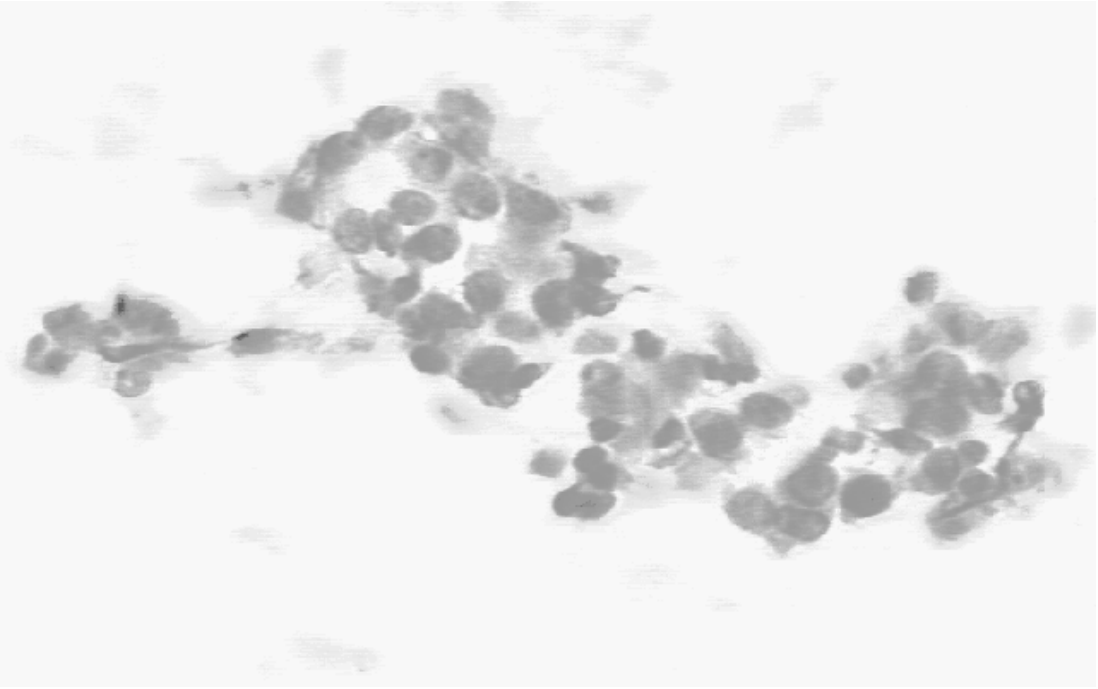
BREAST CANCER DIAGNOSIS

Group 6:

Feng Qingduo, Shen Zhentao, Sun Wenjia, Lin Jialiang

CATALOG

- Goal & Dataset
- Our Progress
- Major Issues
- Corresponding Methods



GOAL & DATASET

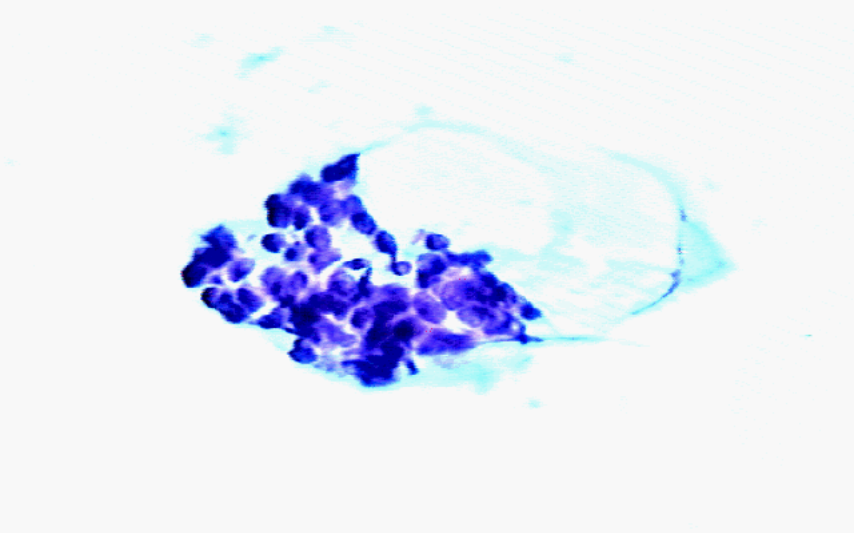
- Classification Task:
  - determine whether a new record data is related with benign breast cancer or malignant breast cancer
- Dataset: Wisconsin Diagnostic Breast Cancer
  - From UCI Machine Learning Repository (UCI-MLR)
  - Can also be found in Kaggle
- 32 Attributes: ID, diagnosis, and 30 real-valued input attributes
  - Diagnosis = {benign, malignant}
  - 10 features obtained from digitized image of FNA (fine needle aspirate) of a breast mass.

1)radius
2)texture
3)perimeter
4)area
5)smoothness
6)compactness
7)concavity
8)concave points
9)symmetry
10)fractal dimension
and so on …

MINI-PROJECT PLAN:

- Stage 1 [Classification Task]
  - data exploration & data preprocessing
  - divide training-set & test test
  - feature selection
  - build classifier
  - analyze the result
- Stage 2 [Image-Processing]
  - extract features from original images
  - image process to get 30 features in WDBC
  - explore more features
  - …

	id ID number	diagnosis The diagnosis of breast tissues (M = malignant, B = benign)	radius_mean mean of distances from center to points on the perimeter
	8670 … 911m	B 63% M 37%	6.98 … 28.1
1	842302	M	17.99
2	842517	M	20.57
3	84300903	M	19.69
4	84348301	M	11.42



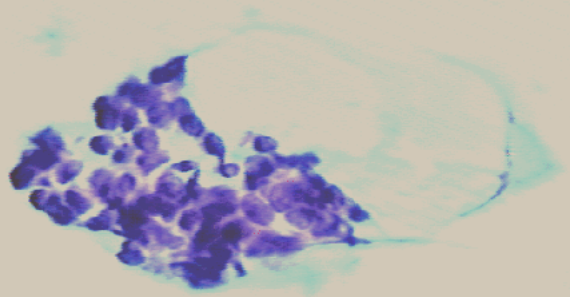
OUR PROGRESS: STAGE 1

- Stage 1
  - data exploration & data preprocessing
  - divide training-set & test test
  - feature selection
  - build classifier
  - analyze the result

	id ID number	gnosis The d… of breast tissues… alignant, B = benign	radius_mean mean of distances from center to points on the perimeter
	8670 … 911m	B 63% M 37%	6.98 … 28.1
1	842302	M	17.99
2	842517	M	20.57
3	84300903	M	19.69
4	84348301	M	11.42

- Stage 2
  - extract features from original images
  - image process to get 30 features in WDBC
  - explore more features
  - …

work on it



OUR PROGRESS: STAGE 1

DATA EXPLORATION & DATA PREPROCESSING

DIVIDE TRAINING-SET & TEST TEST

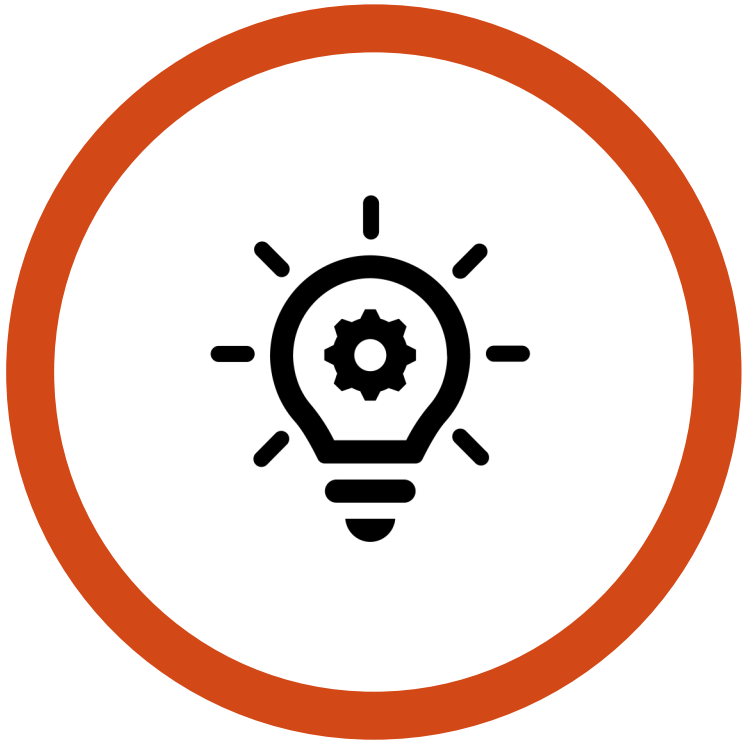
FEATURE ENGINEERING

BUILD CLASSIFIER

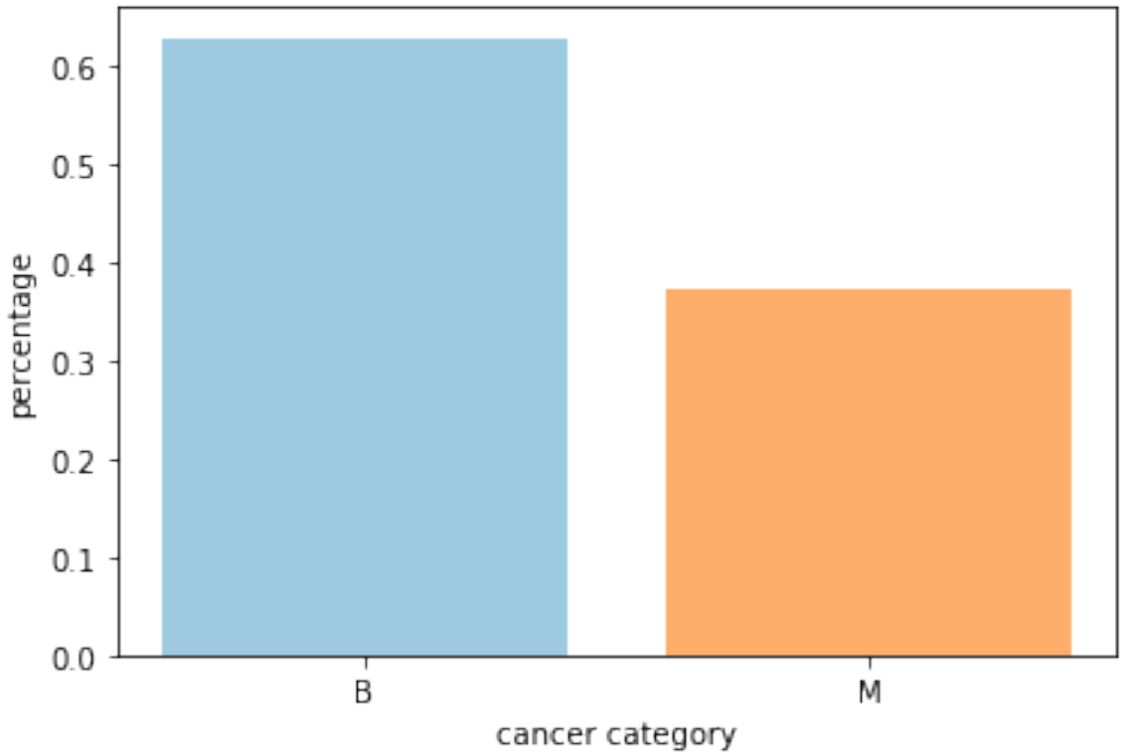
ANALYZE THE RESULT

MAJOR ISSUES:

- imbalanced class distribution
- feature importance
- feature redundancy

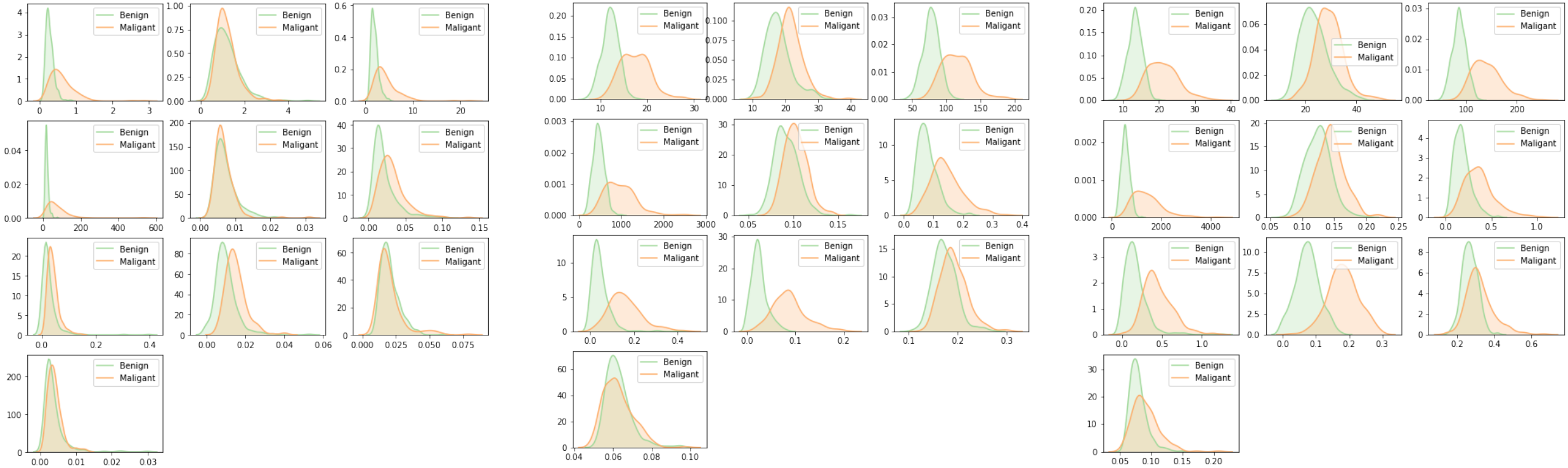


1. IMBALANCED DATASET

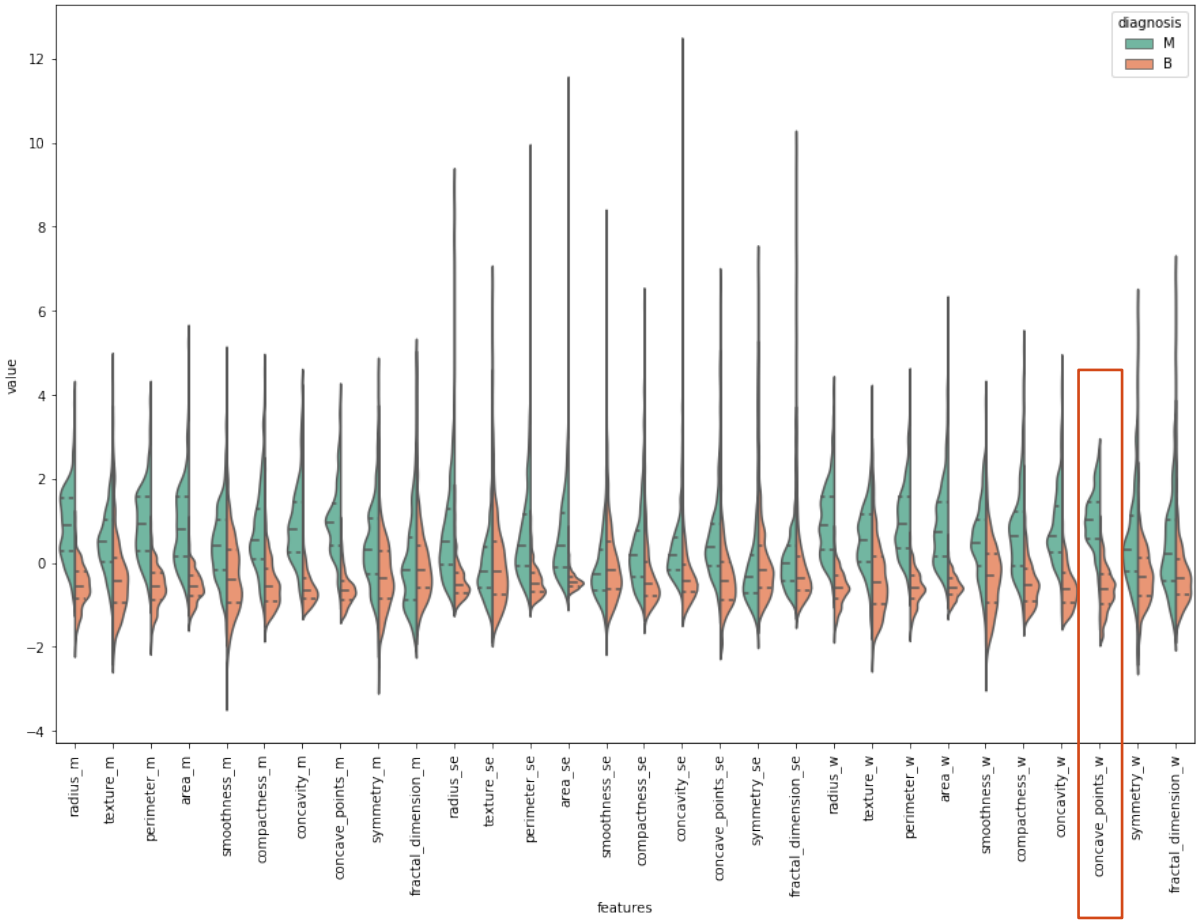


- Notice that it is an imbalanced class distribution.
- We need to handle this problem by using over- and under-sampling, or it may affect the performance of the model.
- Later we will build models under different sampling methods to compare their performance.

2. FEATURE IMPORTANCE



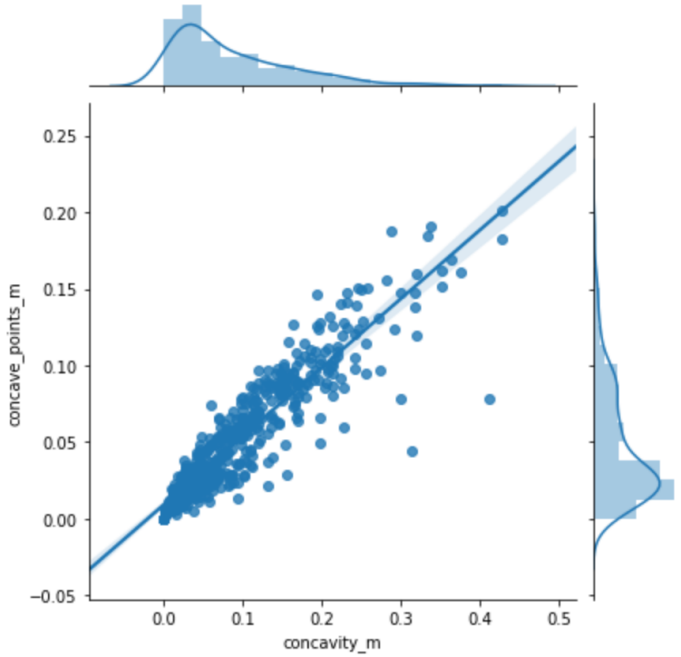
2. FEATURE IMPORTANCE

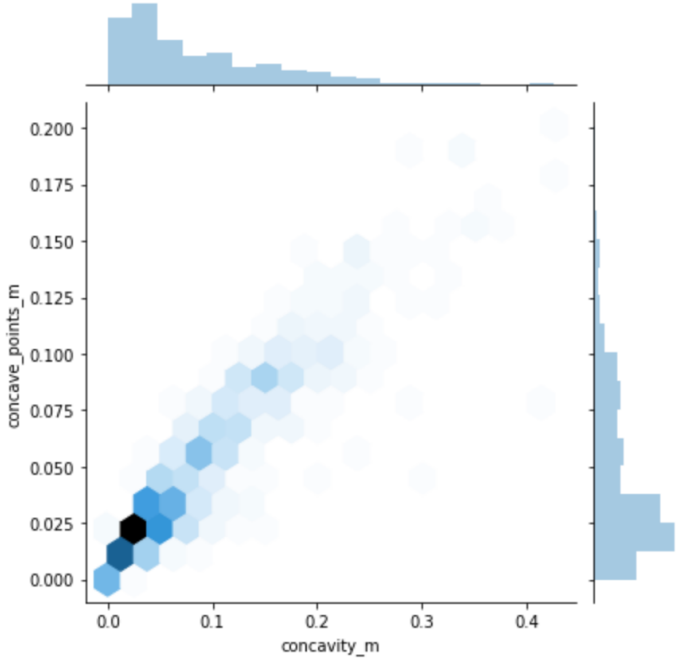


Feature Distribution

- To see the distribution of feature values in two different classes.
- Some attributes have good distinctive performance, like "concave_points_w", while others do not.
- So different features may have different importance in terms of classification.

3. FEATURE REDUNDANCY

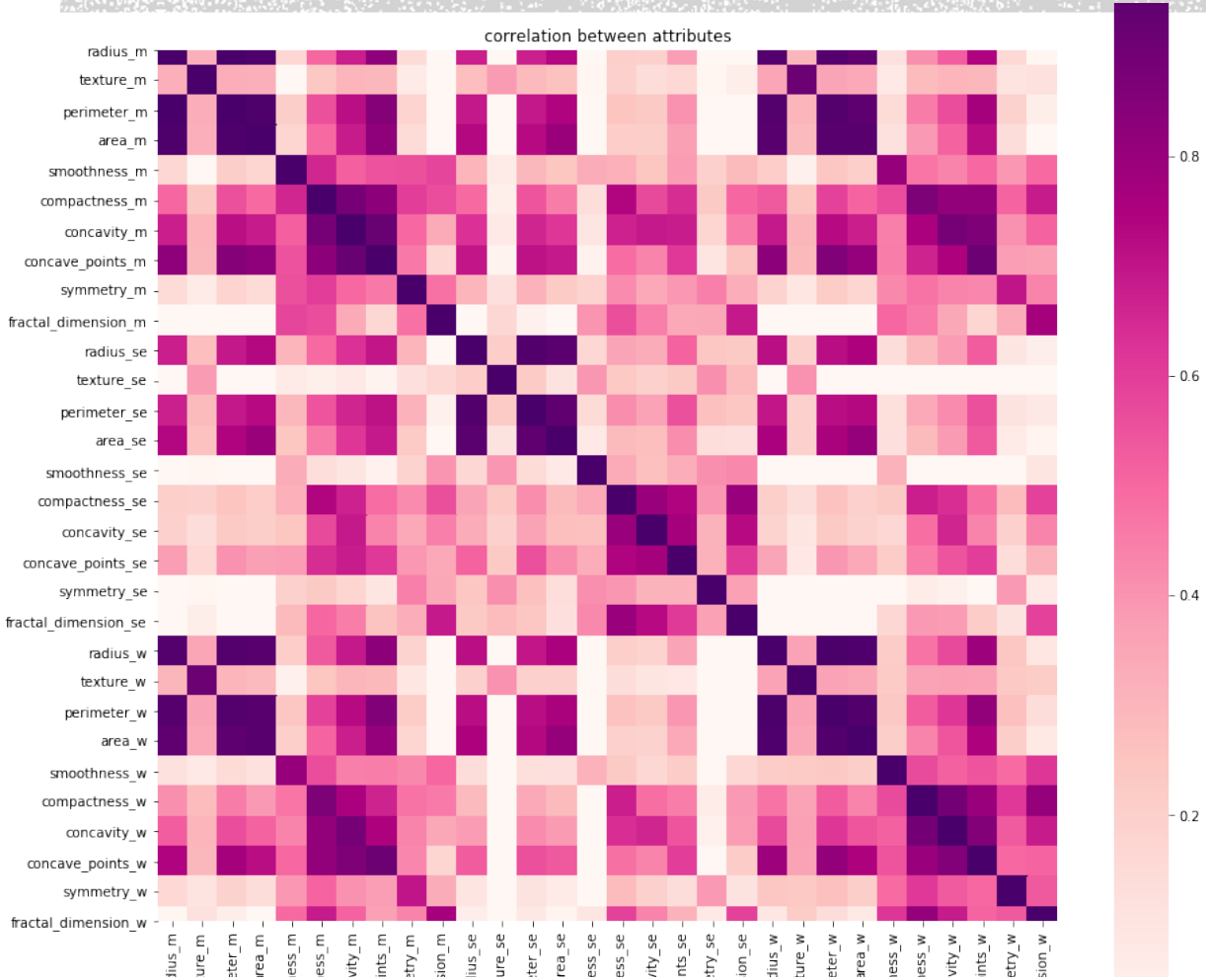




Relation between Features

- Compute the correlation to see whether some of them are very relevant.
- There are some attributes with high relations like the 'concavity_m' *and* 'concave_points_m'.
- Later we will utilize PCA (Principal Component Analysis), an unsupervised-learning method, to see whether we can reduce the number of attributes into a few features that have captured the most info.

3. FEATURE REDUNDANCY

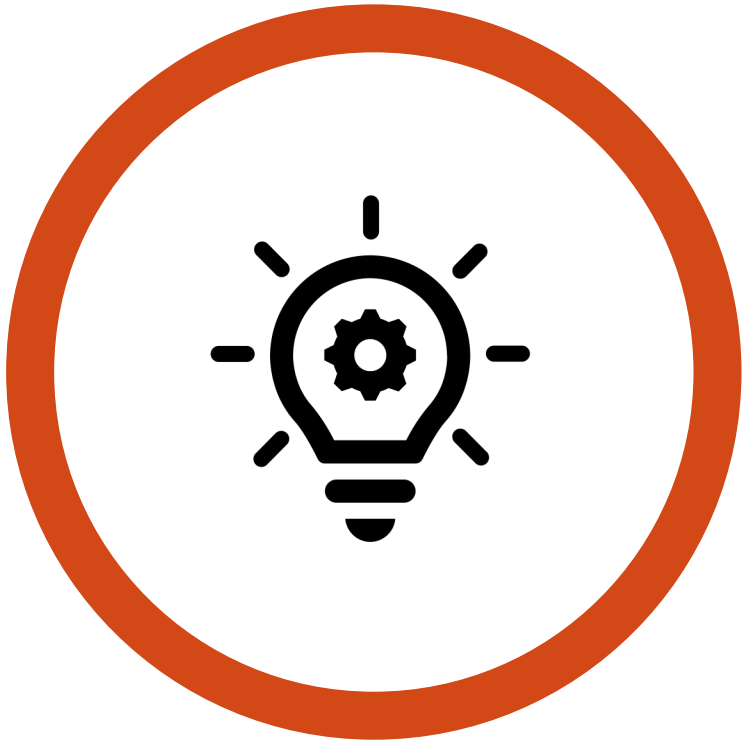


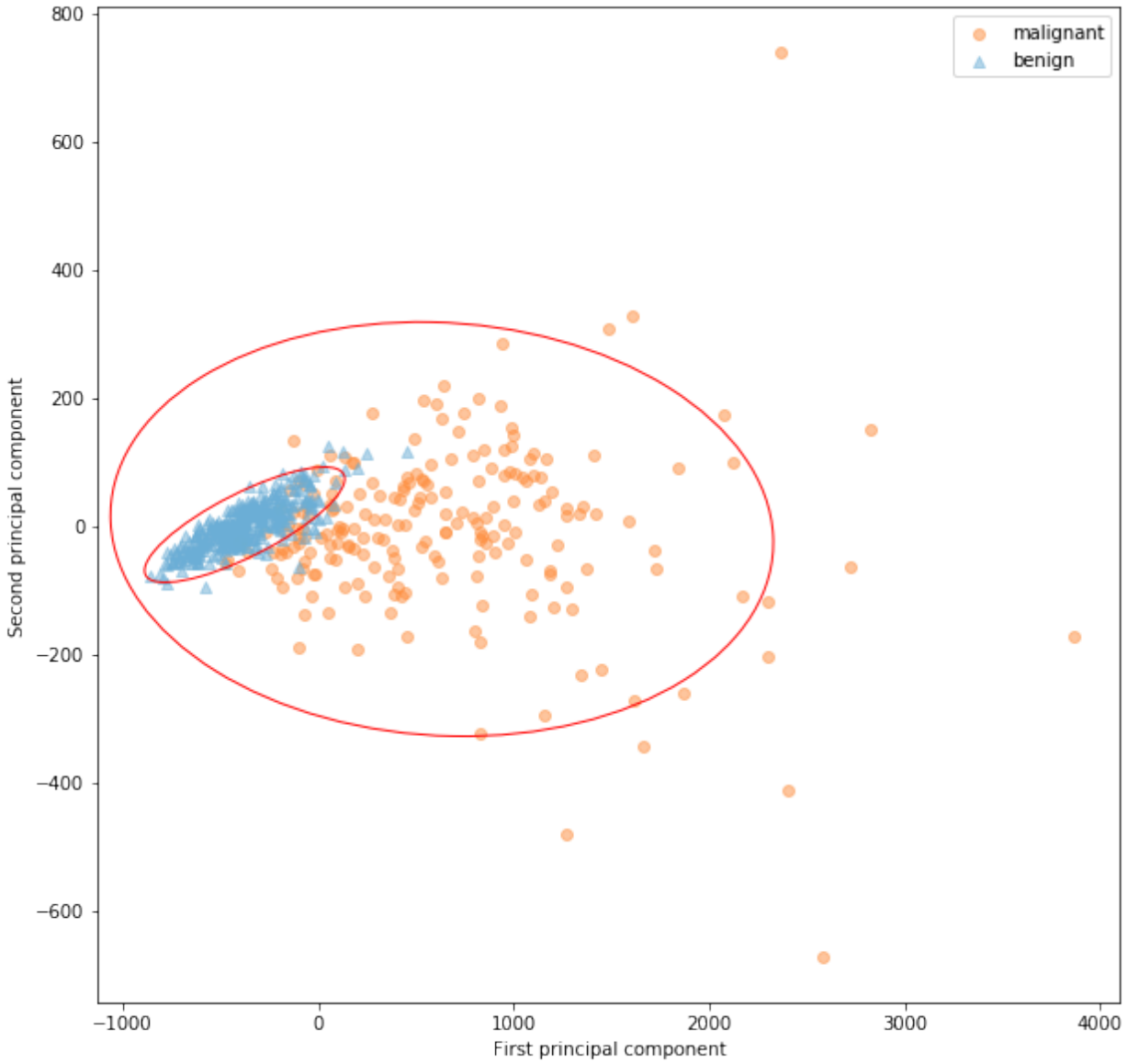
Relation between Features

- Compute the correlation matrix to see the correlation between pairwise features.
- Notice that there are some attributes with high relations like perimeter and area.
- So there may be redundancy within the features, which we can try to solve.

MAJOR ISSUES & CORRESPONDING METHODS

- feature redundancy
  - PCA method
- feature importance
  - information theory
- imbalanced class distribution
  - sampling method





1. SOLVE FEATURE REDUNDANCY

Reduce the dimension into 2-D with PCA method

Here is the data visualization diagram.

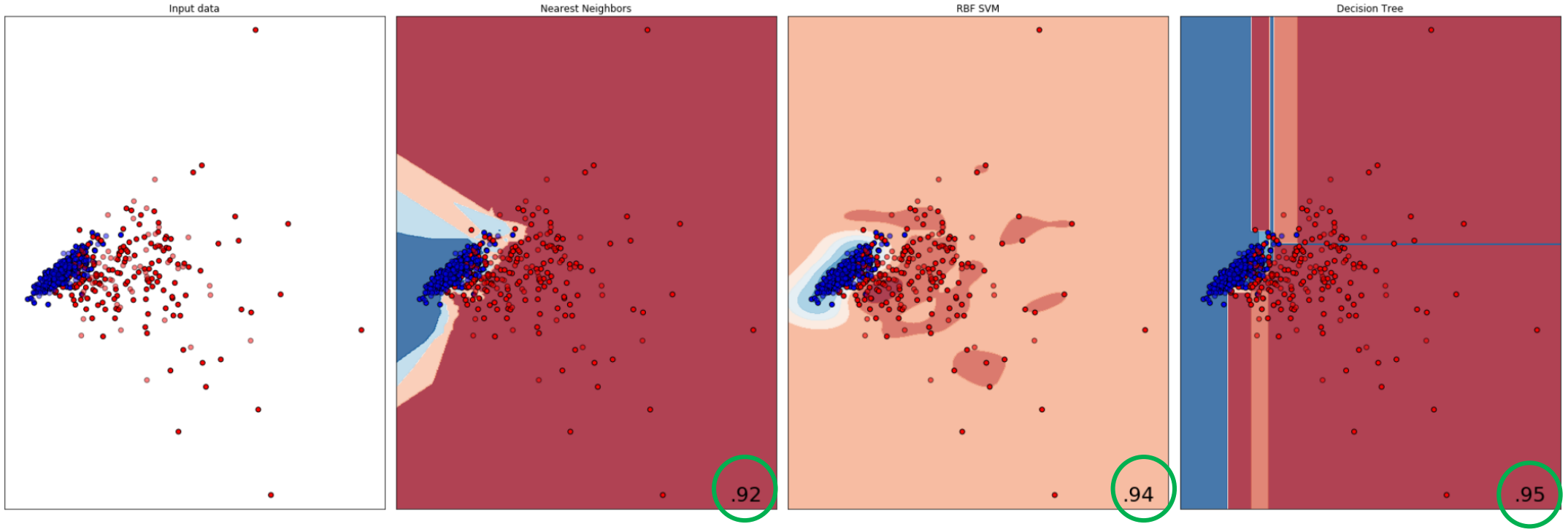
we can see how well these the two principal components work and the corresponding confidence.

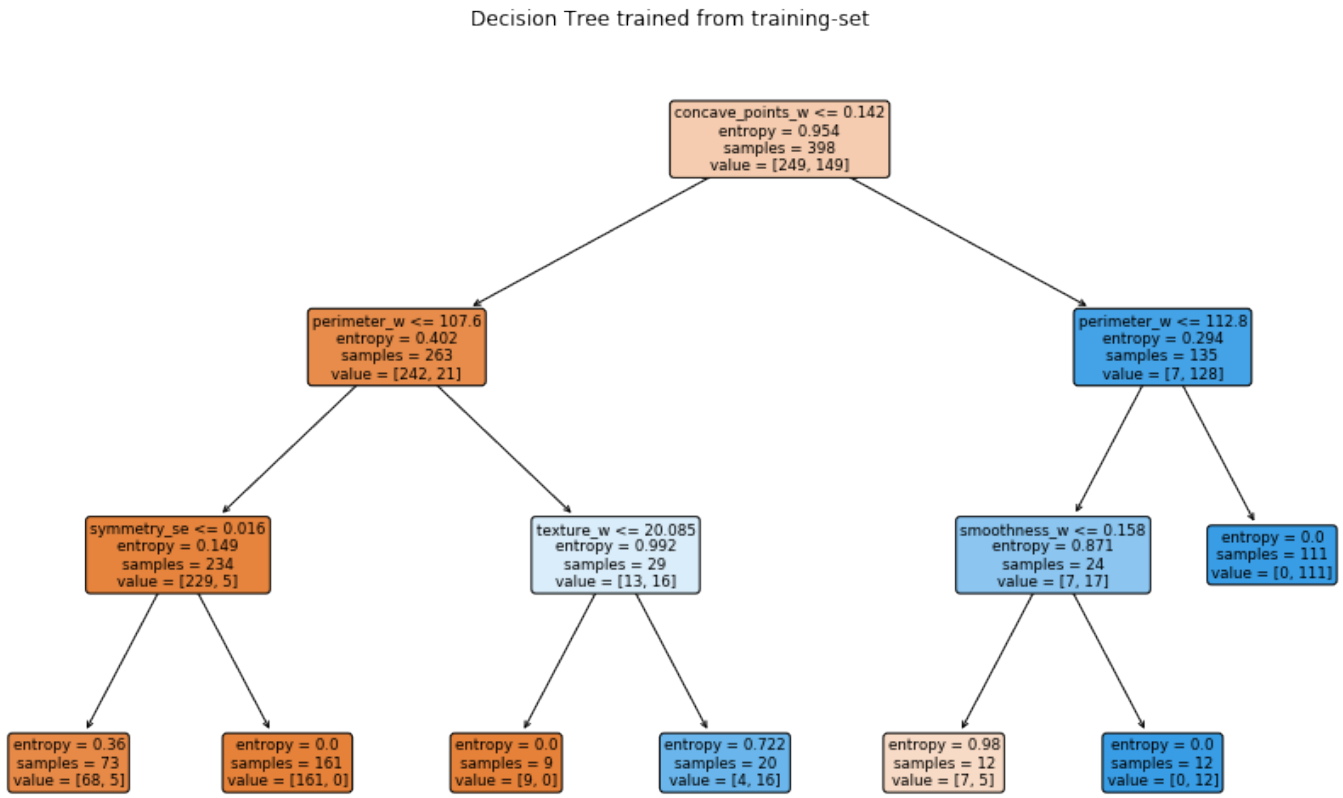
From the diagram with confidence ellipse (95%), we can see that the benign cases have a compact distribution while the malignant ones appear seperately.

PCA METHOD

Construct Model after PCA

- K Nearest Neighbors
- Support Vector Machine
- Decision Tree



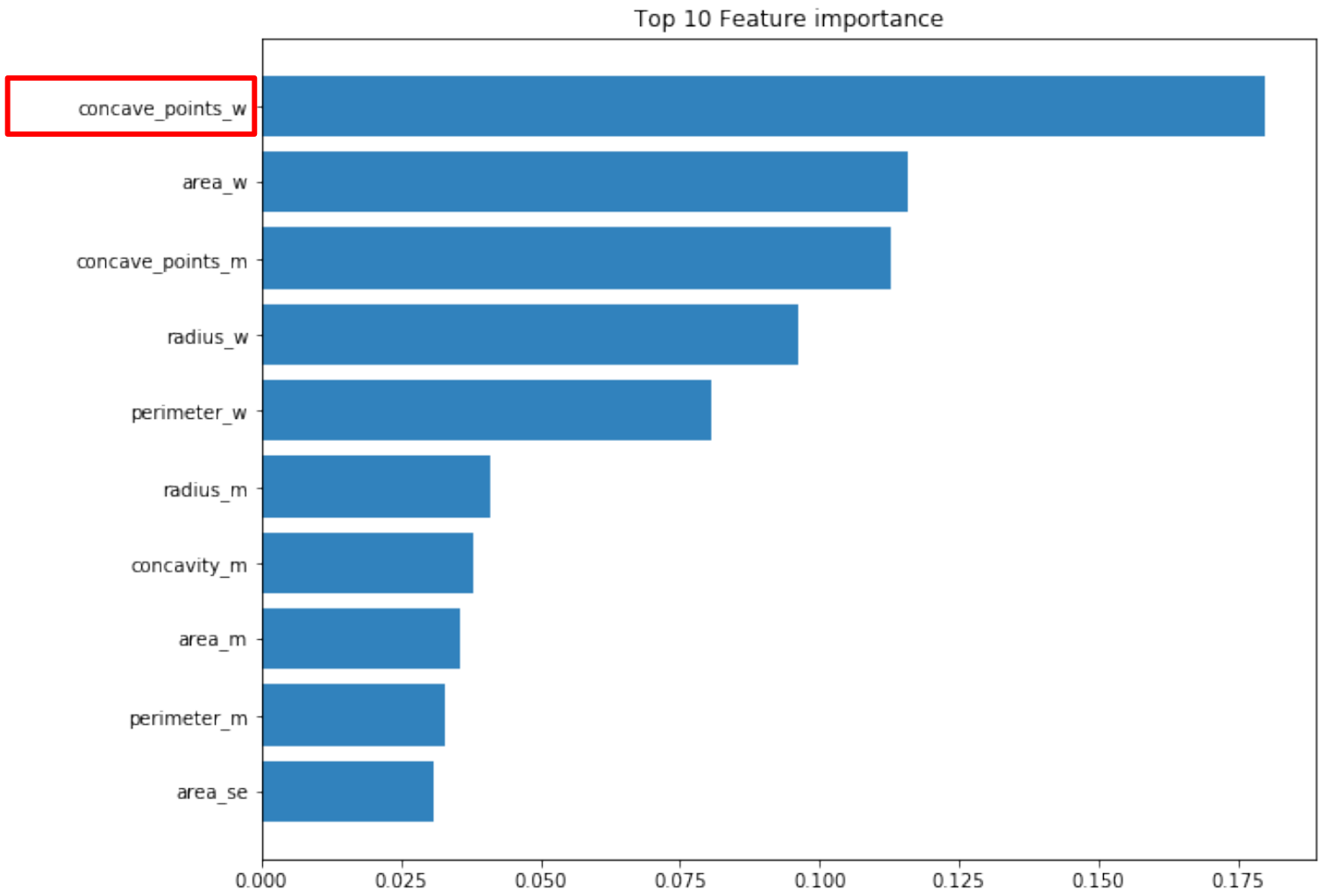


2. BUILD TREE-TYPE CLASSIFIER

Construct Tree-Type Model

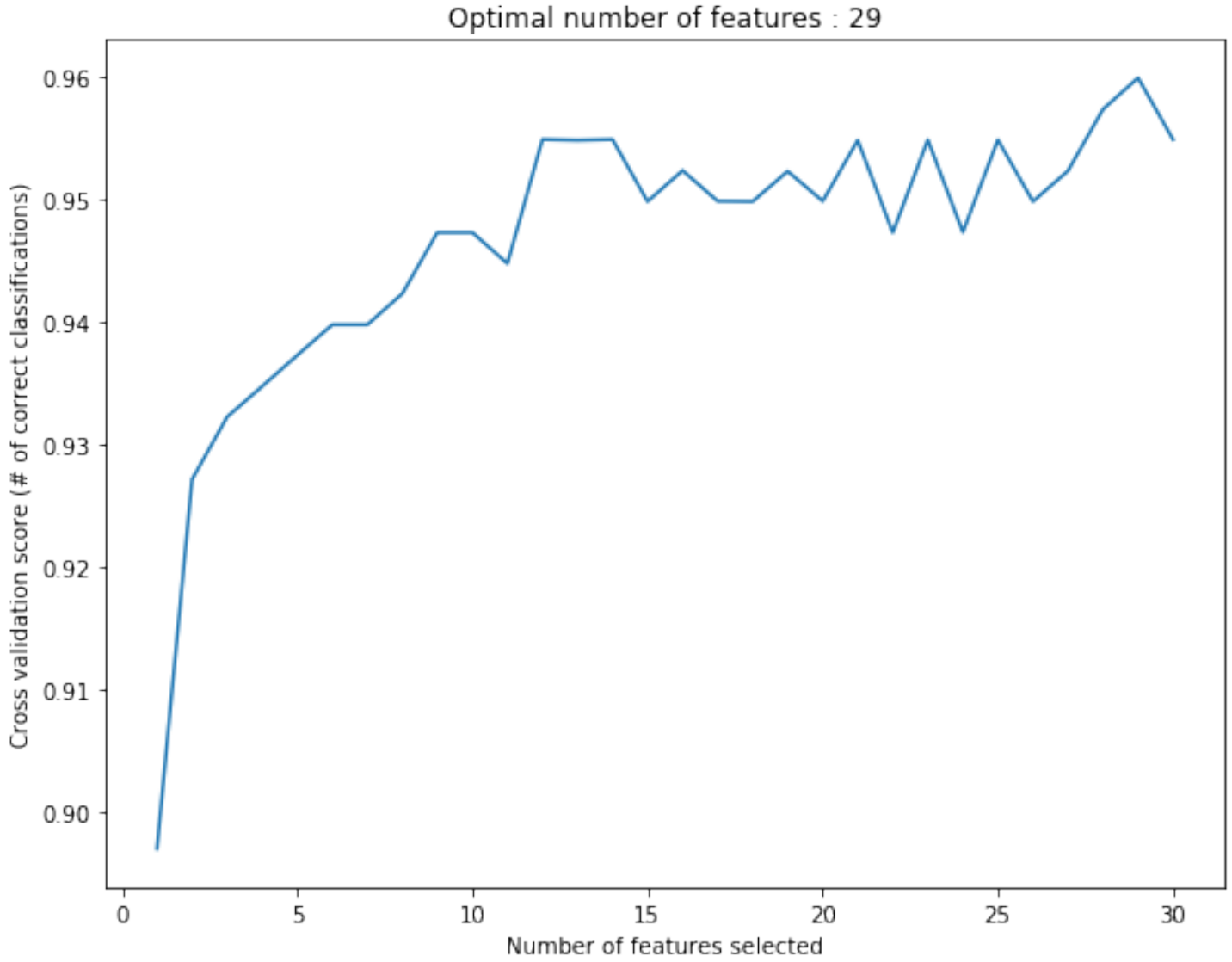
- Decision Tree
- Random Forest

- Form the tree plot diagram, we can see that most data records can be successfully classified into one category



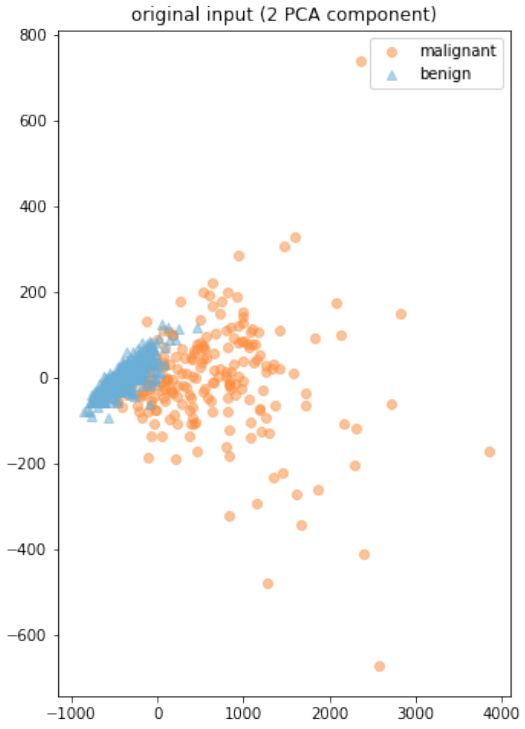
FEATURE IMPORTANCE

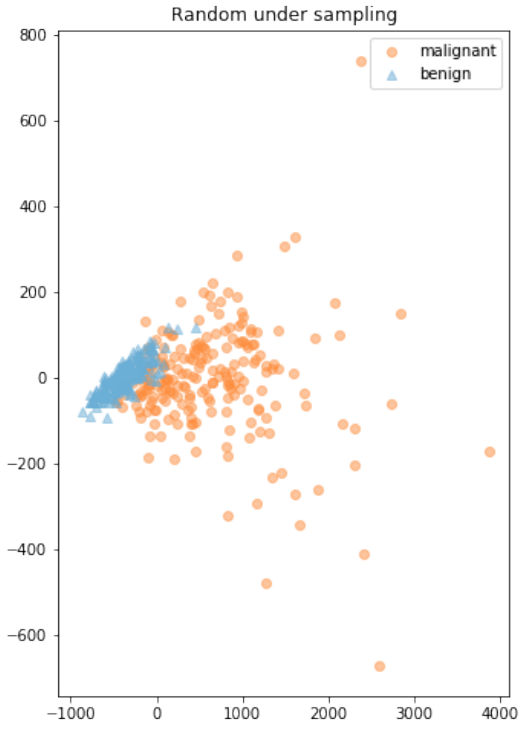
- Information theory:
- i.e. entropy, gini index …
- “concanve_points_w” are the most important feature in Random Forest with 100 trees
- it is consistent with our observation in violin plot of feature distributions

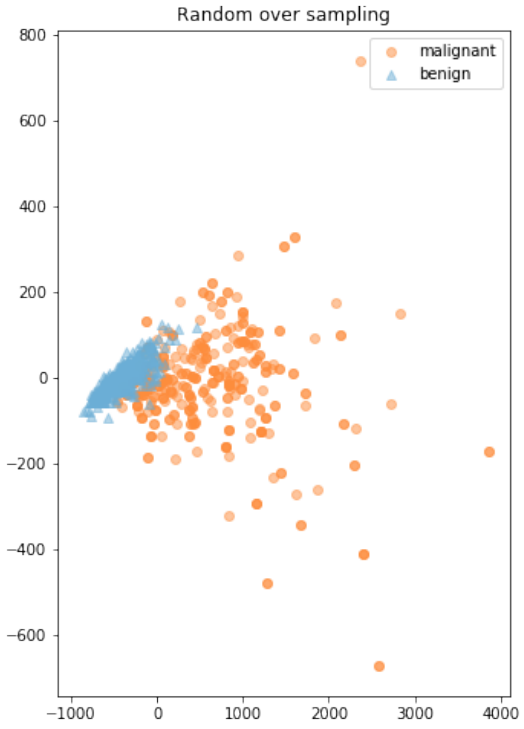


FEATURE IMPORTANCE

- The Recursive Feature Elimination (RFE) method is a feature selection approach.
- It uses the model accuracy to identify which attributes and the combination of attributes are most important.
- Determine the most suitable number of features that we should select.





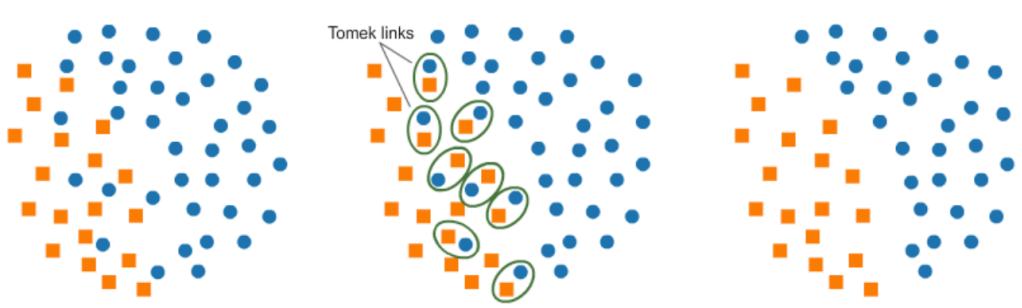


3. OVERCOME IMBALANCED DATA

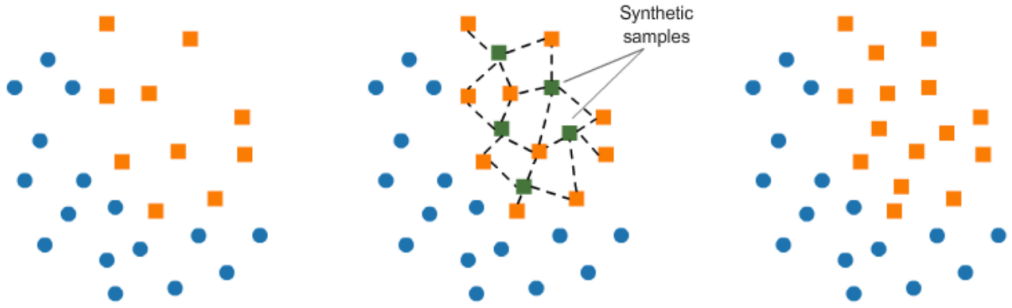
Sampling method:

- random under-sampling
- random over-sampling

- Tomek-links
- SMOTE
- SMOTE + Tomek-links



Tomek links

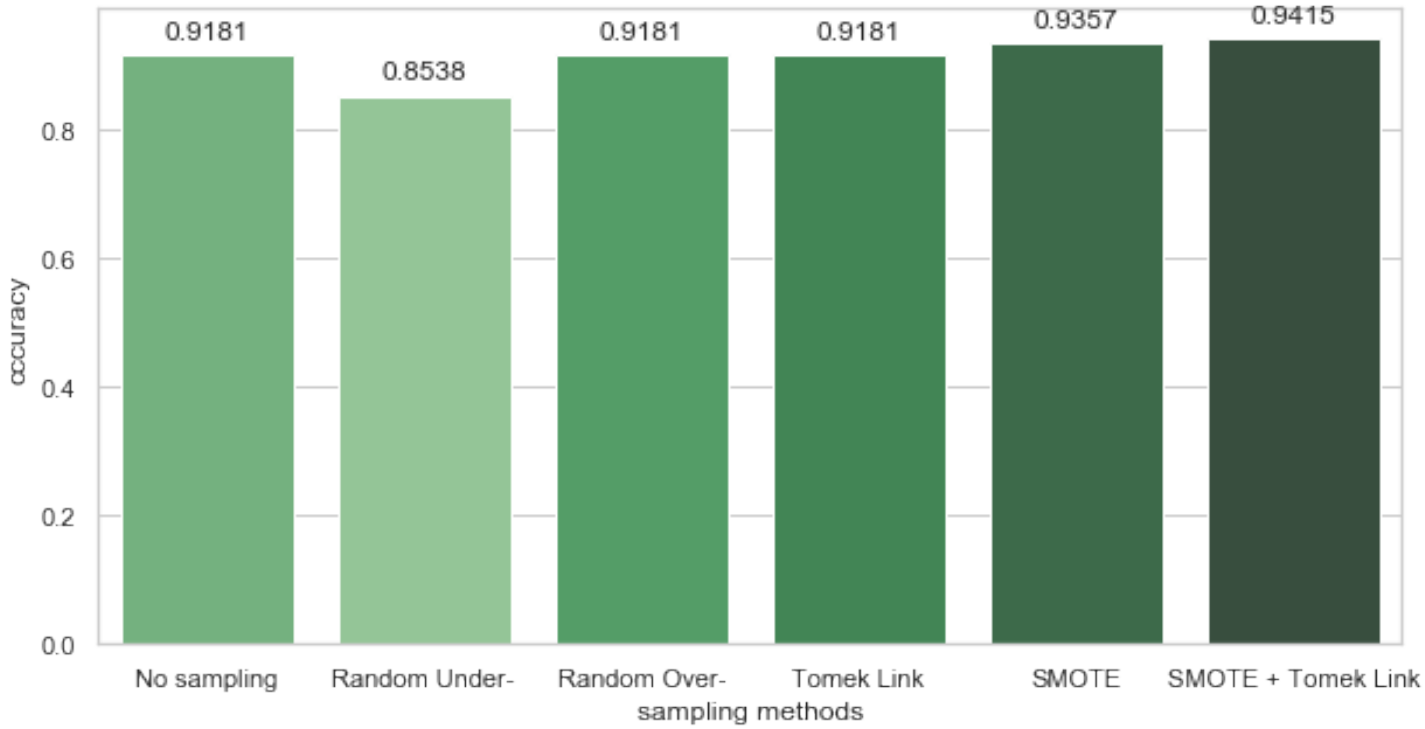


SMOTE

OVERCOME IMBALANCED DATA

- Tomek-links (under-sampling)
- SMOTE (over-sampling)
- SMOTE + Tomek-links

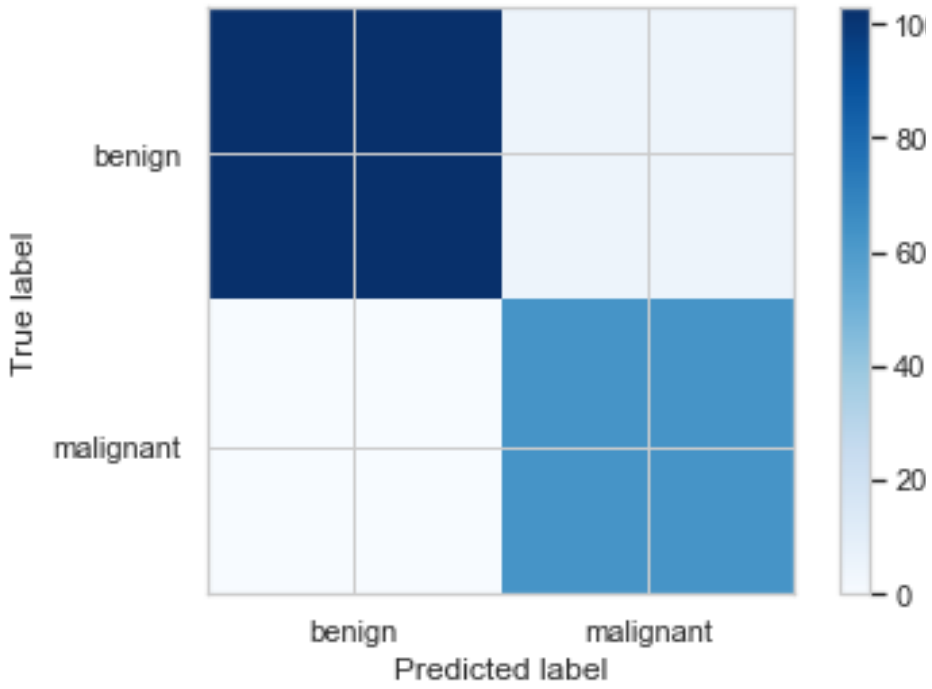
FINAL RESULT



Compare the performance

- Use the best features selected
- Variable-controlling approach
- SMOTE + Tomek Link achieves the best performance

FINAL RESULT



Build final model

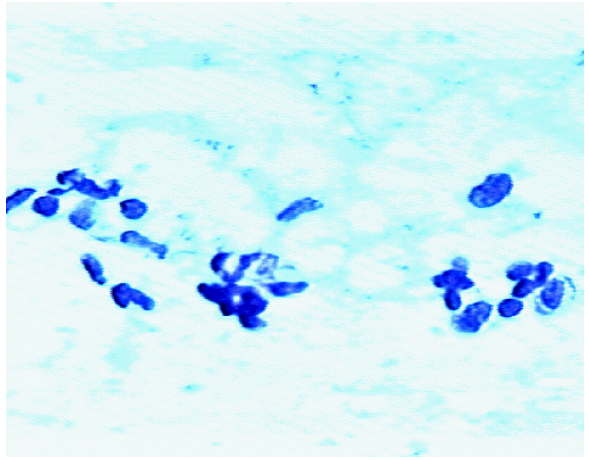
- SMOTE + Tomek-Link sampling method
- Use the best features selected
- Utilize Random Forest Model

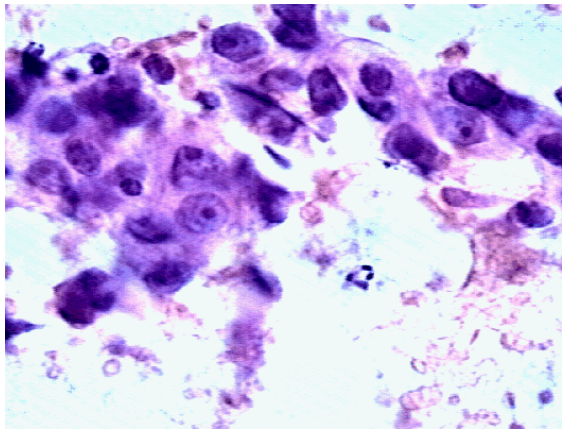
Accuracy: 97.08%

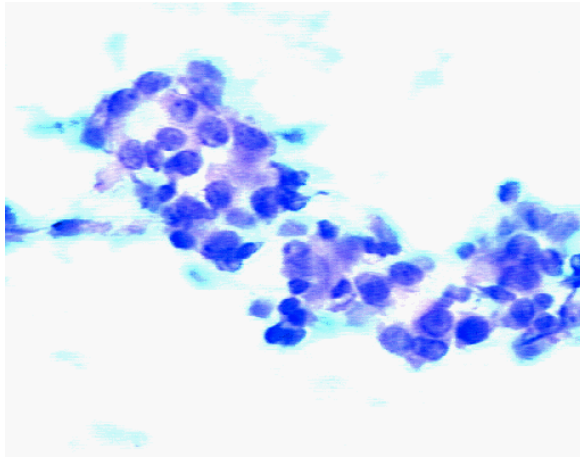
- From the confusion matrix, the prediction is balanced

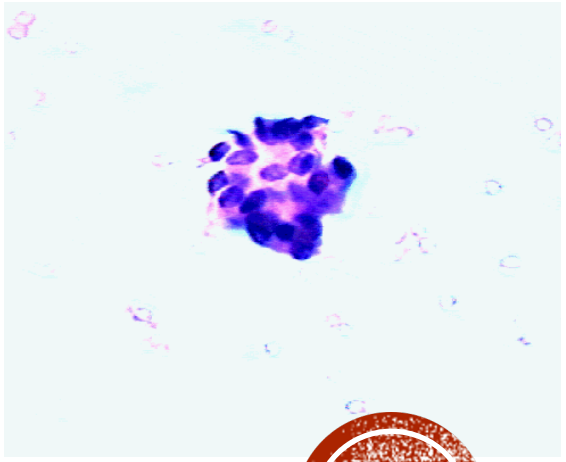
FUTURE WORK

- **Extract features from raw images**









REFFERENCE

- Wisconsin Diagnostic Breast Cancer dataset: [https://www.kaggle.com/uciml/breast-cancer-wisconsin-data]
- Breast Cancer Image: [ftp://cs.wisc.edu/math-prog/cpo-dataset/machine-learn/WDBC/]
- pandas: [https://pandas.pydata.org/docs/]
- scikit-learn: [https://scikit-learn.org/stable/]
- matplotlib: [https://matplotlib.org]
- seaborn: [https://seaborn.pydata.org]
- imblearn: [https://imbalanced-learn.readthedocs.io/en/stable/api.html]

THANK YOU

-- BY GROUP 6