Dr. Sudhir Dhage
Mentor

Minor Project

Breast Cancer Prediction using Machine Learning

Jash Jain
2019130021

Kashish Jain
2019130022

Manthan Juthani
2019130028

Problem Statement

To enhance early detection of Breast Cancer in order to improve the prognosis and chances of survival significantly using state-of-the-art machine learning algorithms like Random Forest Classification ,Support Vector Machines etc.

DATA PREPARATION

Using the UCI Machine Learning Repository for breast cancer dataset.

MACHINE LEARNING ALGORITHMS

Logistic Regression,K Nearest neighbours,Support Vector Machines,Decision Trees,Naive Bayes,Random Forest

RESULTS

Using Confusion matrix,precision,recall and accuracy parameters

Scope of Work

ASSUMPTIONS

Assuming radiologists data is accurate which may not always be the case

Assuming dataset provided is sufficient and apt for the research

BOUNDARIES

Limited to data in form of variables rather than digitally accurate images

Non-availability of high computational power resulting in not using the approach of Neural Networks

Scope of Work

Objectives

FEATURE SELECTION

Selecting the most prominent features contributing to the detection of breast cancer

CHOICE OF ALGORITHM

Choosing from the Machine learning algorithms the one which gives the highest classification between benign and non benign

DEPLOYMENT

Converting the jupyter notebook to a WebApp thereby providing a convenient GUI for visitors

Implementation Details

TECH STACK USED

Jupyter Notebook,
Numpy,Pandas,Matplotlib

Django,Html,CSS,Bootstrap,
Python,

Sklearn

API'S USED

Amazon Web Services(AWS) for deployment and providing a web API and deploying the machine learning model

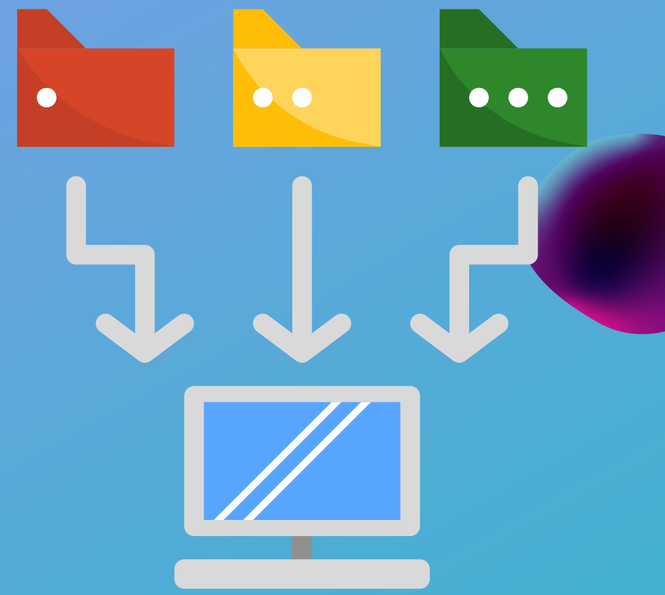
Data Preparation

```
dataset = pd.read_csv('cancer_modified.csv')
X = dataset.iloc[:, 1:31].values
Y = dataset.iloc[:, 31].values
```

```
dataset.head()
```

	id	diagnosis	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	mean_compactness	mean_concavity	mean_concave points	...
0	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
1	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
2	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
3	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...
4	843786	M	12.45	15.70	82.57	477.1	0.12780	0.17000	0.1578	0.08089	...

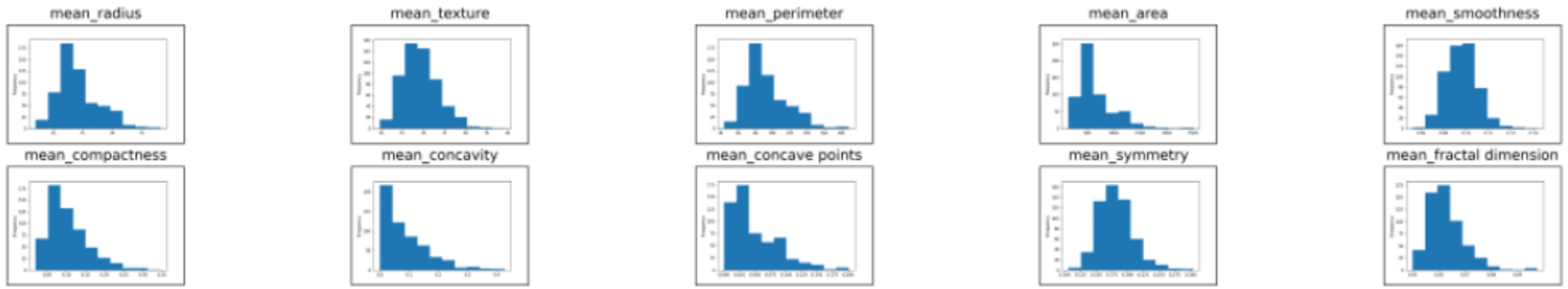
5 rows × 32 columns



Downloading the data from the Uci repository-adding labels-creating headers-converting to csv

Plotting graphs

```
i=0
fig = plt.figure(figsize=(25, 4),dpi=400)
for idx in np.arange(10):
    ax = fig.add_subplot(2, 10/2, idx+1, xticks=[], yticks=[])
    mpimg_img = mpimg.imread(os.path.join('graph_images', images[i]+'.png'))
    ax.imshow(mpimg_img)
    ax.set_title(images[i])
    i+=1
```



Plotting histograms for the attributes and deriving important conclusion from the same

Nan value check,

```
In [24]:  1  dataset.isnull().sum()
          2  dataset.isna().sum()
```

```
Out[24]: id                    0
         diagnosis             0
         mean_radius           0
         mean_texture          0
         mean_perimeter        0
         mean_area             0
         mean_smoothness       0
         mean_compactness      0
```

Checking for Nan Values which add noise to the dataset

Categorical to numerical data

```
1 from sklearn.preprocessing import LabelEncoder
2 labelencoder_Y = LabelEncoder()
3 dataset['diagnosis'] = labelencoder_Y.fit_transform(dataset['diagnosis'])
```

```
1 dataset.tail()
```

	id	diagnosis	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	mean
563	926424	1	21.56	22.39	142.00	1479.0	0.11100	
564	926682	1	20.13	28.25	131.20	1261.0	0.09780	
565	926954	1	16.60	28.08	108.30	858.1	0.08455	
566	927241	1	20.60	29.33	140.10	1265.0	0.11780	
567	92751	0	7.76	24.54	47.92	181.0	0.05263	

5 rows × 32 columns

Conversion of M and B to 1 and 0 for training of the machine learning model

Trying various machine learning models

Using Logistic Regression Algorithm to the Training Set

In [36]:
```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, Y_train)
```

Out[36]:
```
LogisticRegression(random_state=0)
```

Predicting the results on the test dataset

In [37]:
```
Y_pred = classifier.predict(X_test)
```

Testing via the confusion matrix

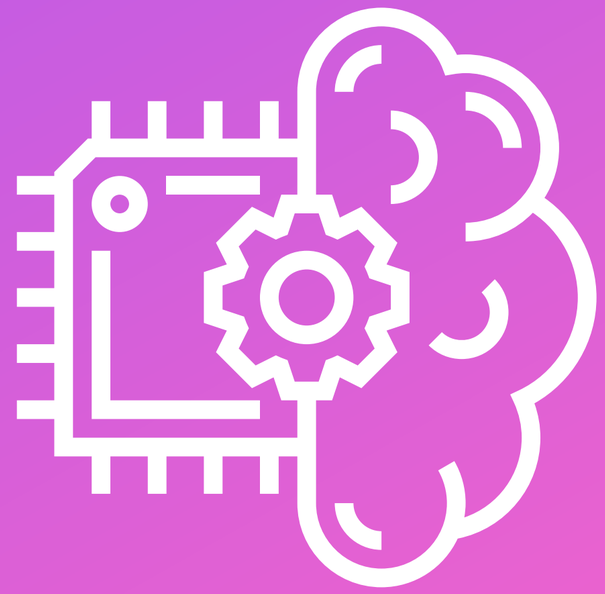
In [39]:
```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred)
```

In [40]:
```
cm
```

Out[40]:
```
array([[91,  1],
       [ 3, 47]], dtype=int64)
```

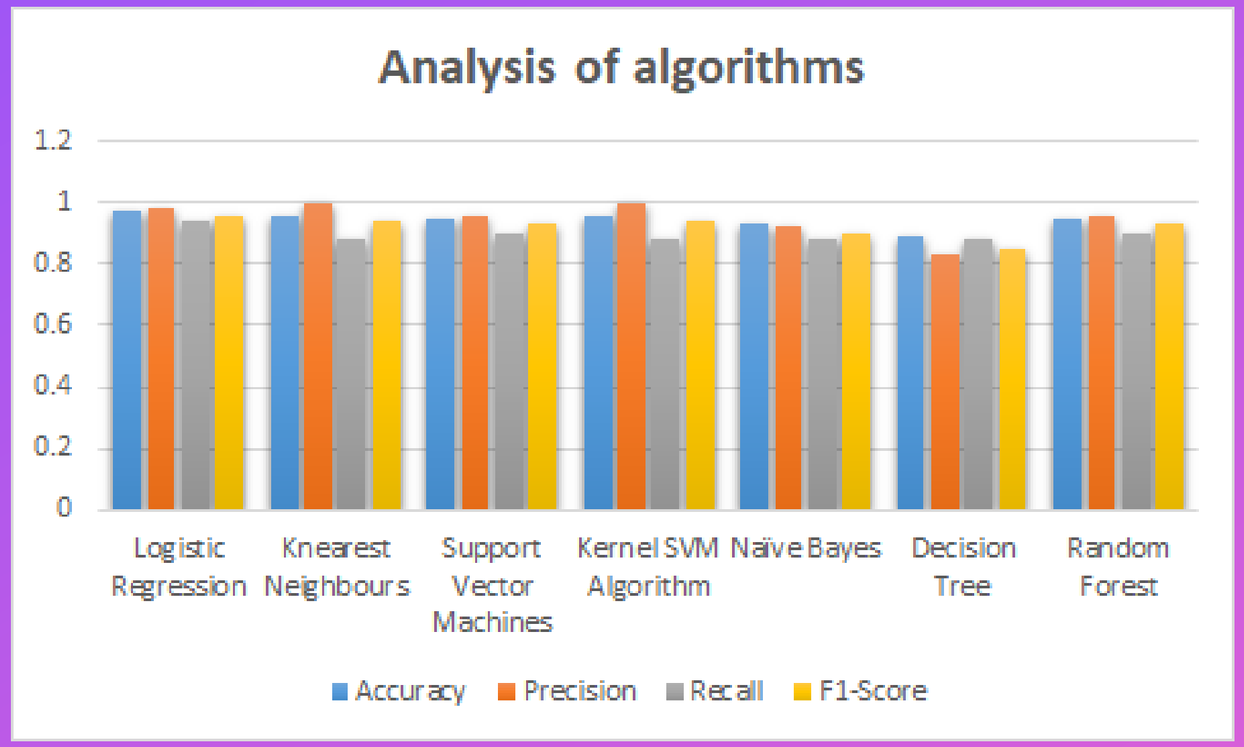
In [47]:
```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
print('\nAccuracy: {:.2f}'.format(accuracy_score(Y_test, Y_pred)))
print('Precision: {:.2f}'.format(precision_score(Y_test, Y_pred)))
print('Recall: {:.2f}'.format(recall_score(Y_test, Y_pred)))
print('F1-score: {:.2f}\n'.format(f1_score(Y_test, Y_pred)))

```

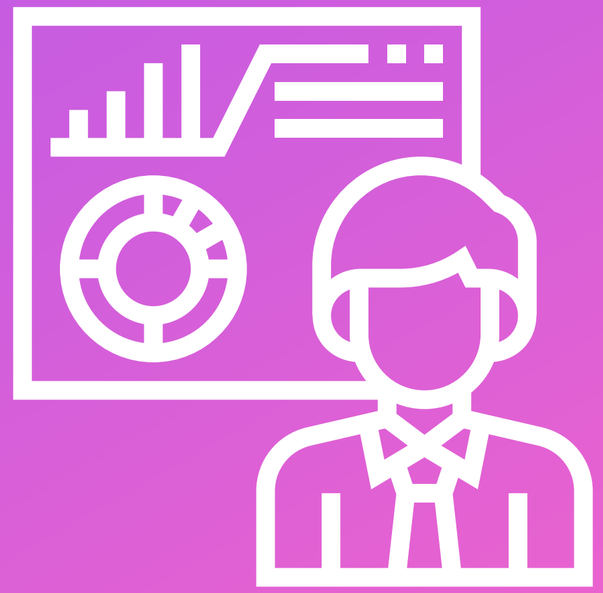
```
Accuracy: 0.97
Precision: 0.98
Recall: 0.94
F1-score: 0.96
```



Logistic regression,KNN,SVM,Random Forest ,Decision Tree,Naive Bayes

Analysing Results





Using various parameters of confusion matrix to analyse the algorithms on various parameters and selecting the most appropriate one

Project Plan and Timeline

CURRENT STATUS

Data exploratory analysis,choice of the machine learning model,choice of scoring completed

FURTHER WORK

Creating an easy to use UI and deploying to AWS

Reference research papers

PAPER 1

R. Potter, “Comparison of classification algorithms applied to breast cancer diagnosis and prognosis”, advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, pp. 40-49, (2007) July

PAPER 2

Van de stolpe, Anja.van.de.stolpe@philips.com & Hartskamp, Michael & Consoli, Sergio & Verhaegh, Wim & Petkovic, Milan. (2019). Artificial Intelligence in Clinical Health Care Applications: Viewpoint. Journal of Medical Internet Research. 21. e12100. 10.2196/12100.

PAPER 3

Ms. Manjiri Mahadev Mastoli1, Dr. Urmila R. Pol2, Rahul D. Patil3 Machine Learning ClassificationAlgorithms for Predictive Analysis in Healthcare International Research Journal of Engineering and Technology (IRJET)

PAPER 4

Ebru Aydındag Bayrak, Pınar Kırcı ,Tolga Ensari Comparison of Machine Learning Methods for Breast Cancer Diagnosis ©2019 IEEE

PAPER 5

Rani, K. & G.Naga Rama Devi, Dr & Doddipalli, Lavanya. (2015). Importance of Feature Extraction for Classification of Breast Cancer Datasets – A Study. International Journal of Scientific and Innovative Mathematical Research. 3. 763-768,Tirupati,India.

Reference blogs and videos

BLOG 1

https://www.rxdatascience.com/blog/machine-learning-and-healthcare-breast-cancer-diagnosis-

BLOG2

https://www.hindawi.com/journals/jhe/2019/4253641/

VIDEO 1

https://www.youtube.com/watch?v=QUT1VHiLmmI

VIDEO 2

https://www.youtube.com/playlist?list=PL-osiE80TeTvipOqomVEeZ1HRrcEvtZB_

VIDEO 3

https://www.youtube.com/watch?v=2WztaC6kyLs&list=PLS1QulWo1RIa7ha9SewcZlsTQVwL7n7oq

QUESTIONS?