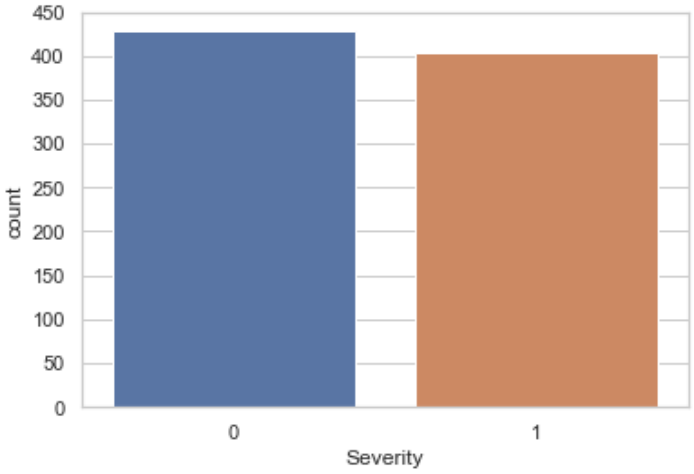
1. Data Exploratory

After dropping column “BI-RADS”, there are 5 columns left. Among these columns, “Age”, “Shape”, “Margin” and “Density” are the independent variables, column “Severity” is the dependent variable.

There are missing values in each column except column “Severity”. I dropped all the instances which include one or more than one missing values, then I got a new data frame which includes 831 rows and 5 columns (the original data set has 961 rows, 13% data has been dropped).

This new data frame is a balanced dataset according to the following graph, which shows the total number of “yes” (1) and “no” (0) based on column “Severity”.



2. Accuracy

Here is the accuracy by using different techniques:

	Accuracy
Decision tree	78.365% (single split)
Decision tree	79.188% (mean value of 10 K-Fold)
Random forest	80.872% (mean value of 10 K-Fold)
KNN (K = 10)	79.185% (mean value of 10 K-Fold)
KNN (the highest accuracy after trying K from 1 to 50)	80.630% (when k = 23) (mean value of 10 K-Fold)
Naive Bayes (MultinomialNB)	69.795% (mean value of 10 K-Fold)

PS: I did not specify the random seed, you may get slightly different results every time you run the model.

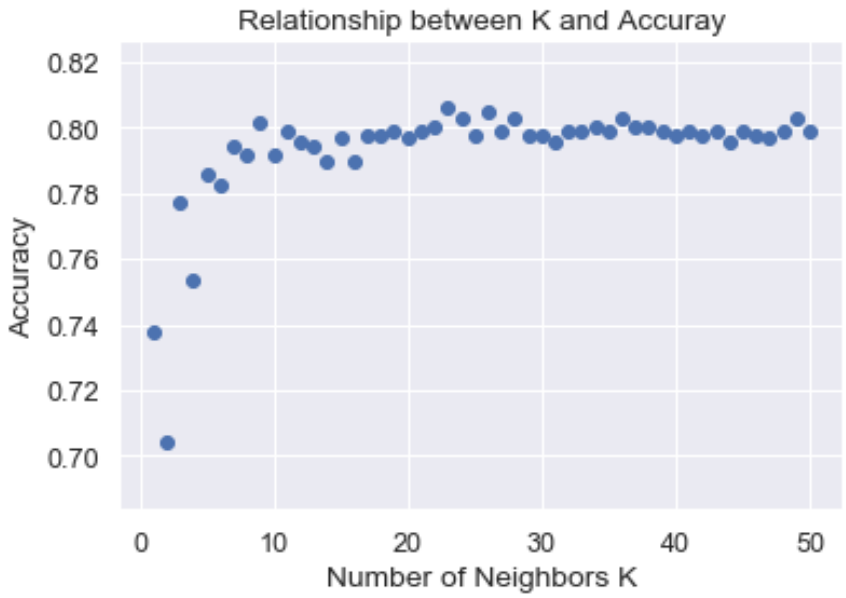
3. Conclusion

(1) Among the above techniques, Random Forest gives the highest accuracy: 80.872%.

(2) About KNN

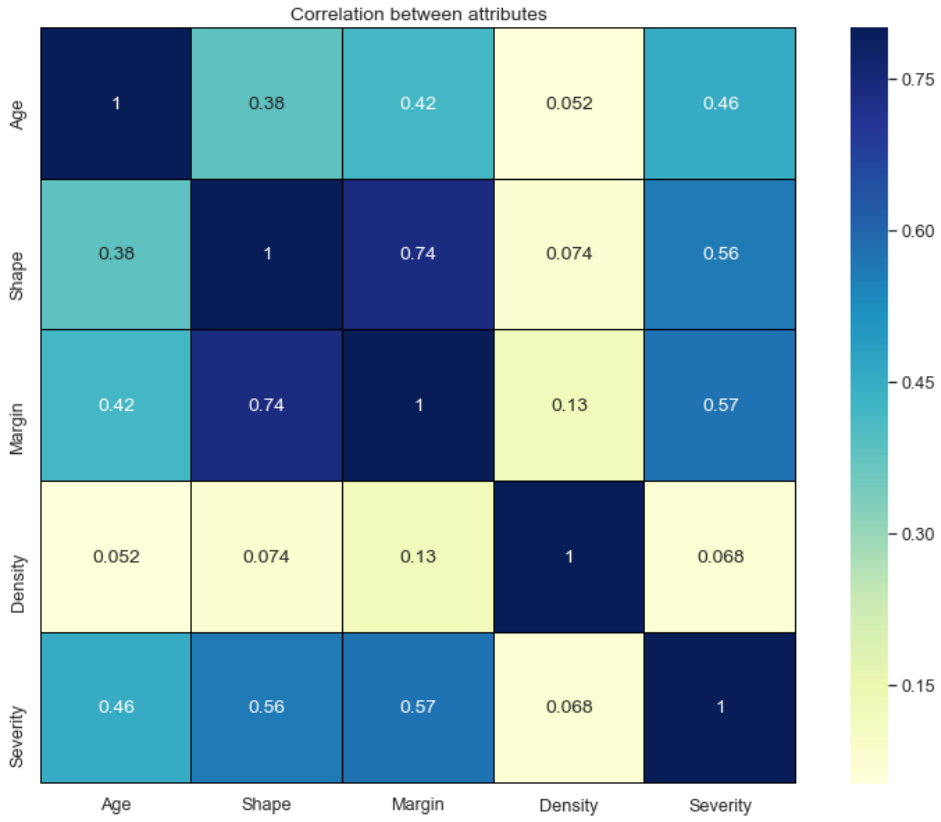
- After trying K from 1 to 50, I found the model's accuracy keeps rising when k increases from 1 to 10, then the accuracy stays stable.

- Also, different K will not make a substantial difference in accuracy, especially when 10 <= K <=50. The mean value of the accuracy (when k is from 1 to 50) is 79.327%, and the standard deviation is 0.017, which is very small.



4. Future work

According to the following correlation matrix heat map:



- Independent variables “Shape” and “Margin” have relatively high correlation with the dependent variable “Severity”. “Density” has very low correlation with “Severity”.

- “Shape” and “Margin” have high collinearity. To improve the model, we need to overcome this high collinearity.