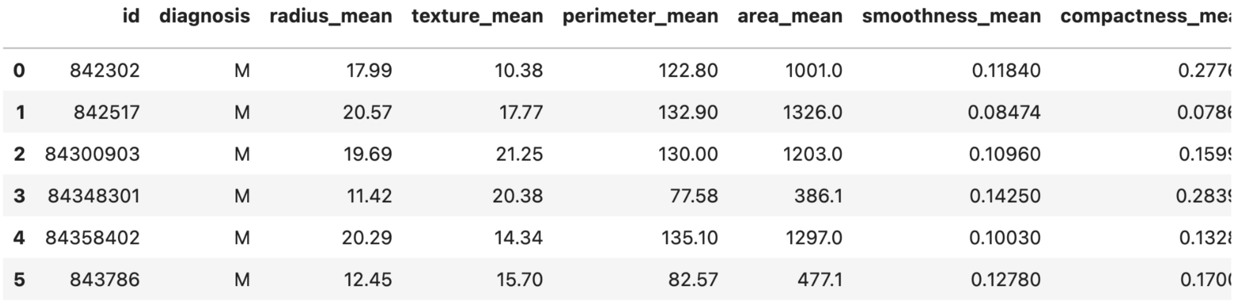
Introduction

Analysis of cell nuclei characteristics from Fine Needle Aspirate images (FNA) is one method used to diagnose cancer. By analysis of a digitized dataset of breast cancer cell FNA images provided by the University of Wisconsin, Clinical Sciences Center, the report aims to develop a machine learning model to classify and predict if the tumor is benign or malignant and the possibility of breast cancer development. After comparing the accuracy and confusion matrix of five models, which are logistic regression, SVM, decision tree, and random forest with and without feature selections, the logistic regression model with an accuracy of 94% is the best fit for breast cancer classification from numeric features of FNA images.

Data

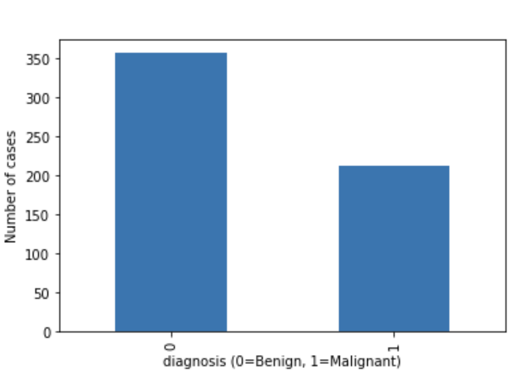
The breast cancer dataset was obtained from Kaggle and was initially provided by the University of Wisconsin Clinical Sciences Center. The data set provides 32 features in text or numeric measurements from FNA image of breast cancer cell. Features includes case id, diagnosis, radius of cell, texture, area, etc.

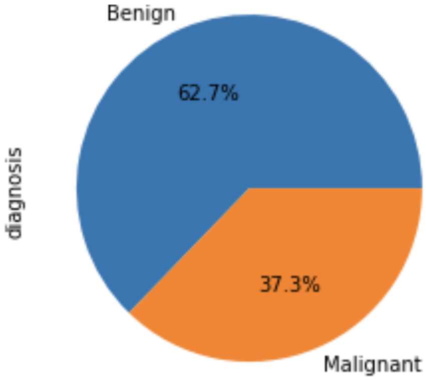


	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_me
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.277
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.078
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.159
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.283
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.132
5	843786	M	12.45	15.70	82.57	477.1	0.12780	0.170

Methodology

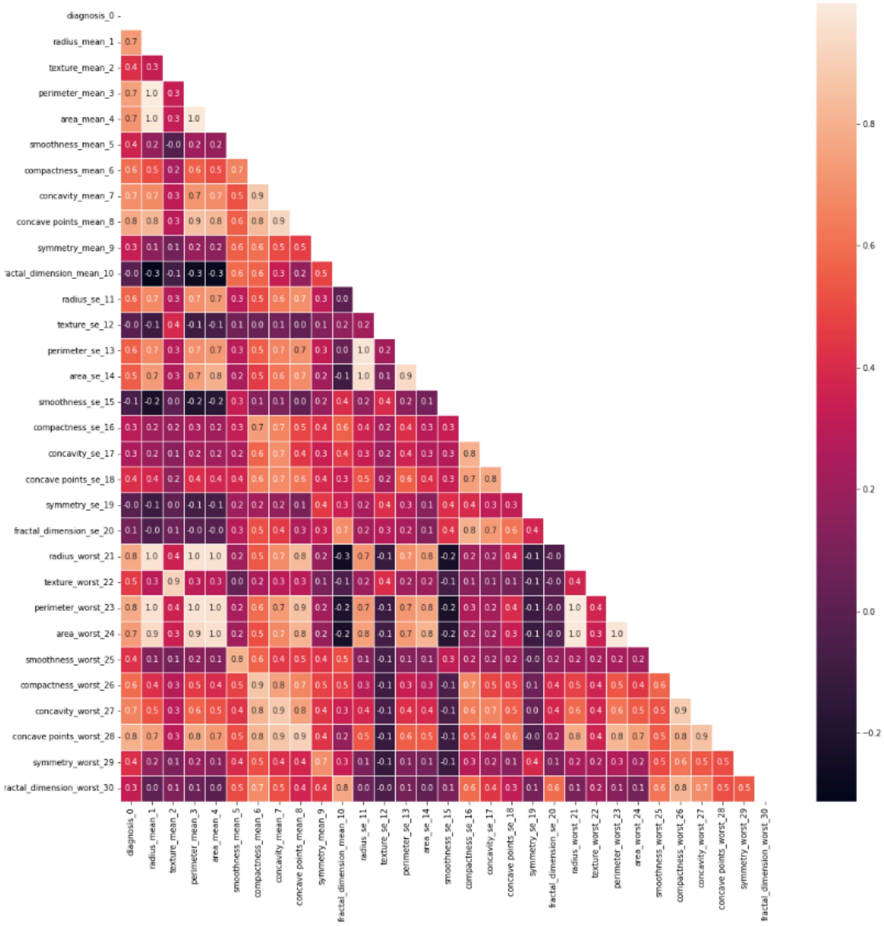
There are 569 rows (cases) and 33 columns (features), which 32 features are numerical and one is in text. Original diagnosis feature is present in text, as M refers to Malignant and B refers to Benign. For analysis purpose, the diagnosis feature is converted to numbers with one refers to malignant and zero refers to benign. The feature “Unnamed: 32” with null values were dropped.
The total of 569 cases are consist of 212 cases of Malignana and 357 cases of Benign, which are 62.7% and 37.3% of total cases, relatively.





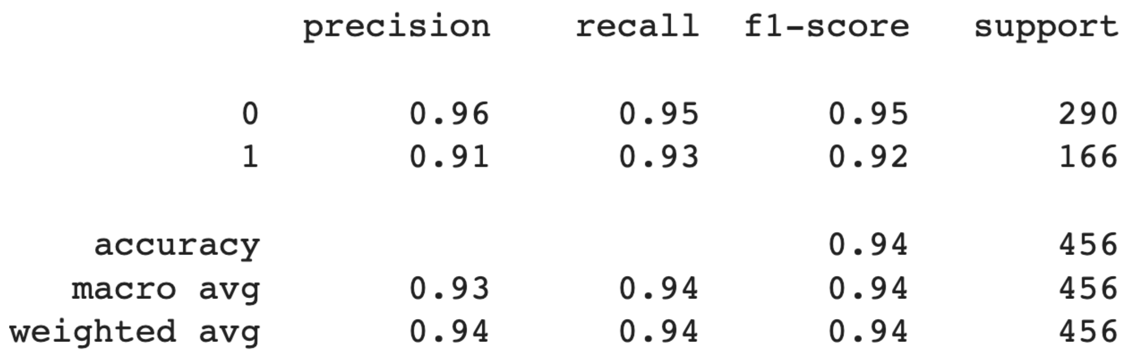
After data exploration and clean up, 70% data was split into training set, and 30% data was split into testing set. Five different tunned models were tested and compared with accuray and confusion matrix. Randon search cross validation is used to validate the model fitting.

From the correlation heatmap, some features are strongly correlated to digagnosis result, such as "texture_mean", "fractal_dimension_mean", "texture_se", "smoothness_se", "symmetry_se"...etc.



The feature importance were visualized as the below graph, and top six features, witch are “concave points_mean”, “concave points_worst”, “perimeter_worst”,”area_worst”, “concavity_mean”, “radius_worst”, were selected for random forest model.

Among the five models, which are logistic regression, SVM, decision tree, and random forest with and without feature selections, the logistic performed the best, which able to predict and classify tumer breast cancer with an accuracy of 94%.



```
              precision    recall  f1-score   support

           0       0.96      0.95      0.95       290
           1       0.91      0.93      0.92       166

    accuracy                           0.94       456
   macro avg       0.93      0.94      0.94       456
weighted avg       0.94      0.94      0.94       456
```

Conclusion

The purpose for the project aims to developing a machine learning model to classify the breast tumor to benign or malignant based on numeric features from FNA images. After processing and clean up raw data, five different machine models were tuned and tested on the dataset. The model logistic regression, which has an accuracy of 94%, is the best fit model to predict and classify tumer as malignant or begine. However, the model is limited by the amount of data present in the dataset, which only consist of 569 cases. A larger collection of data may be needed to further validate the model.