Predicting the Diagnosis of Breast Tumor based on Cell Nucleus Measurements.

An accurate prediction of a tumor is of upmost importance for hospitals and patients. This project determines which, if any, measurements of a cancer cell nuclei contribute to an accurate diagnosis of an existing breast tumor. The two classifications of tumor in this case are malignant and benign. This project will be using data provided by the following study:

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

Using the provided data set, we will determine which if any of the independent variables are reliable in predicting the classification of a breast tumor. Because of the severity of a type II error in the diagnosis of a tumor, considerations will be taken that may limit the acceptable models.

The ten independent variables that will be examined in this analysis are:

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

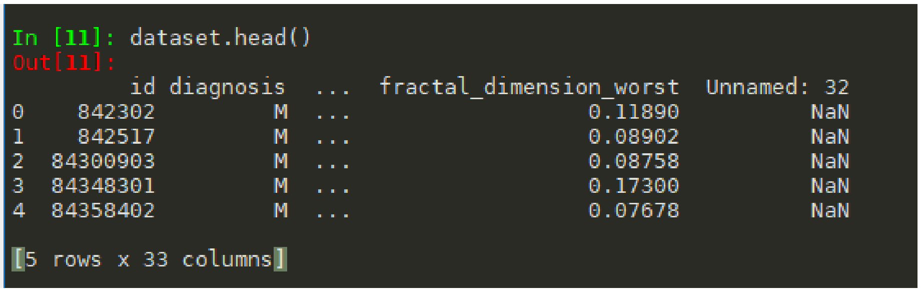
i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The null hypothesis states that there is no correlation between the successful diagnosis of a breast tumor and measurements of a breast cancer cell nuclei that is minimal in type II error. In order to reject the null hypothesis, there not only needs to be correlation, but correlation that will not lead to missed diagnosis of a malignant tumor.

Data Cleaning

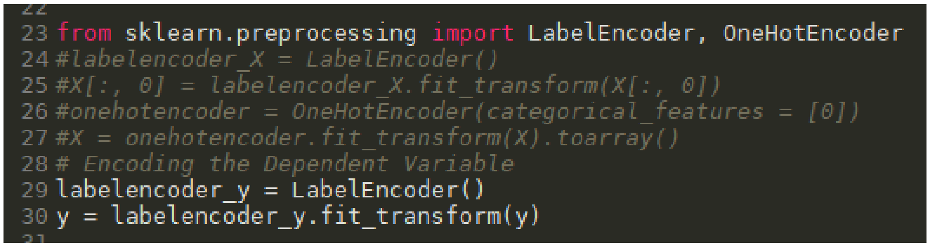
All fields are present and consistently formatted for each variable, so no data cleaning is needed. However, the independent variable needs to be encoded to conduct analysis.:



```
In [11]: dataset.head()
Out[11]:
         id diagnosis  ...  fractal_dimension_worst  Unnamed: 32
0    842302         M  ...                  0.11890          NaN
1    842517         M  ...                  0.08902          NaN
2  84300903         M  ...                  0.08758          NaN
3  84348301         M  ...                  0.17300          NaN
4  84358402         M  ...                  0.07678          NaN

[5 rows x 33 columns]
```

We will use the following code:



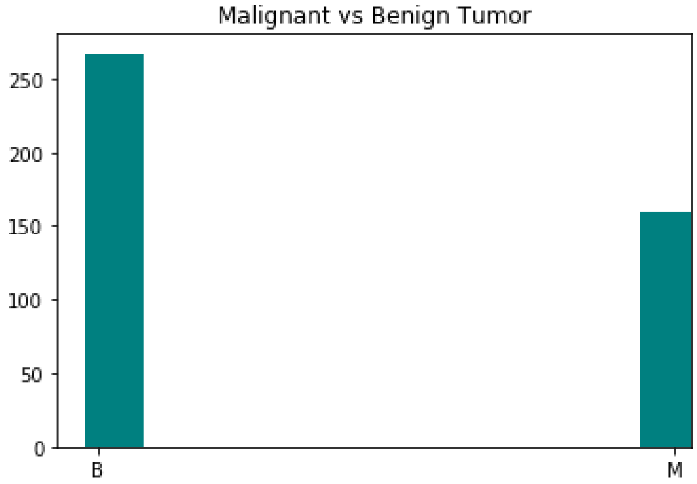
```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
#labelencoder_X = LabelEncoder()
#X[:, 0] = labelencoder_X.fit_transform(X[:, 0])
#onehotencoder = OneHotEncoder(categorical_features = [0])
#X = onehotencoder.fit_transform(X).toarray()
# Encoding the Dependent Variable
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
```

To convert the values ‘M’ and ‘B’ into ‘0’ and ‘1’, allowing the use of machine learning models.

Data Analysis

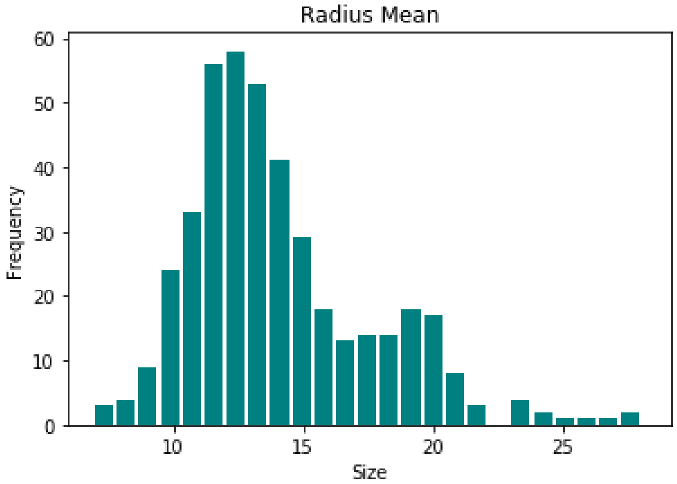
Before we begin analysis of predictive modeling, we will first look at each variable to gain insight into the data we are working with.

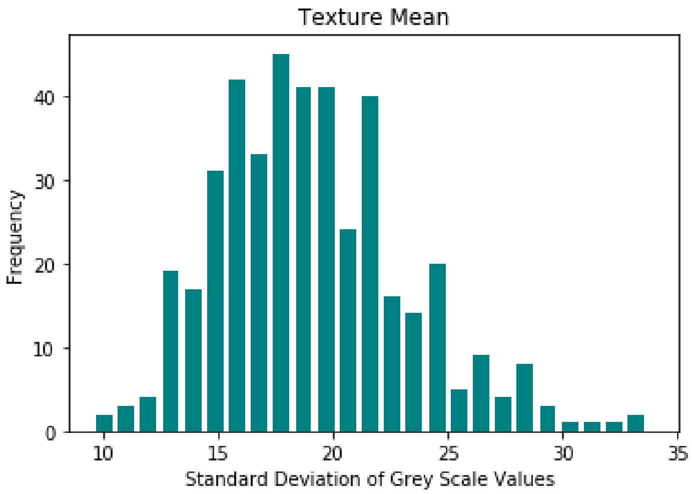
First, we look at the dependent variable, the diagnosis of malignant vs benign tumor. This dataset contains 569 entries, which will be separated into a training set of 426 entries and a test set of 143 entries. The following graphs represent data contained in the training set:

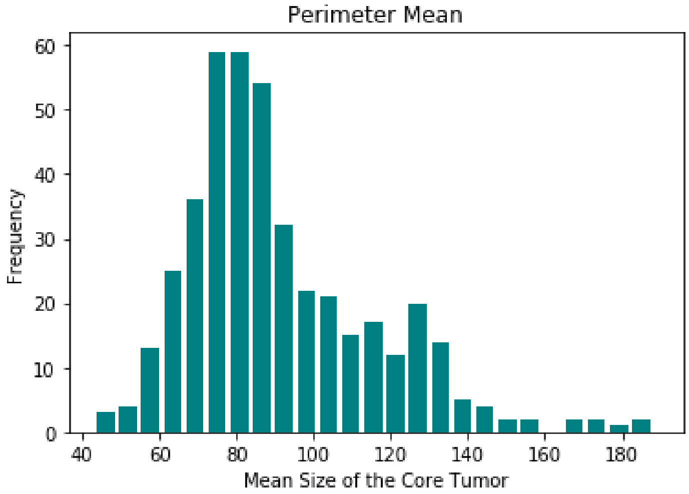


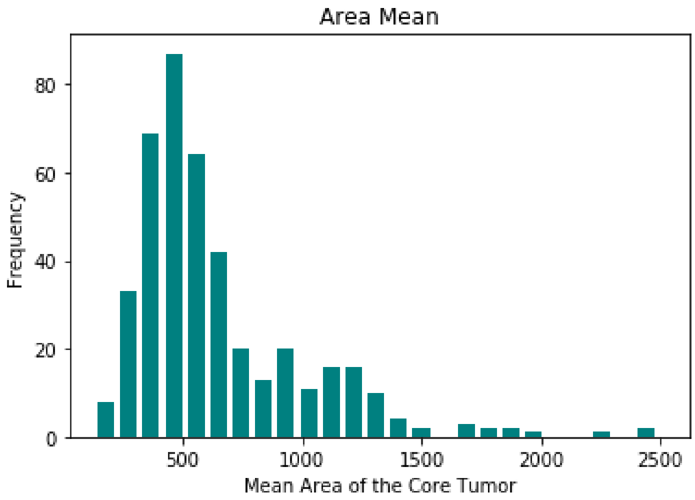
As indicated, there are more diagnosis of benign than malignant tumors. Out of the training set, there were 274 benign diagnosis, and 152 malignant diagnosis.

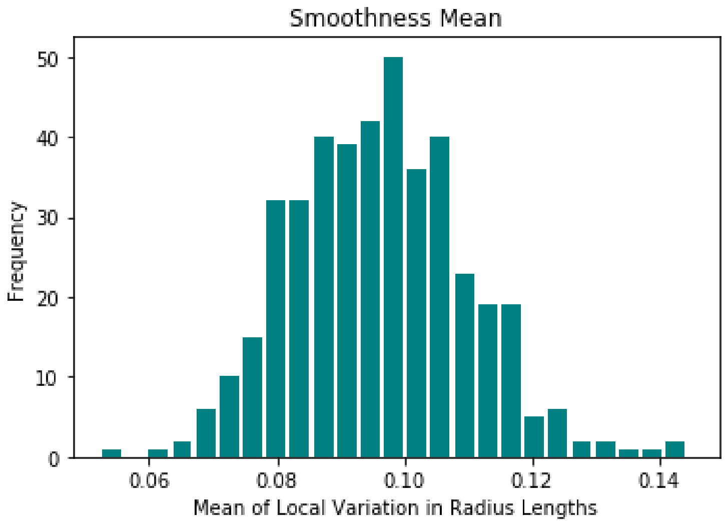
The following are the distributions of values for each independent variable:

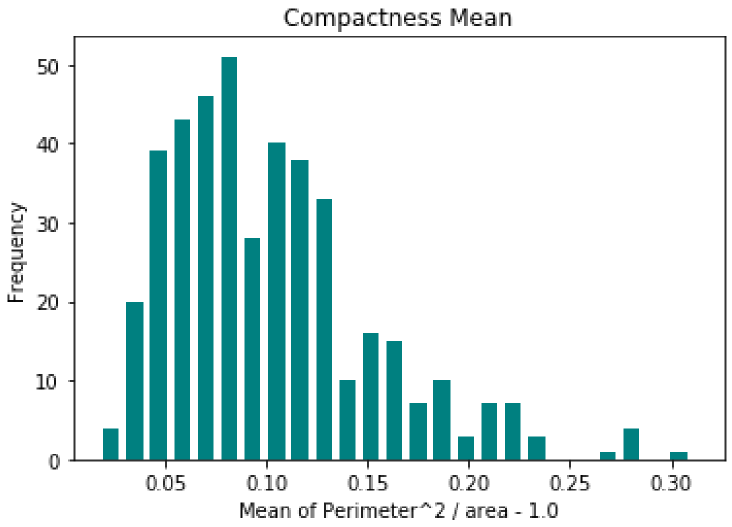


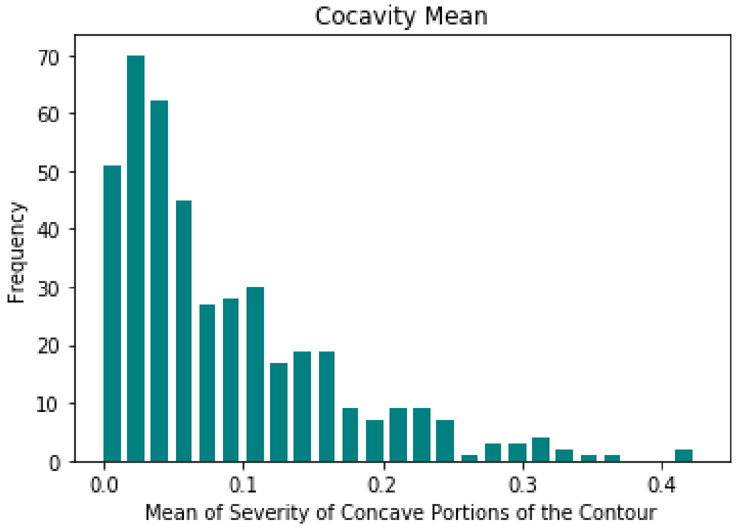

Texture Mean
Frequency
0
10
20
30
40
10
15
20
25
30
35
Standard Deviation of Grey Scale Values

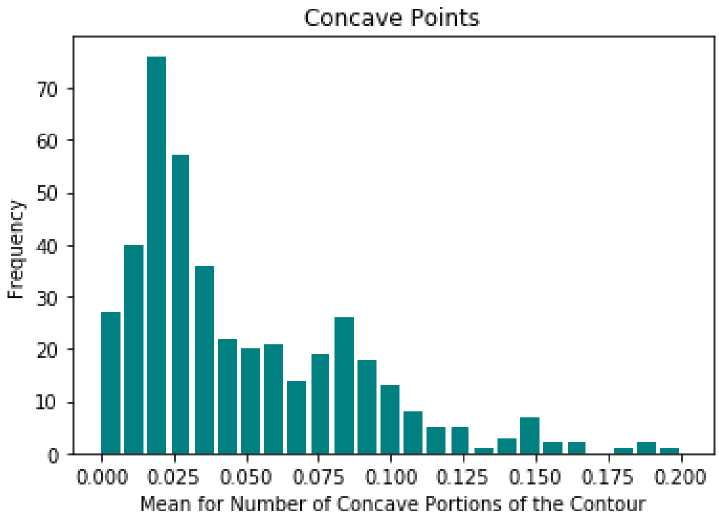

Perimeter Mean
Frequency
0
10
20
30
40
50
60
40
60
80
100
120
140
160
180
Mean Size of the Core Tumor

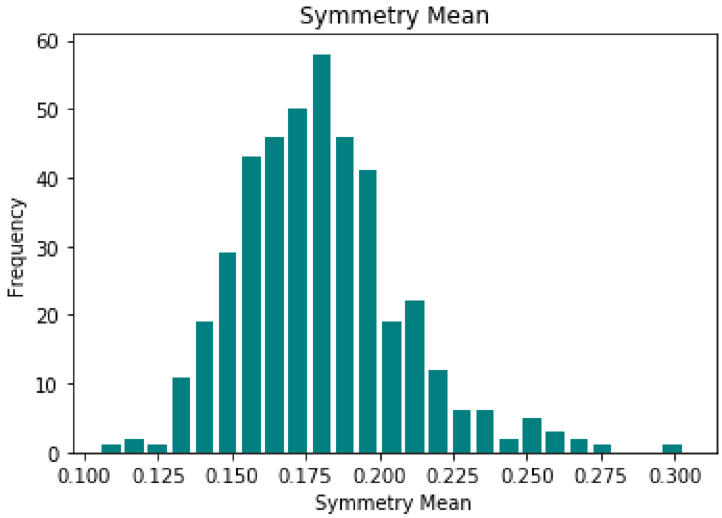

Area Mean
Frequency
0
20
40
60
80
500
1000
1500
2000
2500
Mean Area of the Core Tumor


Smoothness Mean
Frequency
0
10
20
30
40
50
0.06
0.08
0.10
0.12
0.14
Mean of Local Variation in Radius Lengths


Compactness Mean
Frequency
0
10
20
30
40
50
0.05
0.10
0.15
0.20
0.25
0.30
Mean of Perimeter^2 / area - 1.0


Cocavity Mean
Frequency
0
10
20
30
40
50
60
70
0.0
0.1
0.2
0.3
0.4
Mean of Severity of Concave Portions of the Contour





Now that we have a visual of the distribution for each variable, we can begin the process of running different models with different combinations of variables.

We will start by using all ten independent variables which each model, then try combinations with the most accurate model to see if we can increase accuracy.

Logistic Regression

Using logistic regression to predict the test set, the following results occurred:

Benign	Malignant
91 Correct Guesses	3 Incorrect Guesses (Type 2 Error)
2 Incorrect Guesses (Type 1 Error)	47 Correct Guesses

The overall accuracy was 97%, with a 94% accuracy in diagnosing malignant tumors.

K Nearest Neighbors

Benign	Malignant
85 Correct Guesses	5 Incorrect Guesses (Type 2 Error)
2 Incorrect Guesses (Type 1 Error)	53 Correct Guesses

The overall accuracy was 95%, with a 92% accuracy in diagnosing malignant tumors.

Support Vector Machine

Benign	Malignant
88 Correct Guesses	4 Incorrect Guesses (Type 2 Error)
4 Incorrect Guesses (Type 1 Error)	47 Correct Guesses

The overall accuracy was 94%, with a 92% accuracy in diagnosing malignant tumors.

Kernel Support Vector Machine

Benign	Malignant
90 Correct Guesses	3 Incorrect Guesses (Type 2 Error)
3 Incorrect Guesses (Type 1 Error)	47 Correct Guesses

The overall accuracy was 94%, with a 92% accuracy in diagnosing malignant tumors.

Naïve Beyes

Benign	Malignant
84 Correct Guesses	6 Incorrect Guesses (Type 2 Error)
9 Incorrect Guesses (Type 1 Error)	44 Correct Guesses

The overall accuracy was 89%, with an 88% accuracy in diagnosing malignant tumors.

Decision Tree

Benign	Malignant
93 Correct Guesses	3 Incorrect Guesses (Type 2 Error)
4 Incorrect Guesses (Type 1 Error)	43 Correct Guesses

The overall accuracy was 95%, with a 93% accuracy in diagnosing malignant tumors.

Random Forest

Benign	Malignant
81 Correct Guesses	5 Incorrect Guesses (Type 2 Error)
3 Incorrect Guesses (Type 1 Error)	54 Correct Guesses

The overall accuracy was 94%, with a 92% accuracy in diagnosing malignant tumors.

The best overall accuracy was from the logistic regression. The test was run 10 times with random selection, to test for overfitting and replicability:

The average scores are as follows:

Logistic Regression (Average of 10 models with consistent dataset)

Benign	Malignant
83.8 Correct Guesses	5 Incorrect Guesses (Type 2 Error)
5.1 Incorrect Guesses (Type 1 Error)	49 Correct Guesses

The overall accuracy was 93%, with a 90% accuracy in diagnosing malignant tumors.

Conclusion

While the overall accuracy was high for predication using all of the models, the accuracy for preventing a false diagnosis of benign tumor was only at 90 percent using the most accurate model. Because of this, we are unable to reject the null hypothesis. Factors that may diminish the likelihood of a type 2 error include having a larger dataset and pairing different regression models together. It may also improve results to remove some of the independent variables.