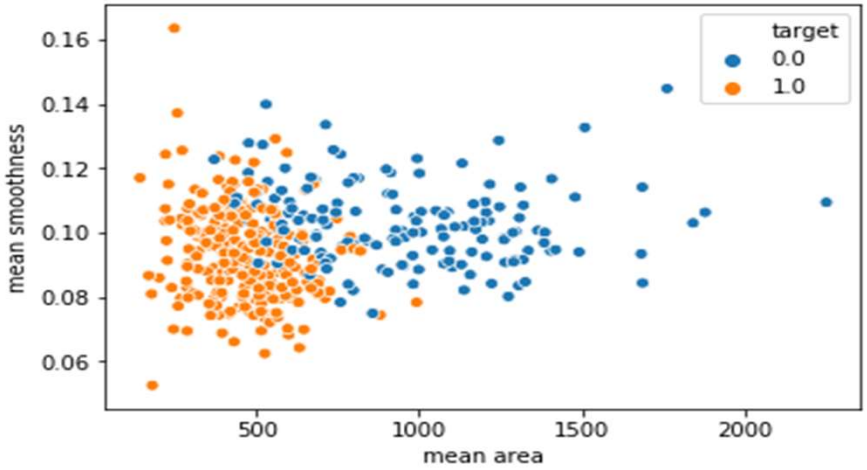


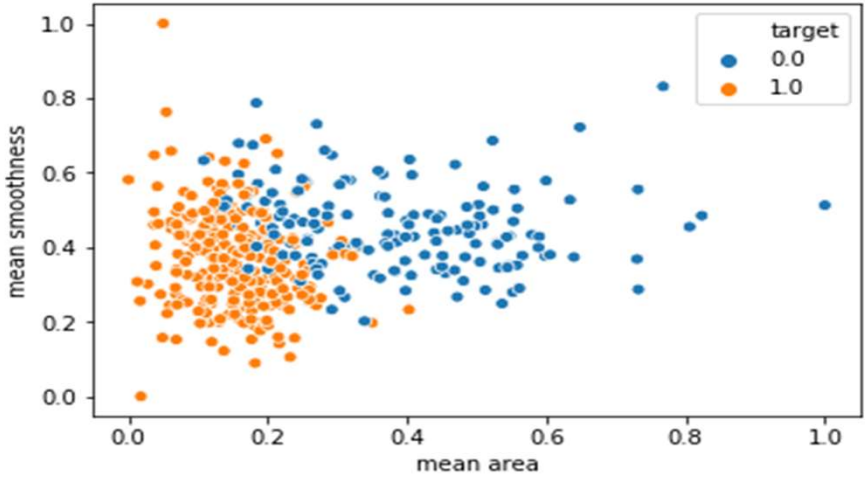

```
print(classification_report(y_test, y_predict))
```

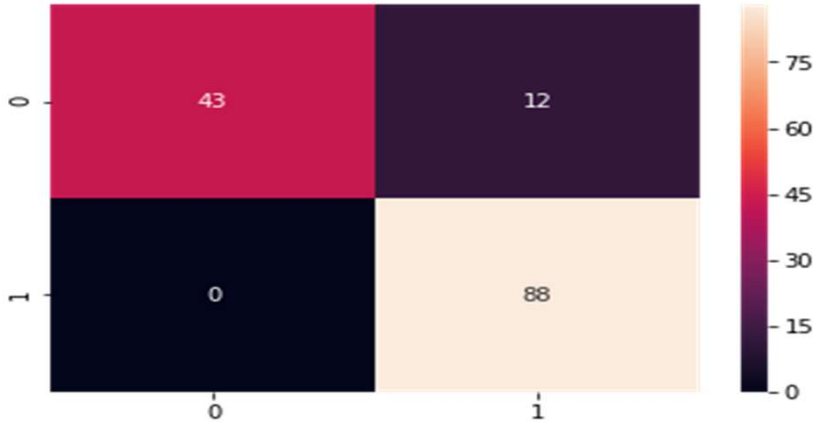
	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	55
1.0	0.62	1.00	0.76	88
avg / total	0.38	0.62	0.47	143

SVM Results(Cont)

- Improving the Model by applying Data Normalization





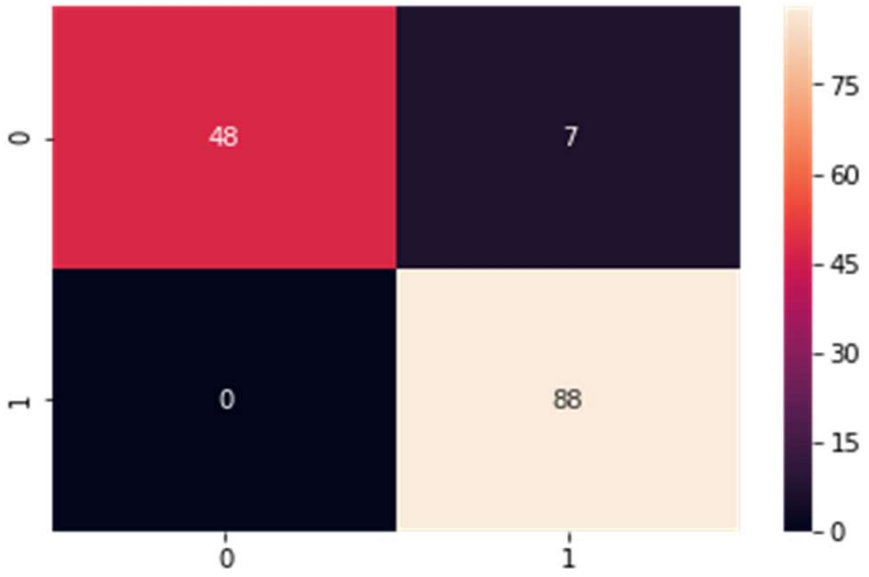



```
print(classification_report(y_test, y_predict))
```

	precision	recall	f1-score	support
0.0	1.00	0.78	0.88	55
1.0	0.88	1.00	0.94	88
avg / total	0.93	0.92	0.91	143

SVM Results(cont):

- Improving the model by using – ‘C’ , ‘gamma’ parameter through by applying kernel = ‘rbf’.
- param_grid = {'C':[0.1,1,10,100], 'gamma':[1,0.1,0.01,0.001], 'kernel':['rbf']}



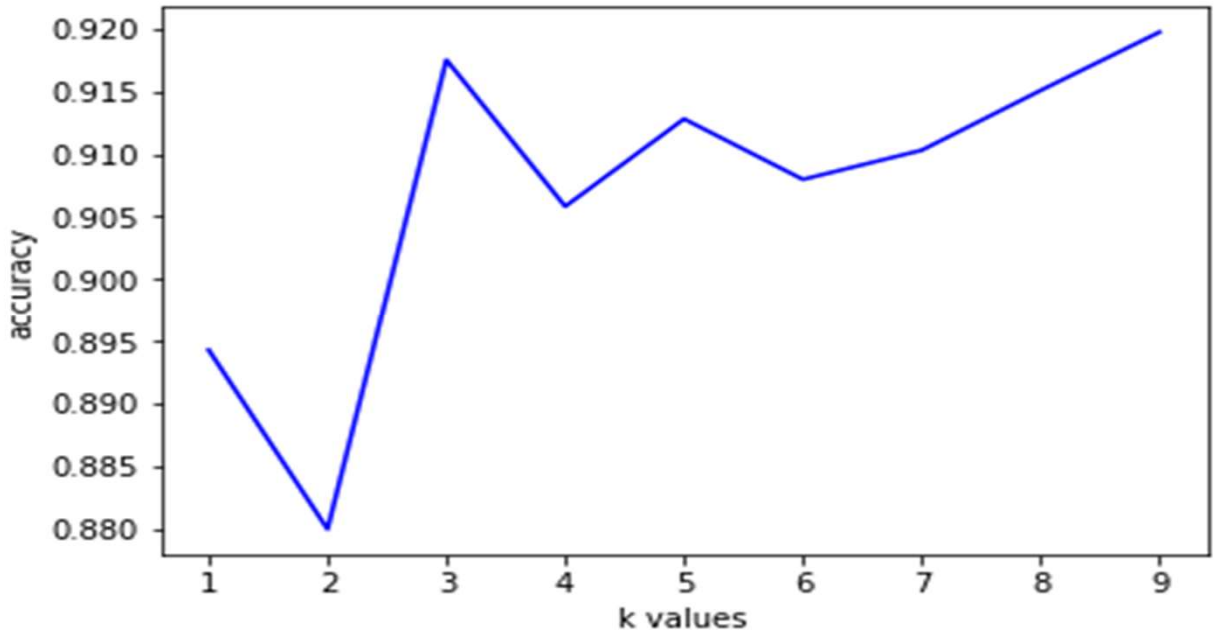

```
print(classification_report(y_test, grid_predict))
```

	precision	recall	f1-score	support
0.0	1.00	0.87	0.93	55
1.0	0.93	1.00	0.96	88
avg / total	0.95	0.95	0.95	143

KNN Results:

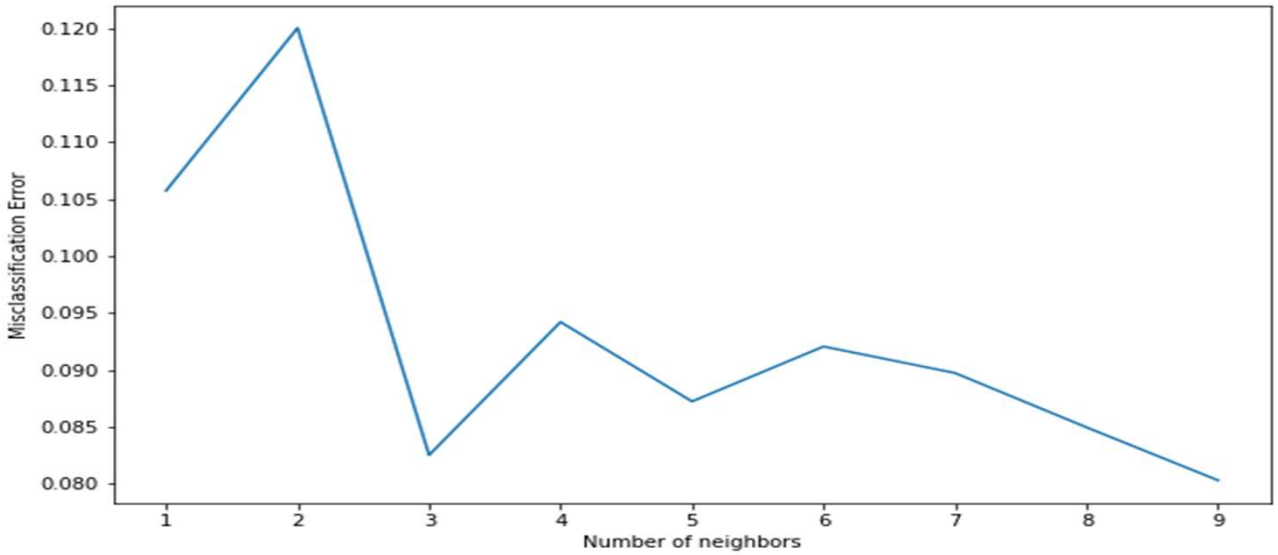
- By applying the cross validation score model

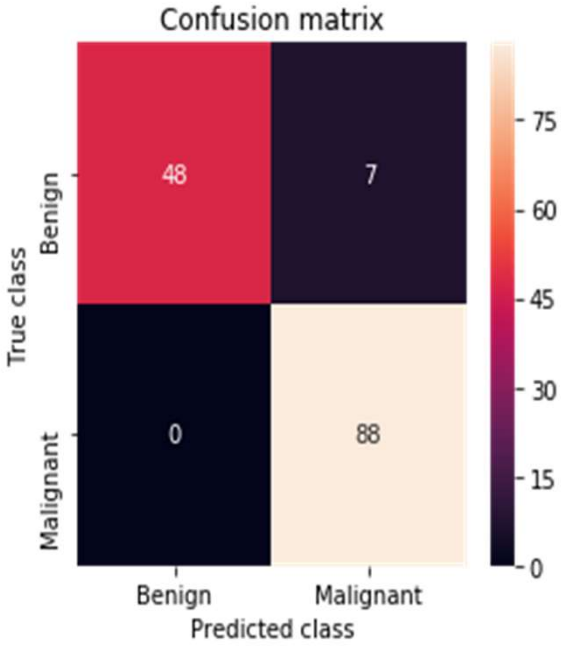
```
[0.894 0.88  0.917 0.906 0.913 0.908 0.91  0.915 0.92 ]
```



KNN Results:

- By applying misclassification error.
- Output : The optimal number of neighbors is 9





KNN Results:

```
print(classification_report(y_test, y_pred))
```

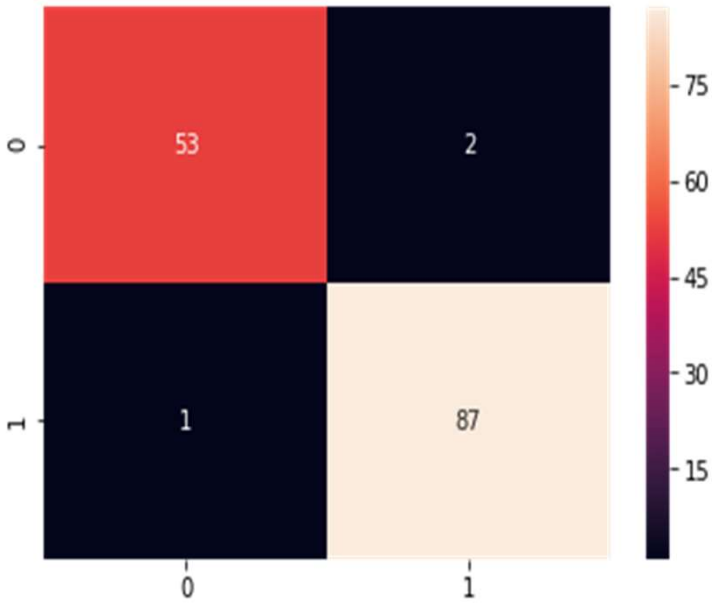
	precision	recall	f1-score	support
0.0	1.00	0.87	0.93	55
1.0	0.93	1.00	0.96	88
avg / total	0.95	0.95	0.95	143

- By applying grid search of accuracy and recall.

```
# Tuning hyper-parameters for accuracy
{'n_neighbors': 11}
0.923
# Tuning hyper-parameters for recall
{'n_neighbors': 35}
0.97
```

Logistic Regression:

- By applying the default things we get




```
print(classification_report(y_test, y_predict))
```

	precision	recall	f1-score	support
0.0	0.98	0.96	0.97	55
1.0	0.98	0.99	0.98	88
avg / total	0.98	0.98	0.98	143

Logistic Regression:

- By applying the parameter grid for accuracy and recall we get

```
# Tuning hyper-parameters for accuracy
{'C': 1000}
0.958
# Tuning hyper-parameters for recall
{'C': 10}
0.974
```

Project Plan For Final :

- Showing the results by using:
    - Decision Tree Classifier & Random Forest Classifier.
- By tuning the parameters also.
- Lastly providing the comparative best one.

References:

- [1] Breast Cancer Detection with Reduced Feature Set. **AhmetMert,1 Niyazi KJlJç,2 Erdem Bilgili,1 and Aydin Akan2** *1Department of Electrical and Electronics, Piri Reis University, 34940 Istanbul, Turke2Department of Electrical and Electronics, Istanbul University, 34320 Istanbul, Turkey(*Hindawi Publishing Corporation Computational and Mathematical Methods in MedicineVolume 2015, Article ID 265138, 11 pages(http://dx.doi.org/10.1155/2015/265138)
- [2] Breast Cancer Classification Using Machine Learning. Meriem AMRANE1Saliha OUKID 2 Computer Science Department, LRDSI Laboratory, University of Blida 1, Blida, Algeria Ikram GAGAOUA3 Tolga ENSAR 4 Computer Engineering, Istanbul University,Istanbul, Turkey[78-1-5386-5135-3/18/$31.00 ©2018 IEEE
- [3] Breast Cancer Classification of Image using Convolutional Neural Network.[978-1-5386-3039-6/18/$31.00©2018 IEEE]
- [4] Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules[International Journal of Computer Applications (0975 - 8887)Volume 62 - No. 1, January 2013]
- [5]Support vector machines combined with feature selection for breast cancer diagnosis. Mehmet Fatih Akay *Department of Electrical and Electronics Engineering, Cukurova University, Adana 01330, Turkey. [0957-4174/$ - see front matter 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.eswa.2008.01.009]
- [6] https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis
- [7] https://www.cancer.org/treatment/understanding-your-diagnosis/tests/testing-biopsy-and-cytology-specimens-for-cancer/how-is-cancer-diagnosed.html

Thank You