# Capstone Project 1 Milestone Report

## Initial Project Ideas

### Predict Energy Requirement based on Building Attributes

Predict the heating and cooling load for a building based on its attributes such as glazing area, compactness, surface area, wall height etc. The data was generated using simulation software 'Ecotect'. This software is owned by Autodesk.

[Link](http://archive.ics.uci.edu/ml/datasets/Energy+efficiency)

### Predict Tumor Classification using biopsy data

Predict the diagnosis (i.e. malignant or benign) of biopsy using the cell attributes from biopsy. The attributes are discrete and have only certain values.

[Link](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29)

### Predict Epileptic Seizures

Time series data containing EEG readings. The EEG data are sampled at 178 datapoints per sec. At the end of each second, the patient's state is recorded in numbers from 1 to 5. Number 1 meaning the patient is having a seizure.

[Link](http://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition)

Predicting tumor classification from biopsy using cell attributes was selected. The author was very familiar with regression methods. This problem was selected in order for the author to gain experience in using classification methods.

## Problem Definition

* Predict if a patient has breast cancer based on the characteristics of cells derived from Fine Needle Aspiration procedure[[1]](#footnote-1).
* Fine Needle Aspiration is a less invasive alternative to full blown surgery. If this procedure is found to be reasonably accurate then it can significantly reduce patient suffering arising from full blown surgery.

## Stakeholders

Building a model to predict the cell classification using results from Fine Needle Aspiration would be of great interest to medical professionals (e.g. pathologists & oncologists). They will use the results to determin treatment plan.

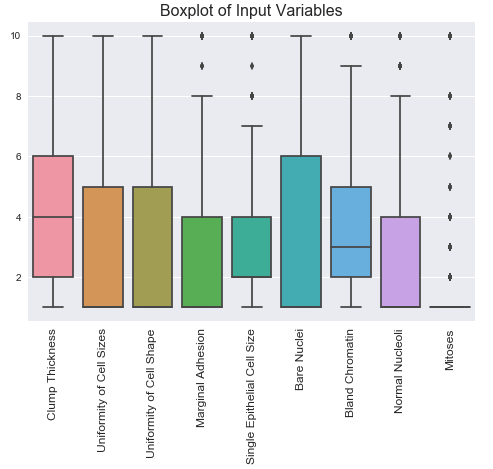
## Data Source

The data was collected from the UCI Machine Learning website: [Link](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29). The raw data was in the csv format. It was copied to excel and then read into a data frame.

## Data Characteristics

* There are 699 Data points, 9 input variables & one output variable. 16 missing values.
* All the input variables are qualitative in nature. They take integer values from 1 to 10.
* Excepting for cell thickness, none of them are randomly distributed. This based on the significant differences between the mean and median and the skewed nature of the population distribution as evident from the box plot.
* Owing to their non-normal distribution & qualitative nature none of the data points that were past 3 standard deviations from the mean were removed.

Figure 1. Boxplot of Input Variables



## Dealing with Missing Values

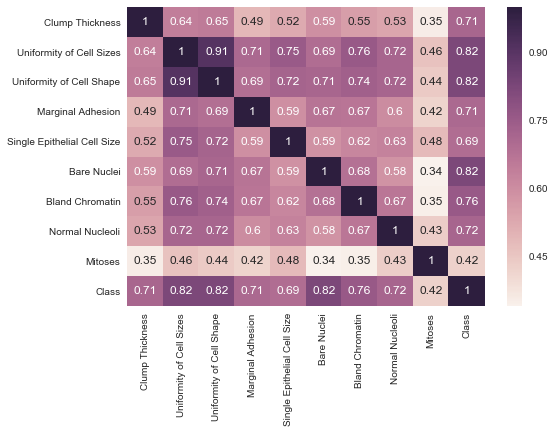
Only 2% of the indices had missing values. Therefore these rows were dropped from the analysis using the drop rows function in pandas.

Table 1. Variables in the Dataset

|  |  |
| --- | --- |
| Input Variables | Output Variable |
| Clump Thickness  Uniformity of Cell Sizes  Uniformity of Cell Shape  Marginal Adhesion  Single Epithelial Cell Size  Bare Nuclei  Bland Chromatin  Normal Nucleoli  Mitoses | Tumor Classification: Malignant or Benign |

## Relationships between Variables

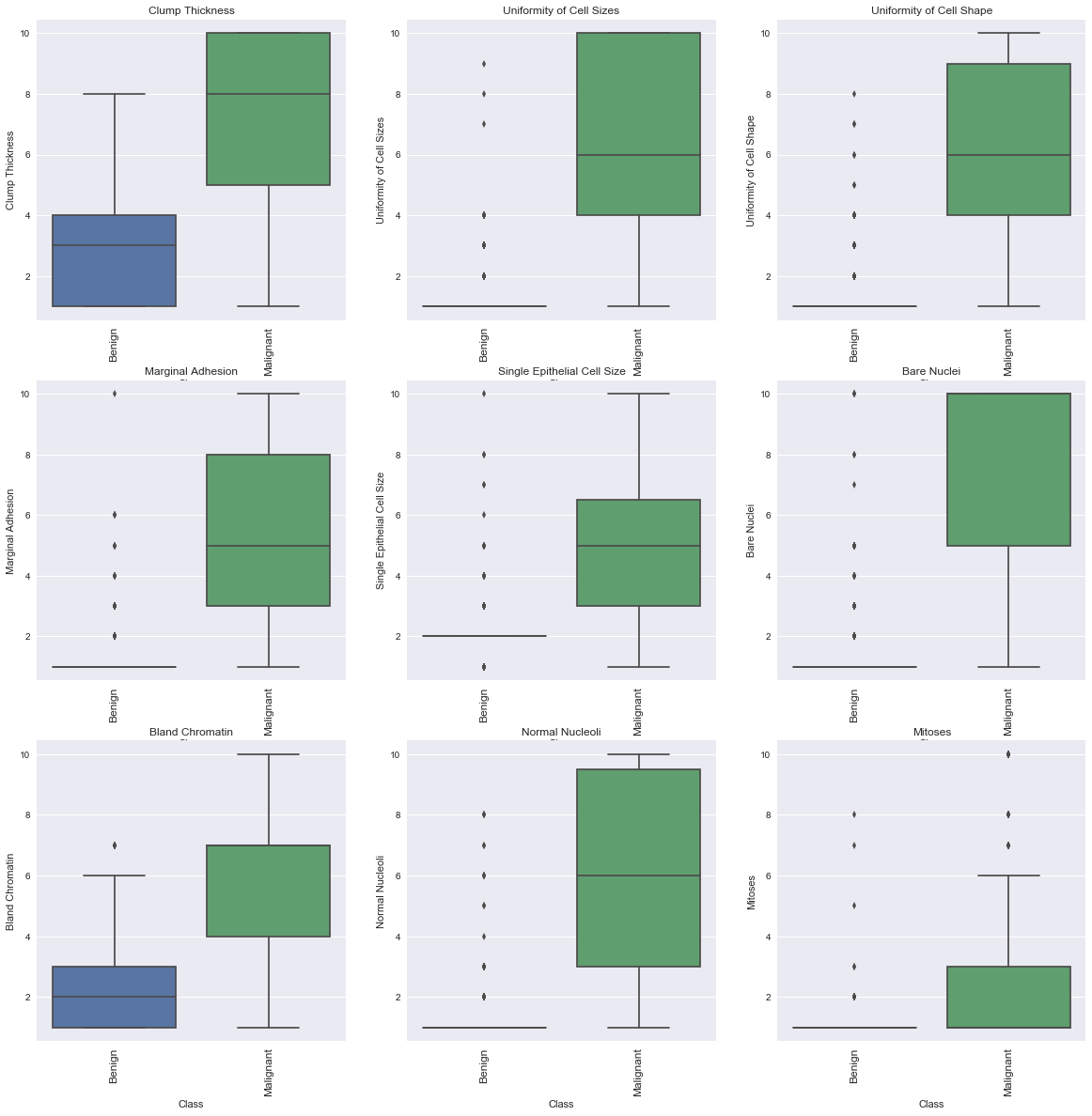
Figure 2. Correlation Matrix



From the above correlation matrix for all the input & output variables, following can be gleaned:

1. Many of the input variables are strongly correlated with Uniformity of Cell shape & size.
2. Uniformity of cell sizes, Marginal adhesion and cell shape are highly correlated.
3. Mitoses has the least correlation with any other input variables.
4. Excepting for Mitoses, most of the input variables have strong correlations (i.e. > 0.7) with output variables as seen in the last row. This is consistent with the trends observed in the box plot shown in Data Story Report (reproduced here as Figure 3 for convenience).
5. Cell Class (i.e. Y Variable) is most correlated with Unformity of Cell Shape/Size & Bare Nuclei. These same variables were also found to be the most significant variables from the analysis of logistic regression coefficients.

Figure 3 Box Plots Showing Relationship between Input and Output variables



## Building Model Using Logistic Regression

* A logistic regression model was built using all the input variables. 90% of the data were used to train the model and the remaining 10% were used to test.
* Table 1 shows the confusion matrix. Resulting Model has good accuracy 87%.
* Table 2 shows the coefficients of logistic regression model. Most of the variables move in the same direction as the Y variable (i.e. classification). This is also consistent with trend observed in the the box plots in Data Story Report.
* Using Alpha of 0.05, only “Uniformity of Cell Shape” has p-value > Alpha. Rest of the variables are relevant for the model.

Table 2. Confusion Matrix from Logistic Regression Model

|  | **Actual Benign** | **Actual Malignant** |
| --- | --- | --- |
| **Predicted Benign** | 41 | 6 |
| **Predicted Malignant** | 3 | 19 |

Table 3. Coefficients from Logistic Regression.

| **Variable Name** | **Coeff** | **std err** | **z** | **P>|z|** | Conf. Interval | |
| --- | --- | --- | --- | --- | --- | --- |
| **[0.025** | **0.975]** |
| **Clump Thickness** | -0.3225 | 0.059 | -5.445 | 0.000 | -0.439 | -0.206 |
| **Uniformity of Cell Sizes** | 0.9437 | 0.139 | 6.809 | 0.000 | 0.672 | 1.215 |
| **Uniformity of Cell Shape** | 0.1804 | 0.112 | 1.617 | 0.106 | -0.038 | 0.399 |
| **Marginal Adhesion** | 0.1780 | 0.079 | 2.239 | 0.025 | 0.022 | 0.334 |
| **Single Epithelial Cell Size** | -0.7894 | 0.105 | -7.514 | 0.000 | -0.995 | -0.583 |
| **Bare Nuclei** | 0.4921 | 0.065 | 7.615 | 0.000 | 0.365 | 0.619 |
| **Bland Chromatin** | -0.5463 | 0.095 | -5.751 | 0.000 | -0.732 | -0.360 |
| **Normal Nucleoli** | 0.3595 | 0.078 | 4.598 | 0.000 | 0.206 | 0.513 |
| **Mitoses** | -0.2500 | 0.089 | -2.814 | 0.005 | -0.424 | -0.076 |

### Significant Variables

* Since all the X variables have the same range (0 to 10), the magnitude of the coefficient for each variable represents its importance.
* Based on the above principle, the top four variables were selected and a model was built based on them. The four variables are:
  + Uniformity of Cell Size
  + Single Epithelial Cell Size
  + Bland Chromatin
  + Bare Nuclei
* The resulting model has 85% accuracy score, same as the original model built using all variables.
* Based on the above result, the four variables shown above are the significant ones. Their coefficients are shown in Table 4.

Table 4. Coefficients of Variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable Name** | **coef** | **std err** | **z** | **P>|z|** | **Conf. Interval** | |
| **2.5** | **97.5** |
| **Uniformity of Cell Sizes** | 1.0600 | 0.110 | 9.604 | 0.000 | 0.844 | 1.276 |
| **Single Epithelial Cell Size** | -0.8451 | 0.098 | -8.608 | 0.000 | -1.038 | -0.653 |
| **Bland Chromatin** | -0.4952 | 0.081 | -6.100 | 0.000 | -0.654 | -0.336 |
| **Bare Nuclei** | 0.4701 | 0.060 | 7.820 | 0.000 | 0.352 | 0.588 |

### Analyzing the Relationships

* Table 4 shows the coefficients & statistics of the input variables. Their p-values show their impact on the model in unambiguous terms. Hence no further tests are pursued.
* 1000 Runs were performed and statistics of the model accuracy were calculated. These are shown in Table 5.

Table 5. Results from 1000 Runs

|  | **2.5 Percentile** | **Median** | **97.5 Percentile** |
| --- | --- | --- | --- |
| **Model Accuracy %** | 77 | 85 | 92 |

## Conclusions

* There is strong correlation present among the input variables and also with output variable.
* A logistic regression model was built using the variables that have the strongest correlation with the output variable. These are variables Uniformity of Cell Sizes, Single Epithelial Cell Size, Bland Chromatin & Bare Nuclei
* The resulting logistic model achieved > 77% accuracy in large number of simulated runs. This validates the model’s accuracy in predictive ability.

1. Fine Needle Aspiration is a diagnostic procedure in which needle is inserted into a body organ (e.g. breast) to collect a sample of cells. The cells are then examined for their characteristics to determine if they are malignant or benign. This procedure is a much safer and less invasive alternative to surgical biopsy. [↑](#footnote-ref-1)