Breast cancer signs and prediction

Chung-I Huang/ 2018-12-05

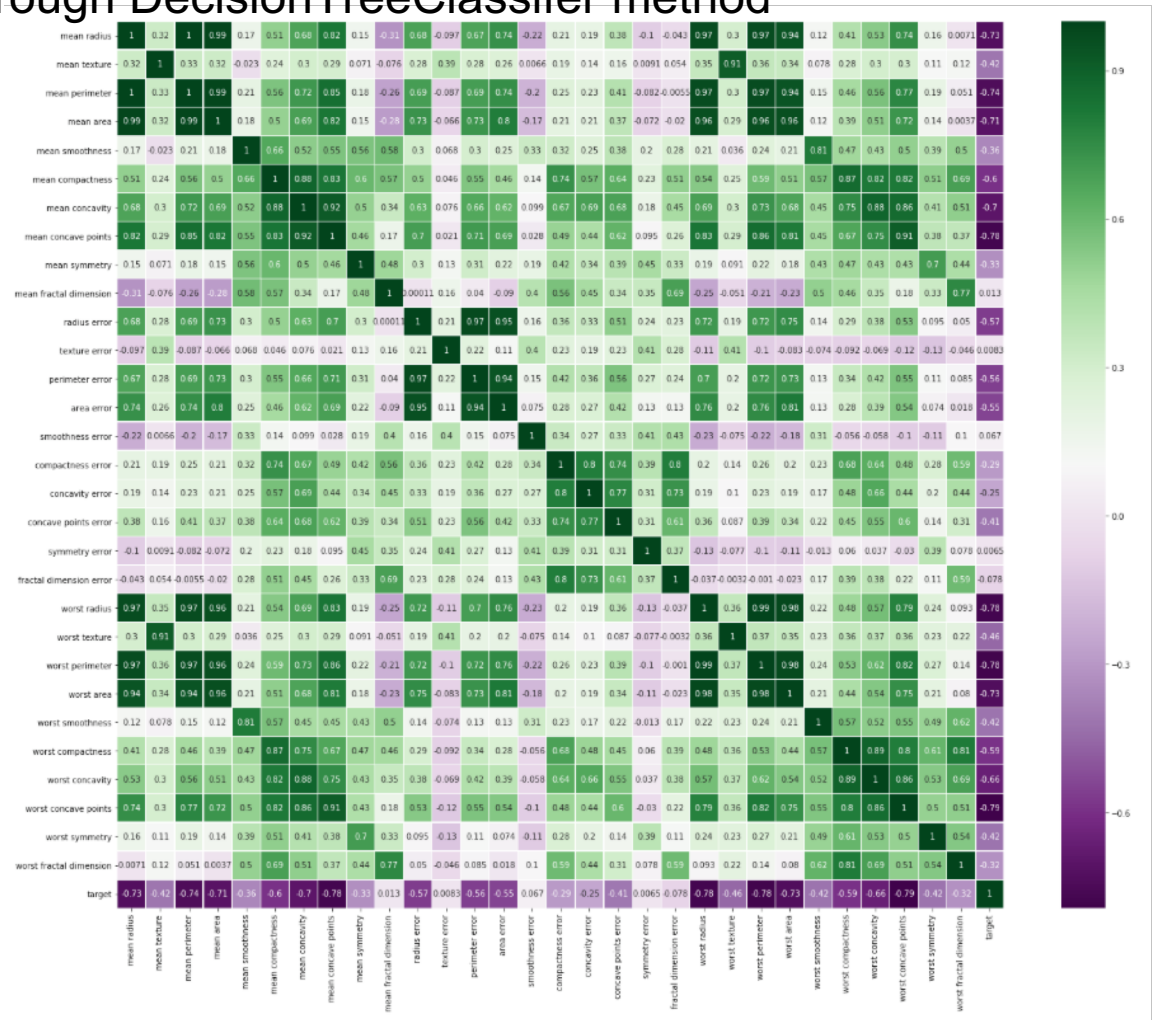
Abstract

I am using sklearn dataset's breast cancer data to figure out what symbols are more associated to breast cancer (benign or malignant)? The approaches I am applying for are DecisionTreeClassifier, KNN, and KMeans. The 1st & 2nd belong to the 'supervised' machine learning, and the 3rd is 'unsupervised'. Without a doubt, the 1st & 2nd have a higher prediction result than the 3rd. In addition, there are 9 characteristics among the total 31 which have a high correlation to breast cancer (benign/malignant).

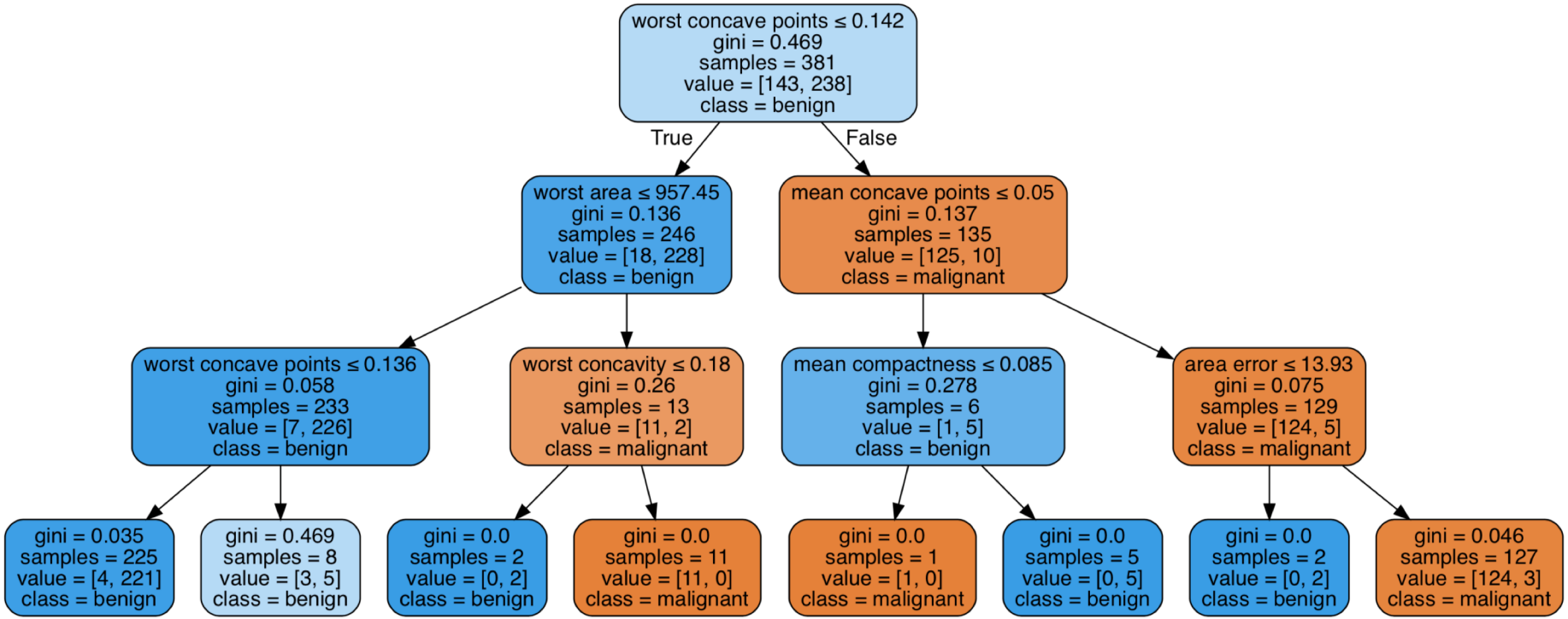
Motivation

To know which characters have a higher chance to lead to breast cancer, I adopted ‘heatmap’ and ‘graphviz’, two visualization tools (see next page). Thus, we can find out the high correlation factors and how the machine distinguishes between ‘benign’ and ‘malignant’.

Heatmap : Through DecisionTreeClassifer method



Graphviz : Through DecisionTreeClassifer method



Dataset(s)

Dataset -> https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

Since the dataset is already in the sklearn.datasets, we can load it through the command of ‘from sklearn.datasets import load_breast_cancer’ in Juypter Notebook.

Data Preparation and Cleaning

First of all, check the data see if any row has null value on it by the command of ‘isnull().any()’. If yes, then use ‘dropna()’ function to clean the data.

In my case, I don’t see any null value.

```
df.isnull().any()
```

```
mean radius                False
mean texture               False
mean perimeter             False
mean area                  False
mean smoothness            False
mean compactness           False
mean concavity             False
                           ...
worst smoothness           False
worst compactness          False
worst concavity            False
worst concave points       False
worst symmetry             False
worst fractal dimension    False
target                     False
Length: 31, dtype: bool
```

Research Question(s)

1. Which characteristics have a high correlation to breast cancer?
2. Comparing two different supervised classification approaches, which one has a higher accuracy rate?
3. Using the unsupervised method, then compare accuracy rate to the above two supervised methods.

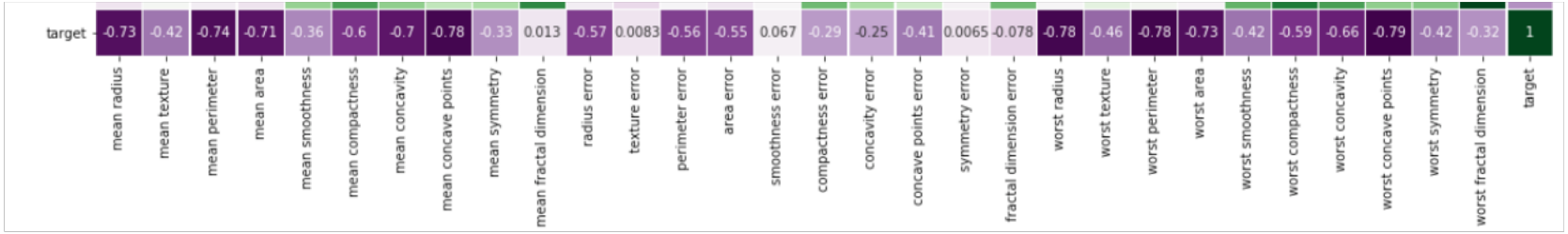
Methods

Supervised approach: DecisionTreeClassifier, KNN

Unsupervised approach: KMeans

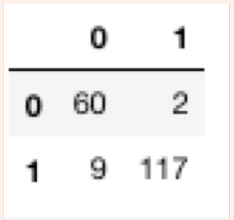
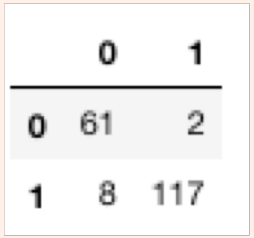
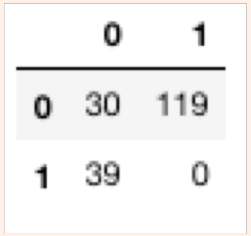
Because they are simple algorithms, and our data only has two labels – benign and malignant.

Findings & Conclusion



According to page 4 the heatmap result, we noticed the darker purple color has a higher correlation to breast cancer. In page 5, we also found the decision tree using these darker purple signs to classify benign or malignant.

Findings & Conclusion

	DecisionTreeClassifier	KNN	KMeans
Accuracy rate	94.14%	94.68%	15.95%
Confusion matrix			

DecisionTreeClassifier and KNN has similar prediction result ~94%, but KMeans is obviously lower due to it is an unsupervised approach.

Limitations

The data only has two results – benign and malignant, which is easier for doing supervised data training.
However, I can foresee these two approaches won’t have that high of an accuracy rate (94%) when it has 3+ different results.

Acknowledgements

My friend suggested to me to use the KNN and KMeans approaches in this case due to our data being simple and belonging to the classification question.

References

Data: sklearn.datasets

I did this presentation on my own.