RoboGarden Python for Machine Learning Capstone Project ML-006

Breast Cancer Diagnosis using Cytology Features

Project Description:

The purpose of this exercise was to determine the best Classification Algorithm to use to determine breast cancer malignancy using 9 cytological features[1] identified from fine needle aspirations (needle biopsy) of 669 patients. The dataset for this project was derived from the Wisconsin Breast Cancer dataset from the University of California at Irvine (UCI) Machine Learning Repository (via Kaggle.com), and was collected from Jan 1989-Nov 1991.

Project Data

Each of the 9 features describe the external appearance and internal chromosome changes, using a 1-10 scale (no need to normalize the data), with 10 being the most abnormal state. The class feature is represented by 2 (benign-non-cancerous) or 4 (malignant-cancerous).

After importing the data into python, the dataset was examined for missing data, anomalous data or duplicated data.

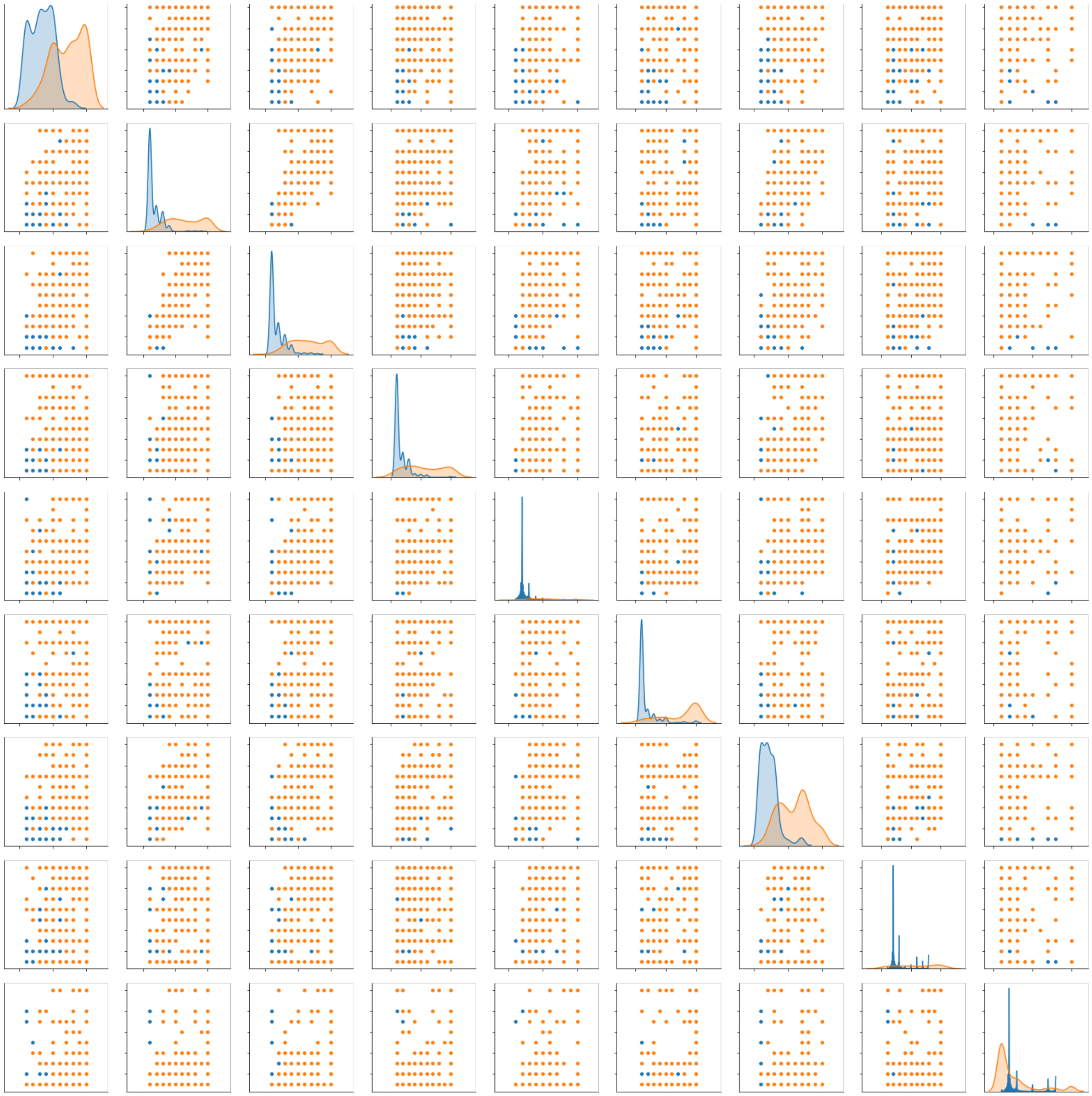
There were 16 missing values in the dataset, all contained in the 'bare_nuclioli' feature, and was represented by '?'. Because we did not know the reason behind the missing data ("unavailable attribute values'), it was decided to remove those patients reports entirely, so as to not unknowingly bias the results.

There were 8 duplicated patient records in the dataset, which were identified using all 11 features, including the patient ID. These records were also deleted from the dataset prior to training the models.

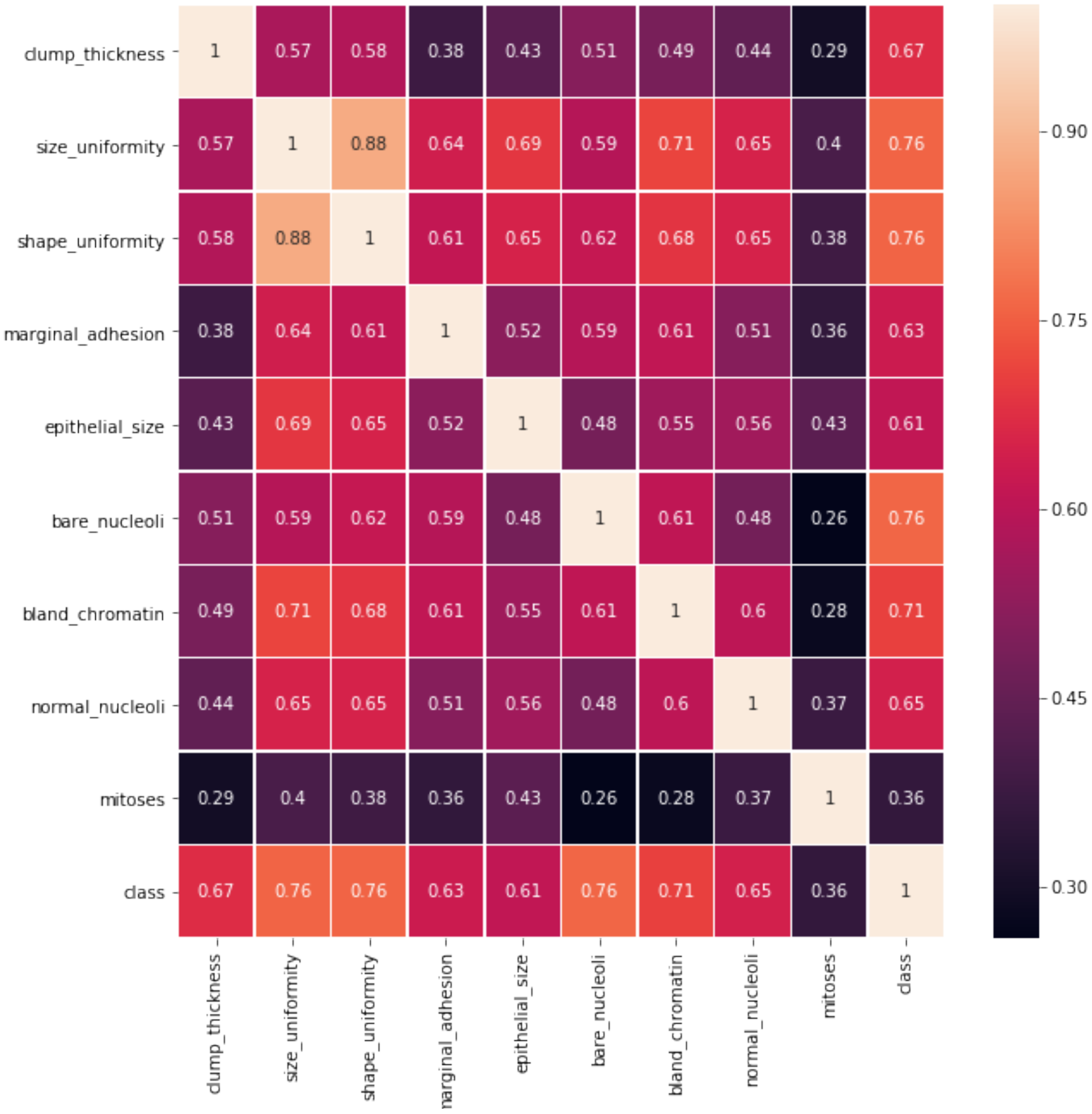
After dropping the ID feature (not necessary for our purpose), duplication analysis was run again. It was determined that there were 226 identical results (every value for every feature equal). Despite representing almost 1/3 of the records, I removed these 226 records prior to training the models. The remaining dataset is now 449 records vs original of 699. (Models were subsequently run to include those 226 duplicated records and did not have significant impact on final classification accuracy).

Data Analysis

Pairplots and a heat map were created to examine the correlations between features. Ideally, we like to use independent variables in our analysis, so we could remove those features that show a strong dependence on others if necessary. While there are some weak correlations between a few of the features, it was not enough to discourage using all of the features to train our models.



PAIRPLOTS COMPARING FEATURE TO FEATURE
LITTLE SCATTER IN PLOTS DUE TO HIGHLY REGULARIZED DATA.

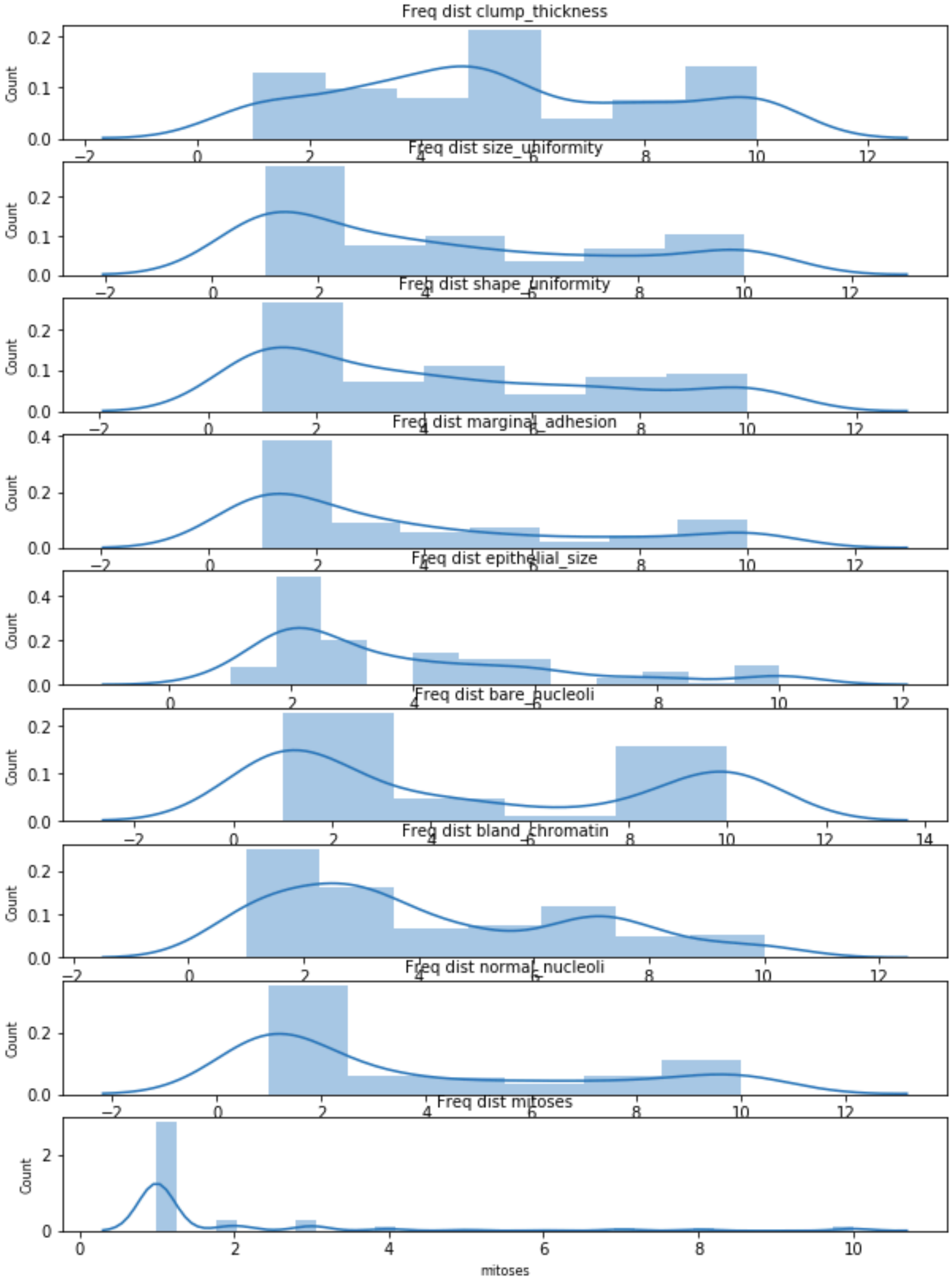


HEATMAP COMPARING FEATURE CORRELATIONS

Mitoses is least correlatable with class.

External features (size, shape, thickness) correlate well to each other as well as class.

From the distribution plots, we can see that most features have relatively good distributions and/or are bi-modal, with the exception of bare_nucleoli and mitoses.



DISTRIBUTION PLOTS FOR EACH FEATURE

Training Models

For this project, there were 6 Classification models trained and tested. For each model, the 2/3 of the dataset was used to train the model, and 1/3 was the test data (random state=42, The Answer to the Ultimate Question of Life, the Universe and Everything)

The six classification models used were:

Decision Tree Classifier

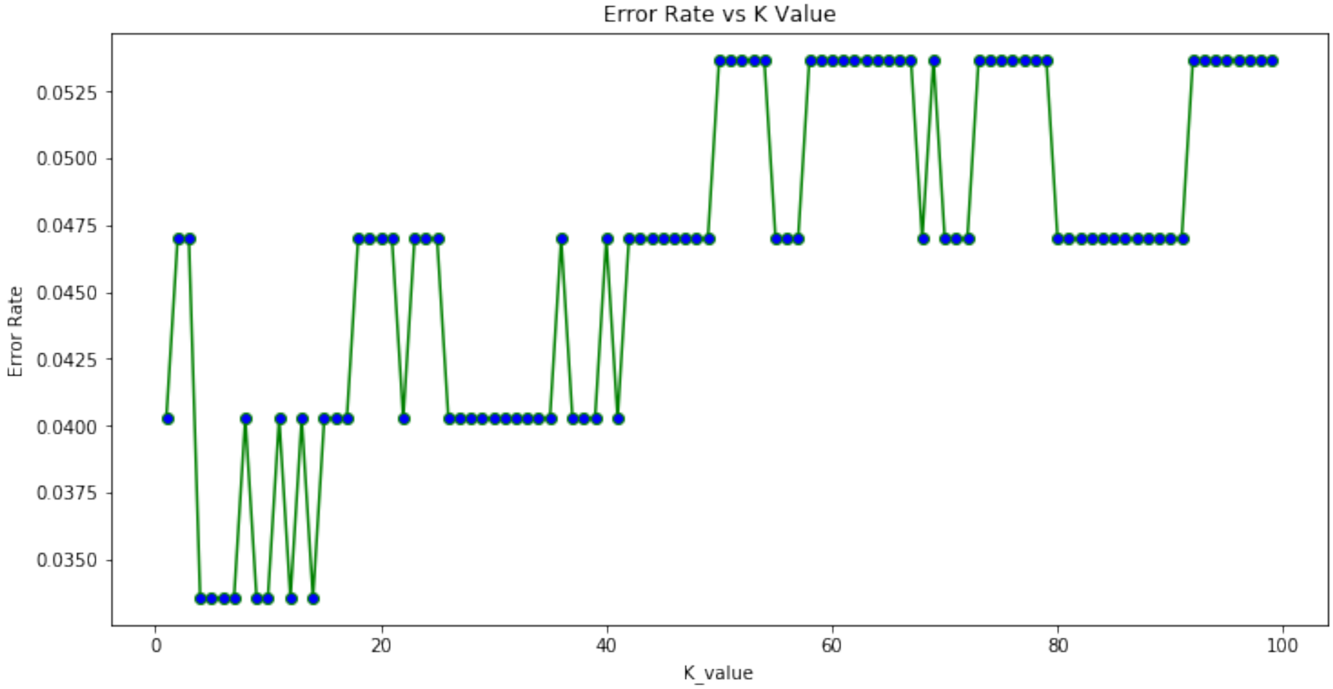
Gaussian Naive Bayes

Logistic Regression

Random Forest

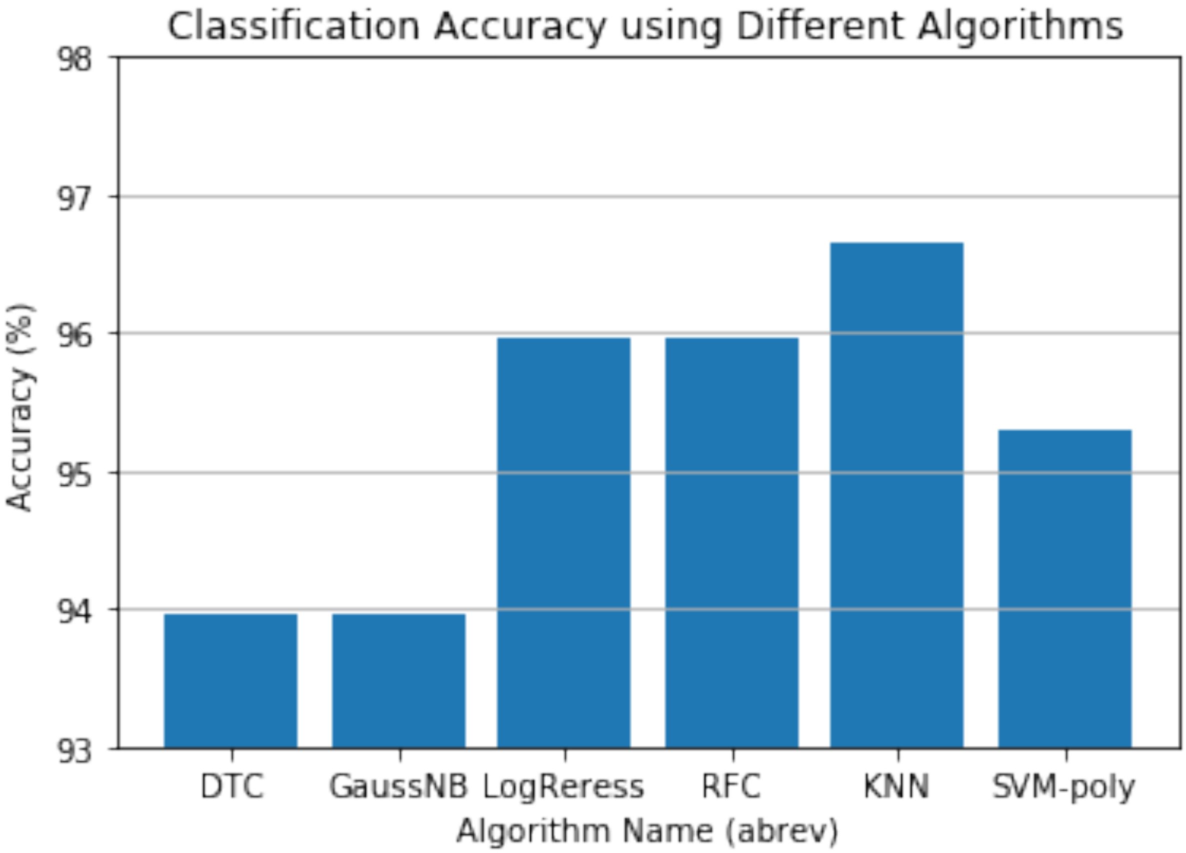
SVM - kernel=poly

K-Nearest Neighbours - n_neighbors=4 (as determined by K vs error rate plot below).



For each classification model, an accuracy score was calculated, using True Positive (TP) and True Negative (TN) outputs. (N=total number of test data points)

$$\text{Accuracy (\%)} = \frac{TP + TN}{N}$$



As we can see from the Accuracy Table, the KNN Classification returned the best classification accuracy, with 96.6%, although all of the models returned accuracies greater than 93%.

However, to determine a better ranking between the Decision Tree Classifier and the Gaussian Naive Bayes Classifier, we can look at their confusion matrices, as different results can lead to similar accuracy statistics. These confusion matrices can be used for more specific assessment of each model.

Confusion matrices are defined as :

```
[True Negative  | False Positive]
[False Negative | True Positive]
```

```
DecisionTreeClassifier       GaussianNB
(Accuracy: 93.9597%)         (Accuracy: 93.9597%)

[[66  5]                     [[63  8]
 [ 4 74]]                     [ 1 77]]

RandomForestClassifier       KNeighborsClassifier
(Accuracy: 95.9732%)         (Accuracy: 96.6443%)

[[66  5]                     [[67  4]
 [ 1 77]]                     [ 1 77]]

LogisticRegression           SVM-poly
(Accuracy: 95.9732%)         (Accuracy: 95.3020%)

[[66  5]                     [[65  6]
 [ 1 77]]                     [ 1 77]]
```

Analyzing the confusion matrices, we can see that the distribution of True/False classifications differs slightly between models, and as such could be used to high-grade/low-grade a model if Sensitivity[2] (rate of correctly classified Positives) or Specificity[3] (rate of correctly classified Negatives) comparison is required.

Conclusion

Using this dataset with these models and parameters, the K-Nearest Neighbours Classifier predicted Breast Cancer Malignancy the best, with 96.6% accuracy, returning just 5 falsely classified records (1+4).

Thank yous

Biggest and bestest **THANK YOU** to Dr. Qazi, for his enthusiastic and entertaining introduction to Python and Machine Learning.

Kanishka the Great easily deserves the next biggest and bestest THANK YOU. He deserves more than a thank-you for trouble shooting, helping, fixing, suggesting, laughing at, researching in a one-on-one setting with all of us!

A final thank-you to Fatemahjoon - for the RoboHelp, mission advice, Github tutorial, mostly laughing at all of my jokes.

1: Description of each feature:

clump thickness: cells die and deposit together in a tissue to form clumps that is clinically presented as swelling.

size uniformity: same in size across the tissue

marginal adhesion: normal cells tend to stick together, where cancer cells tend to lose this ability. loss of adhesion is sign of malignancy

epithelial size: epithelial are a type of cells in our body that makes up the layer of majority of human tissue. The Epithelial cells that are significantly enlarged may be a malignant cell.

bare nucleoli: Describe where nucelus is not surrounded by cell fluid (cytoplasm) nucleus. Nucleus is the core of any cell where the DNA lies.

bland chromatin:describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin (part of chromosome) tends to be coarser (broken fragments of chromosome are seen in malignant cells).

normal nucleoli: in normal cell, the nucleus has small structures. the normal nucleoli are very small and hardly visible. In cancer cell, nucleoi become very prominent and bigger in size.

mitoses: mitosis is defined as nuclear division which produces two identical daughter cells. In cancer cells, this process is disrupted. A cytologist will look at the number of mitoses to determine the state of cancer.

https://pdfs.semanticscholar.org/8e1a/85632e587e1b1fb180e3bdf8aebda0d9f91d.pdf

2:

$$\text{Sensitivity (\%)} = \frac{TP}{TP+FN}$$

3:

$$\text{Accuracy (\%)} = \frac{TN}{TN+TP}$$