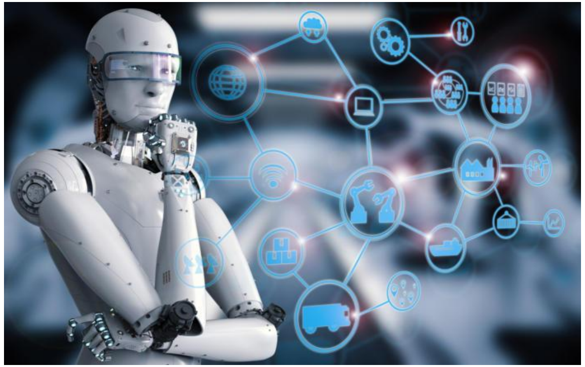
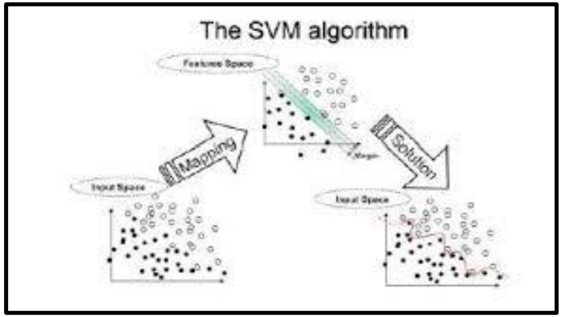
# MOSTEC Machine Learning Final Project

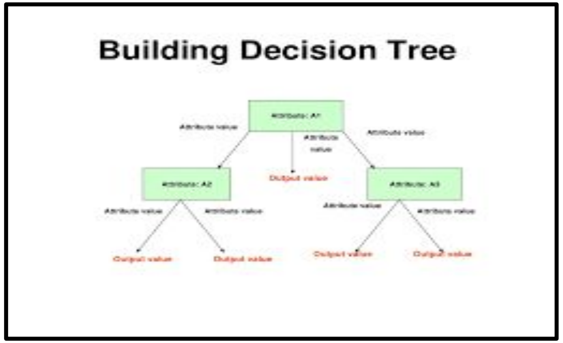
*By: AJ Arnolie and Mohammed Islam*

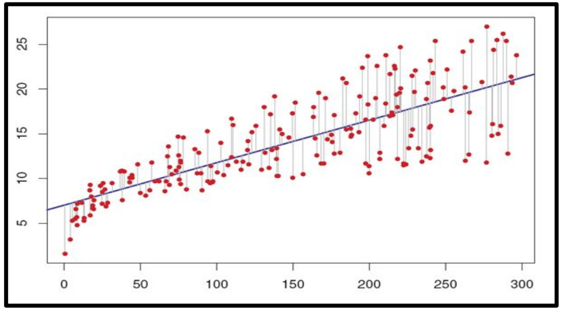
# What is Machine Learning?

- Subset of Artificial Intelligence
- Uses statistical techniques and methods to help machines "learn"
- Doesn't need to be directly programmed
- Similar to pattern recognition and computational learning theory









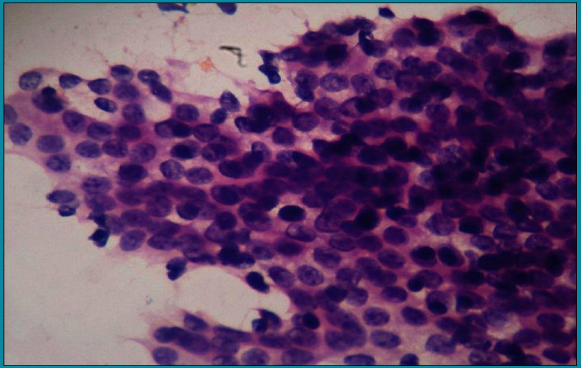
# Project Agenda

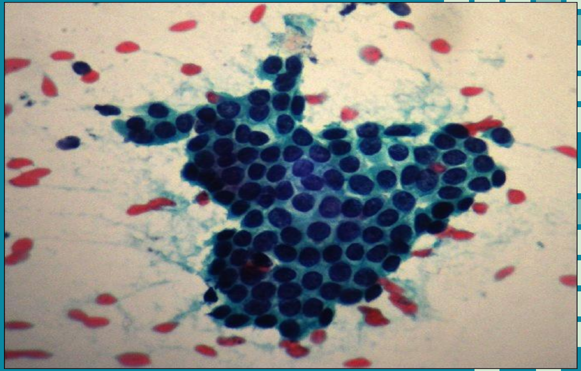
- **Four Trends Production / Analysis**
- **Linear Regression / Predictor Explanation and Discussion**
- **Categorical Techniques**
  - Logistic Regression / Support Vector Machines / Decision Trees
- **Explanation and Discussion**
- **Conclusion w/ Results**

# BREAST CANCER DATA SET

“

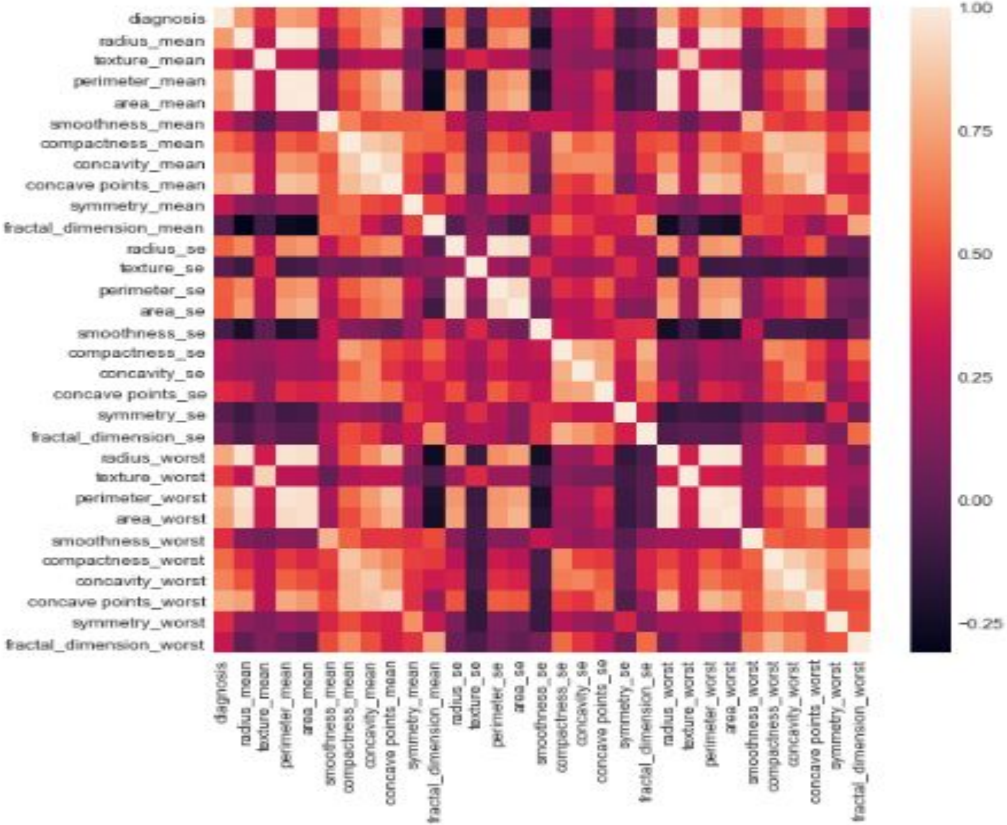
- Study from University of Wisconsin
- Over 600 entries with data based on observations of breast cancer lesions (masses of cells)
  - Variables: Area, Compactness, Diagnosis, Etc.
- Diagnosis is the feature we will be predicting





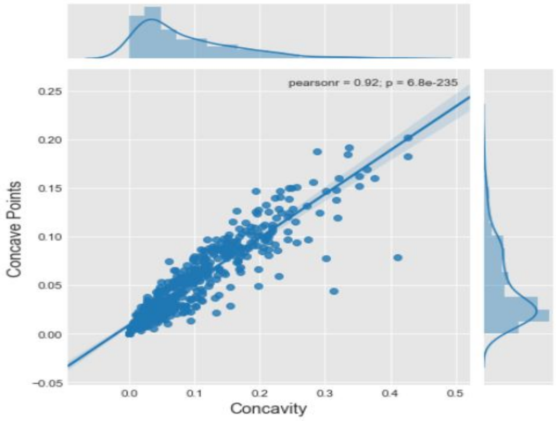
# Pre-Check/Improving Data for Analysis

- Removed unnecessary columns and in the "diagnosis" column, mapped M and B to dummy values 1 and 0
- Used the correlation heat map to select some of the variables with the highest correlation values



# Interesting Trends

### Concave Points vs Concavity

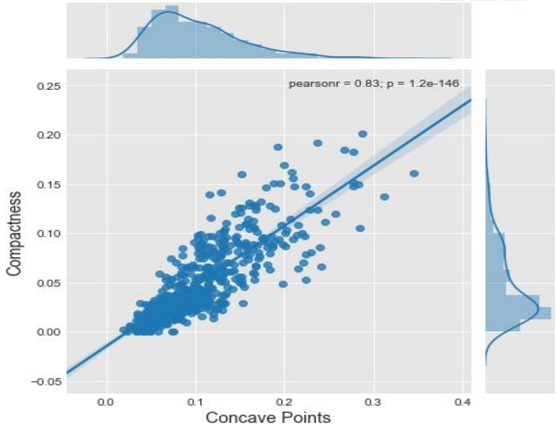


R = .92

*Strong Correlation between Concave Points and Concavity

*Higher R-value

### Compactness vs Concave Points

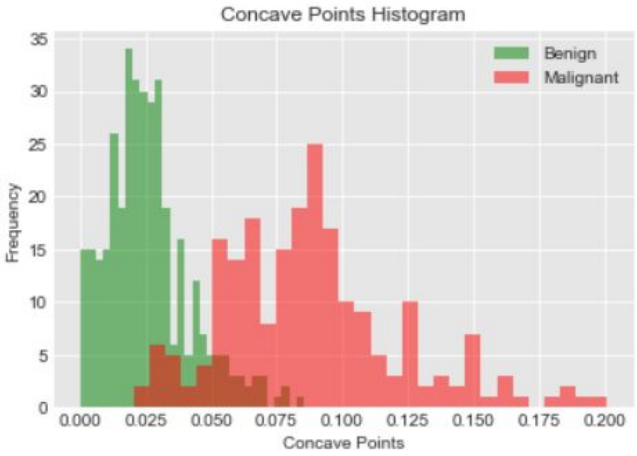


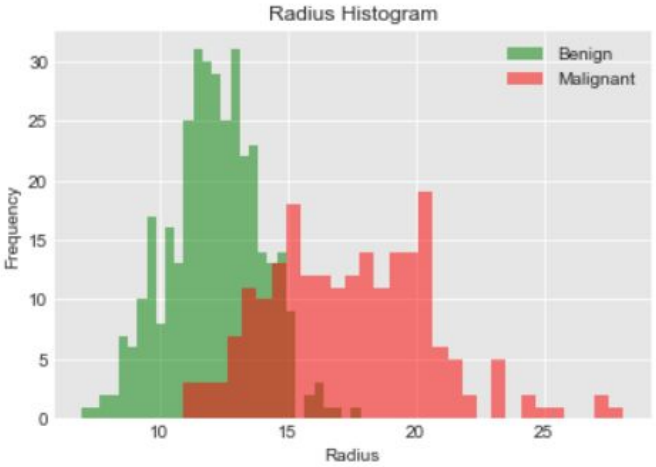
R = .83

*Good Correlation between Compactness and Concave Points

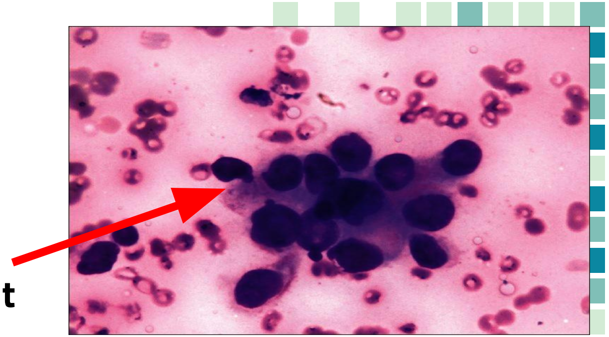
**Lower R-Value

# Interesting Trends Cont.



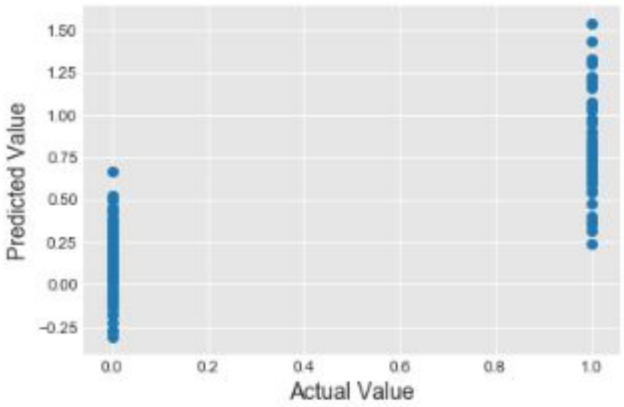


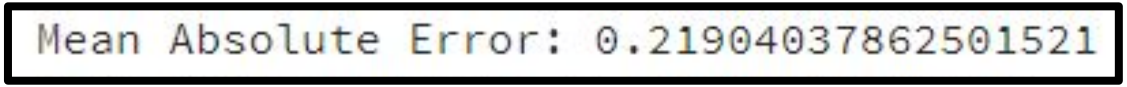
- Shows how higher values for both Concave Points and Radii usually give a Malignant diagnosis while lower values give a Benign diagnosis
- These were both useful features for our predictors



**Concave Point**

# Linear Regression of Data

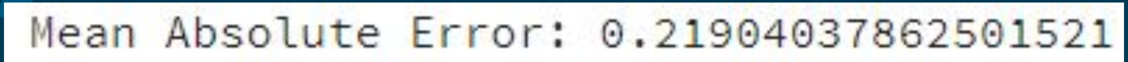




```
Mean Absolute Error: 0.21904037862501521
```

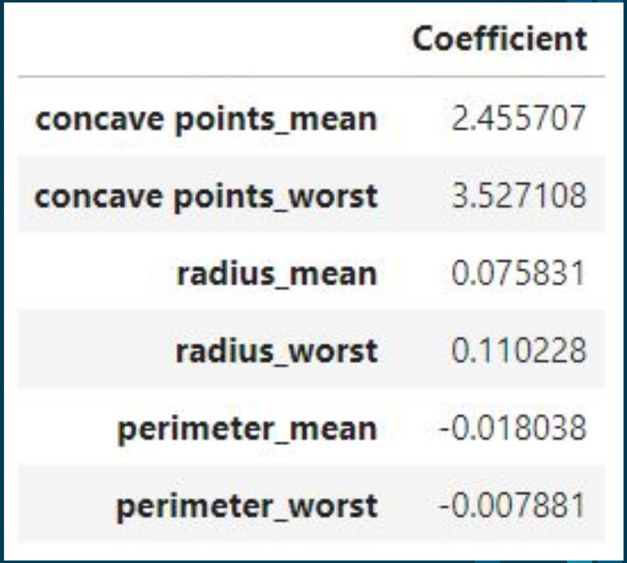
- Most basic regression technique in machine learning
- Tries to find a linear relationship between the dependent and independent variables
- Considering the range of our data is 1, the error is fairly low

# Linear Regression Analysis



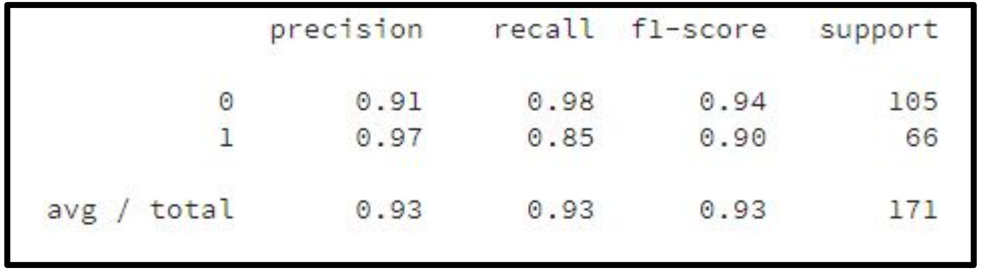
```
Mean Absolute Error: 0.21904037862501521
```

- Concave points are the most important in the regression process
- Larger the coefficient, the greater the effect on the final results
- Concave points have the largest coefficient and therefore affect the final result the most

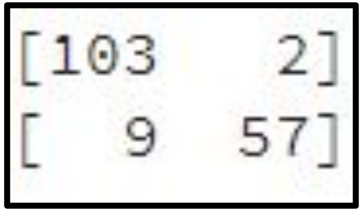


| | Coefficient |
|---|---|
| concave points_mean | 2.455707 |
| concave points_worst | 3.527108 |
| radius_mean | 0.075831 |
| radius_worst | 0.110228 |
| perimeter_mean | -0.018038 |
| perimeter_worst | -0.007881 |

# Logistic Regression of Data



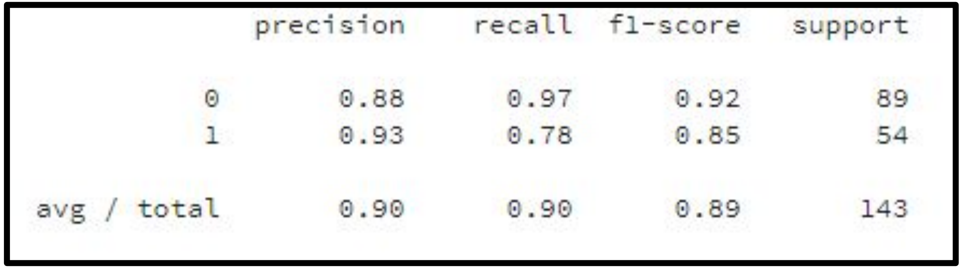
```
             precision    recall  f1-score   support

          0       0.91      0.98      0.94       105
          1       0.97      0.85      0.90        66

avg / total       0.93      0.93      0.93       171
```



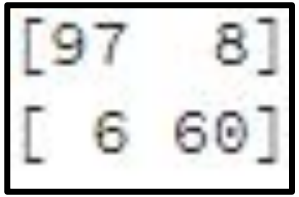
```
[103    2]
[  9   57]
```

- One of the most effective methods for binary classification in machine learning
- Describes relationship between one dependent binary variable and independent variables but uses logistic function
- Predictor gave consistent .93-.94 F1-Score

# Support Vector Machines



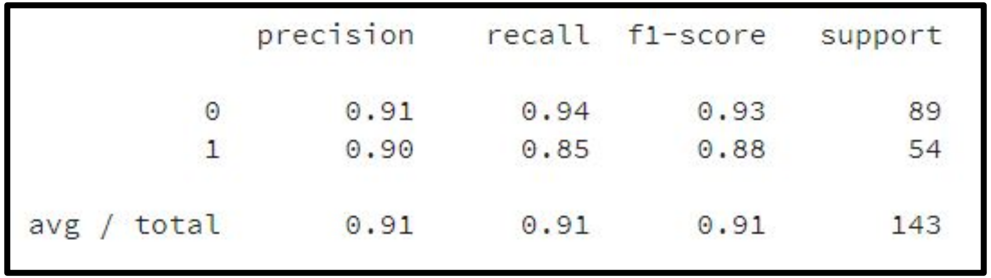
```
             precision    recall  f1-score   support

          0       0.88      0.97      0.92        89
          1       0.93      0.78      0.85        54

avg / total       0.90      0.90      0.89       143
```



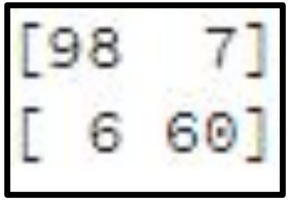
```
[97   8]
[ 6  60]
```

- Discriminative classifier used to act as a “seperation of classes” (distinguishing specific data entries)
- Uses vectors on 2-D coordinate plane to determine a hyperplane line  (line of separation) between the datasets
- Predictor resulted in F1-Score between 0.89 & 0.94

# Decision Trees



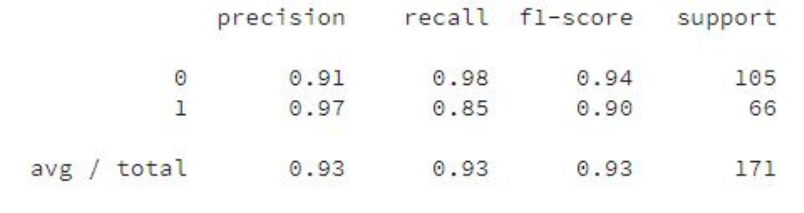
```
              precision    recall  f1-score   support

          0       0.91      0.94      0.93        89
          1       0.90      0.85      0.88        54

avg / total       0.91      0.91      0.91       143
```



```
[98  7]
[ 6 60]
```

- Can be used for both classification and regression
- Tree-like decision making process
- Makes sequential, hierarchical decisions about outcomes based on predictor data until a result is reached
- F1-Scores resulted between two intervals: 0.90-0.94 (mostly) and 0.96-0.97 (occasionally)

# Categorical Feature Analysis



```
             precision    recall  f1-score   support

          0       0.91      0.98      0.94       105
          1       0.97      0.85      0.90        66

avg / total       0.93      0.93      0.93       171
```

- **Logistic Regression** was most effective and consistent
- Advantages
  - Output is easier to interpret
  - Can be updated easily
- Disadvantages
  - Usually requires more data to achieve stable results
  - More dependent on the chosen independent variables
  - Can be overfitted

# THANK YOU!

Any questions?

MOSTEC