Predict breast cancer with machine learning

Fan Yang

Project summary

- Breast cancer is the the second main cause of cancer death in women.
- Fine needle aspiration (FNA) is a common method of breast cancer exam that samples cells from a breast lesion with a fine needle.
- Breast cancer is diagnosed as either benign or malignant based on cell nuclei morphology.

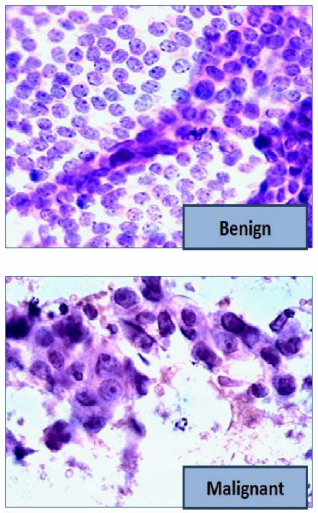


Image cited from Ahmad *et al.* 2013

- Dataset downloaded from Kaggle website contain cell nuclei features from 569 breast cancer patient.
- Goal of project is to build classifier model detecting malignant breast cancer.

Dataset overview

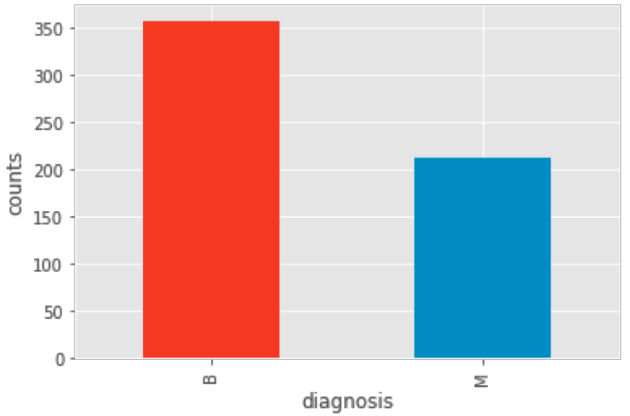
```
df.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

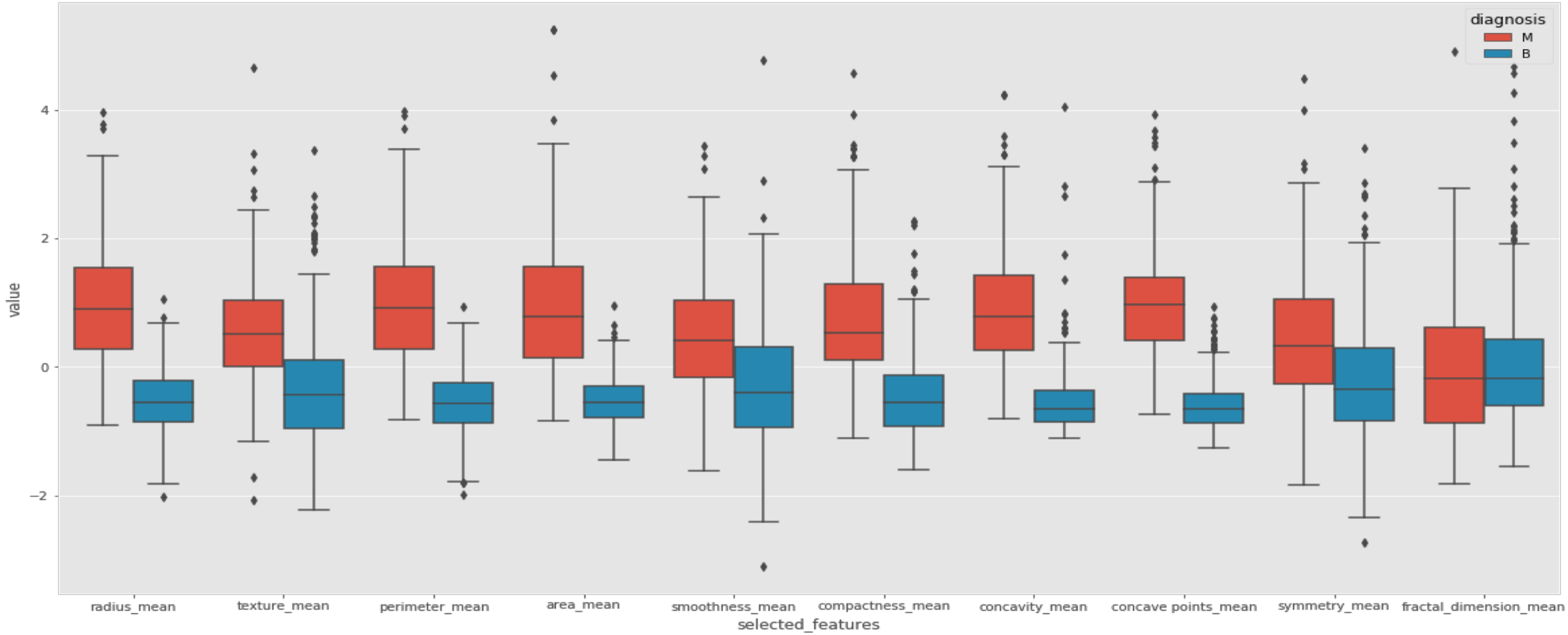
1) ID number
2) Diagnosis (M = malignant, B = benign)

Nuclei morphology is represented by following features:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

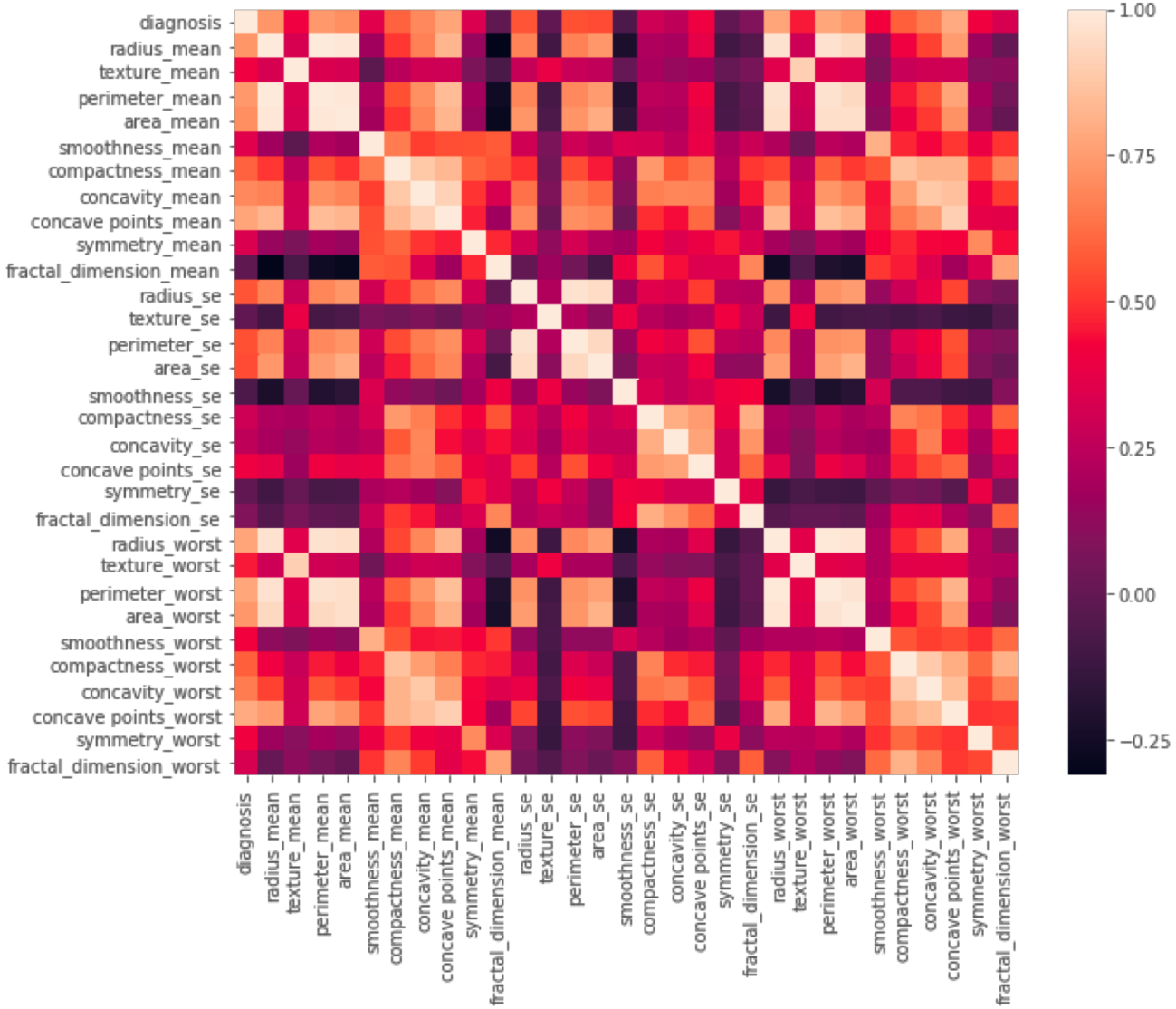


Differential distribution of cell nuclei features between benign and malignant samples



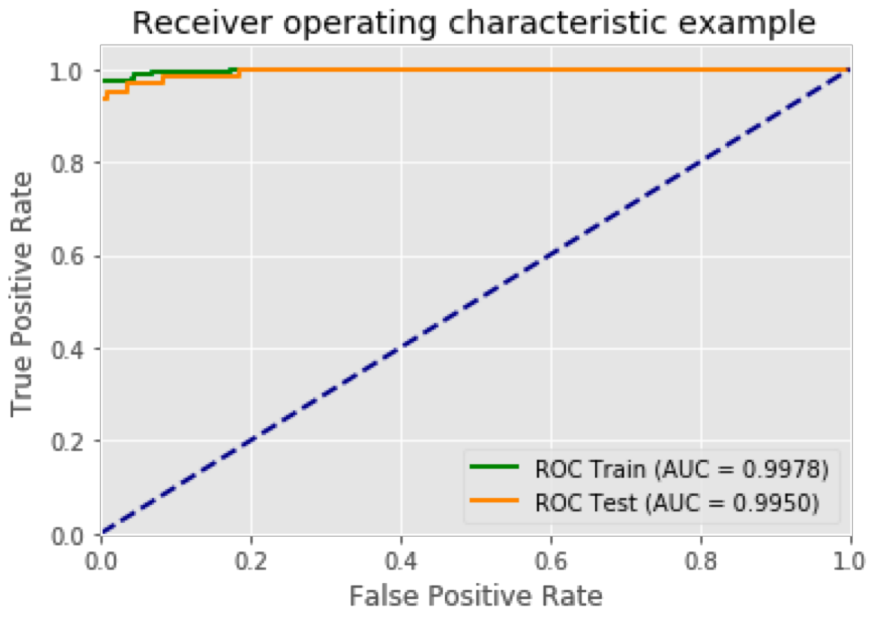
Relationship among numerical variables representing cell nuclei morphology

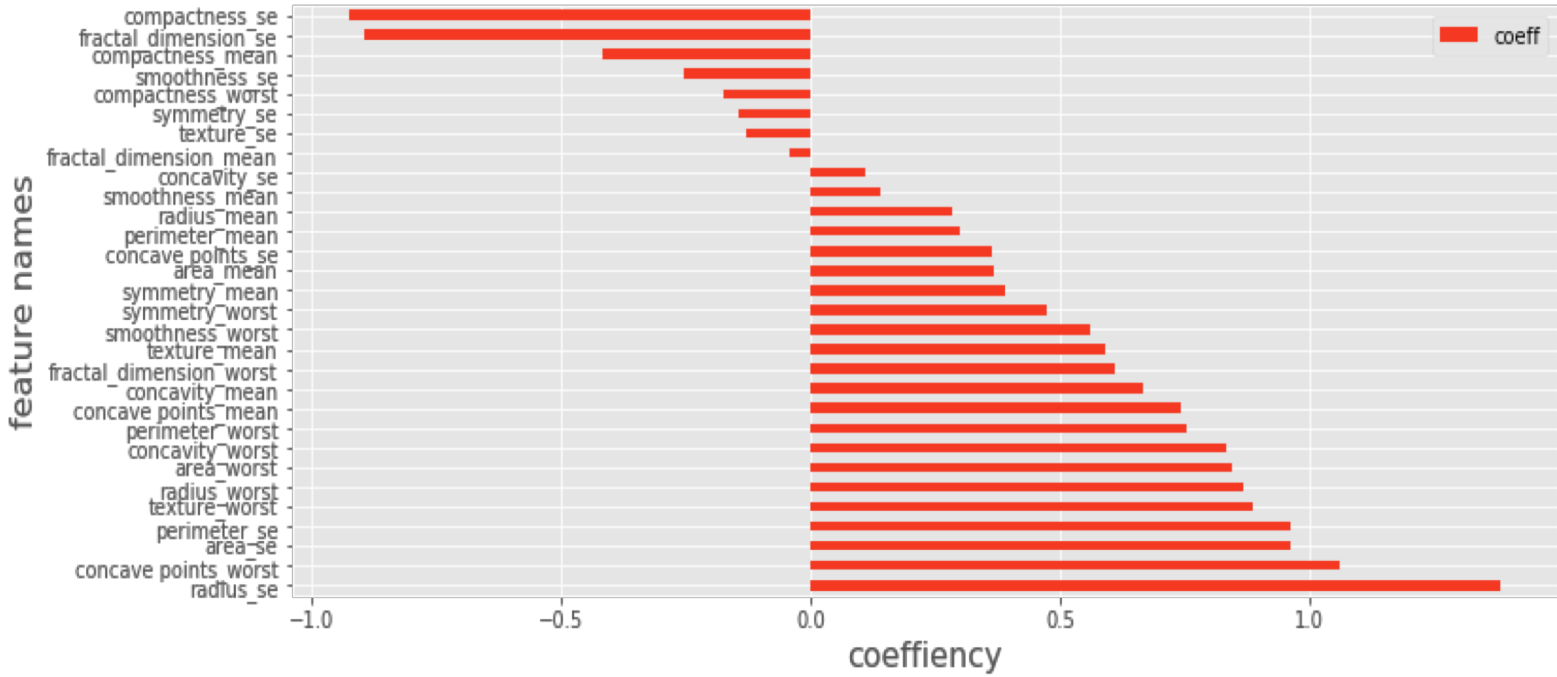
Correlation between breast cancer diagnosis and cell nuclei features



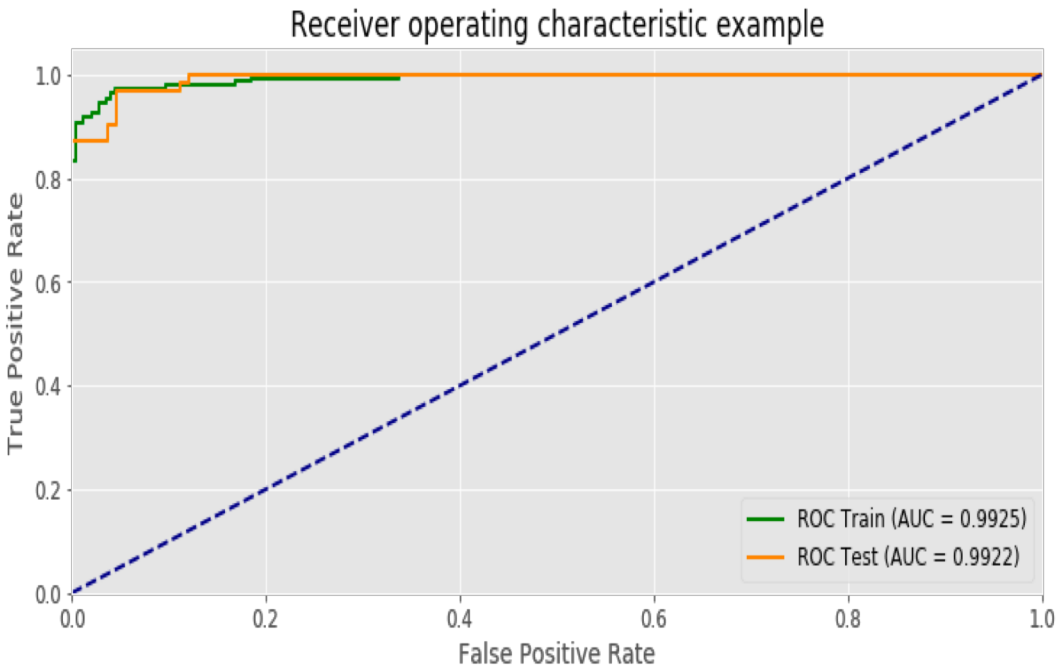
	index	corr_value
0	diagnosis	1.000000
1	concave points_worst	0.793566
2	perimeter_worst	0.782914
3	concave points_mean	0.776614
4	radius_worst	0.776454
5	perimeter_mean	0.742636
6	area_worst	0.733825
7	radius_mean	0.730029
8	area_mean	0.708984
9	concavity_mean	0.696360
10	concavity_worst	0.659610
11	compactness_mean	0.596534
12	compactness_worst	0.590998
13	radius_se	0.567134
14	perimeter_se	0.556141
15	area_se	0.548236
16	texture_worst	0.456903
17	smoothness_worst	0.421465
18	symmetry_worst	0.416294
19	texture_mean	0.415185
20	concave points_se	0.408042
21	smoothness_mean	0.358560
22	symmetry_mean	0.330499
23	fractal_dimension_worst	0.323872
24	compactness_se	0.292999
25	concavity_se	0.253730
26	fractal_dimension_se	0.077972
27	symmetry_se	-0.006522
28	texture_se	-0.008303
29	fractal_dimension_mean	-0.012838
30	smoothness_se	-0.067016

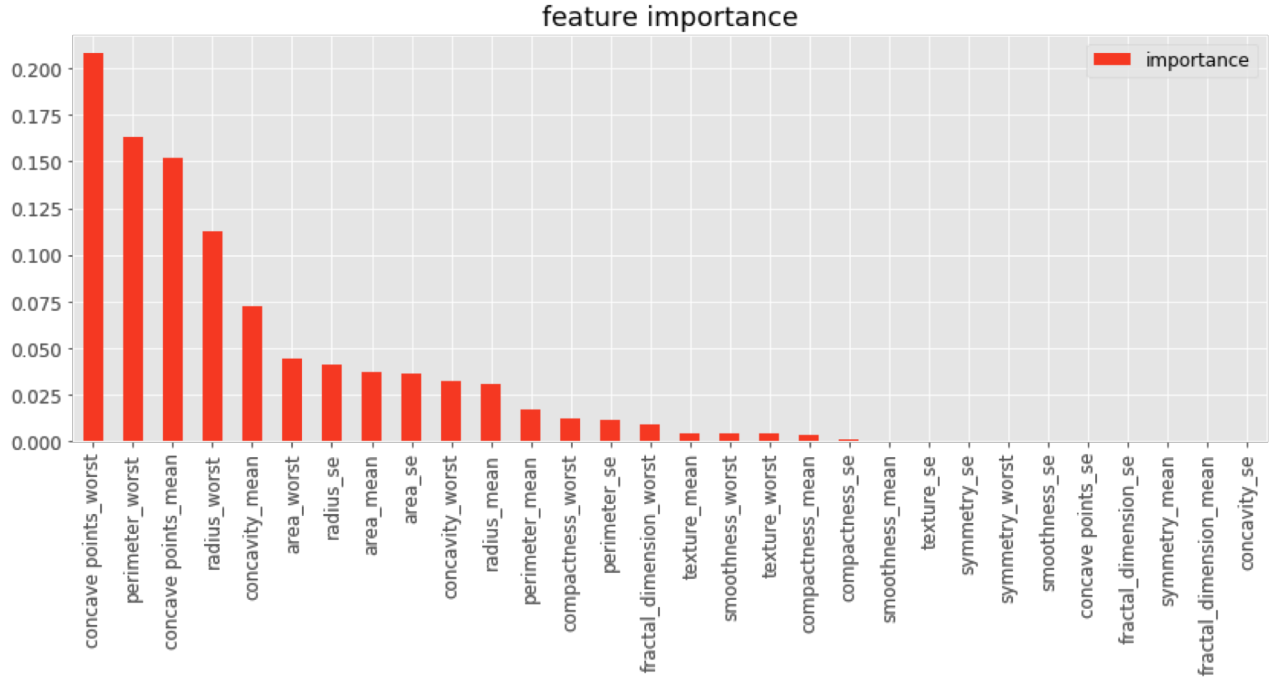
Build Logistic regression model detecting malignant breast cancer





Build Random forest model detecting malignant breast cancer





Project conclusions

1. Morphological features of cell nuclei are criteria for breast cancer diagnosis.
2. This project explores how these features correlate to each other and how they cause a differential distribution between benign or malignant cancer samples
   .
3. Logistic regress and random forest models were built to detecting malignant breast cancer
4. Features importance reveal that concave points and radium are the most useful information for image-based breast cancer diagonosis.