Breast Cancer Predictions

By: Richney Chin-Chap

Problem Statement:

- Breast cancer is one of most common cancers in women after skin cancer and is the second leading cause of death in women.
- Since 2007, breast cancer death rates have been steady in women younger than 50 but have continued to decrease in older women.
- From 2013 to 2018, the death rate went down by 1% per year.

Goal

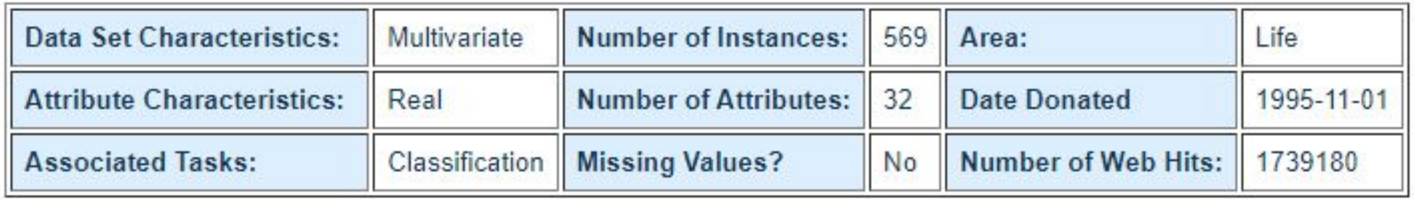
Our goal is to create an accurate predictor based off current symptoms to detect if a diagnosis is malignant or benign, to aid in future diagnosis and decrease the number of breast cancer deaths.

Who Would This Benefit?

- Women
- Doctors
- Medical Experts
- Researchers
- Hospitals

Looking at the Data:

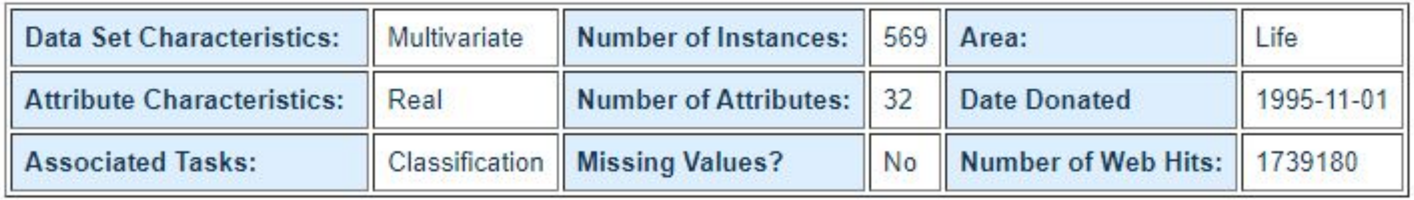
- The data set used for this report is the Breast Cancer Wisconsin Diagnostic Data Set, provided by the University of Irvine Machine Learning Repository



Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1739180

Looking at the Data:

- No missing values
- 569 rows
- 32 columns



Data Set Characteristics:	Multivariate	Number of Instances:	569	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	32	Date Donated	1995-11-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1739180

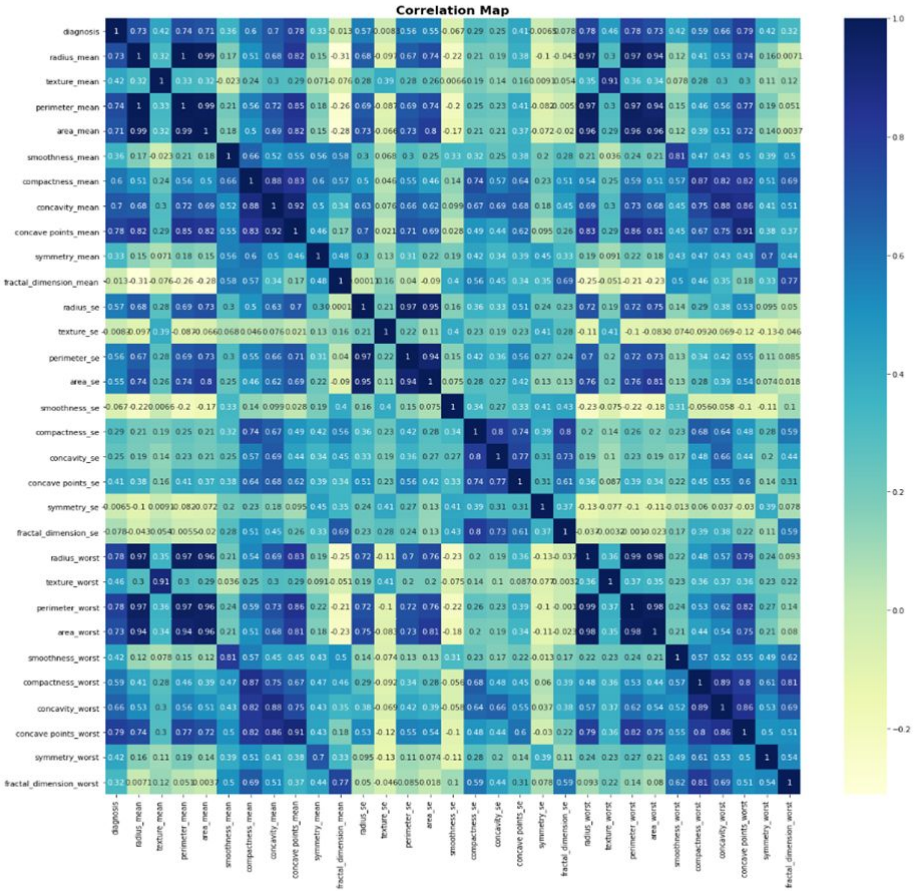
Exploring the Data

Main features to explore:

- ID number
- Diagnosis (M = malignant, B = benign)
- Radius (mean of distances from center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- Smoothness
- Compactness
- Symmetry
- Fractal Dimension

Correlation Map

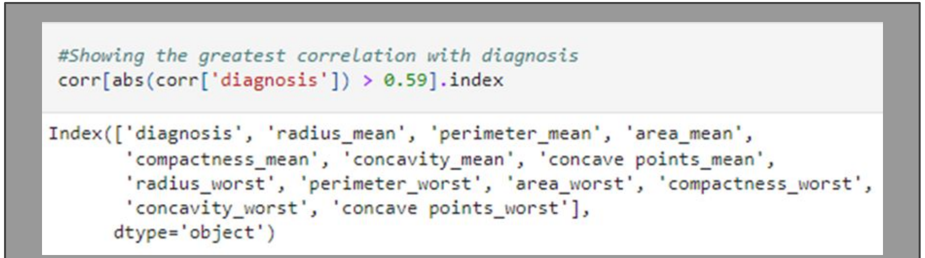
In order to see how each feature is correlated, let's take a look at a correlation map to gain a general sense of which features correlated the most with our target, "diagnosis". Features in darker blue will have a higher correlation, whereas features that are yellow will have the lowest correlation.



Correlated Features:

Out of the 32 features, the following features along with their mean and worst were revealed to have the highest correlation to our target, “diagnosis”:

- Radius
- Perimeter
- Area
- Compactness
- Concavity
- Concave Points



```
#Showing the greatest correlation with diagnosis
corr[abs(corr['diagnosis']) > 0.59].index
```

```
Index(['diagnosis', 'radius_mean', 'perimeter_mean', 'area_mean',
       'compactness_mean', 'concavity_mean', 'concave points_mean',
       'radius_worst', 'perimeter_worst', 'area_worst', 'compactness_worst',
       'concavity_worst', 'concave points_worst'],
      dtype='object')
```

Python code used to determine which features had the highest correlation to “diagnosis”.

Model Selection:

Each model was scored based on accuracy and the models are sorted from highest to lowest accuracy:

Model	**Score**
Support Vector Machines	0.982456
Logistic Regression	0.976608
KNN	0.964912
Random Forest	0.959064
Naive Bayes	0.900585

From this table it appears as though the Support Vector Machines and Logistic Regression Model have the highest accuracy of around 98%.

Key Takeaways and Future Work:

- From the 32 features that were analyzed in this process, the most highly correlated features are related to the size of the cell nucleus.
- Possible deeper analysis on the relationships between each feature.
- Check for any possible multicollinearity between the features to remove any unnecessary features and improve predictions.
- After improving predictions, present data to doctors or medical experts to aid them in improving breast cancer detection in patients.

Thank you!