Final Report:
Breast Cancer Predictor

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. These decreases are believed to be the result of finding breast cancer earlier through screening and increased awareness, as well as better treatments. 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.

Data Cleaning and Data Wrangling:

The data set used for this report is the Breast Cancer Wisconsin Diagnosis Data Set, provided by the University of Irvine Machine Learning Repository. It contains 569 rows and 32 columns with no missing values. With a decently sized data set, no missing values, no outliers, and relatively clean data, one change that was made was to change the diagnosis column to be numerical rather than categorical.

Exploratory Data Analysis:

With 32 different features to explore, I wanted to see which of those features have the highest correlation with our diagnosis.

The 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 (local variation in radius length)
- compactness (number of concave portions of the contour)
- symmetry
- fractal dimension

NOTE: The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

In order to see how each feature is correalted, 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, wheras features that are yellow will have the lowest correlation.

Figure 1: Correlation graph

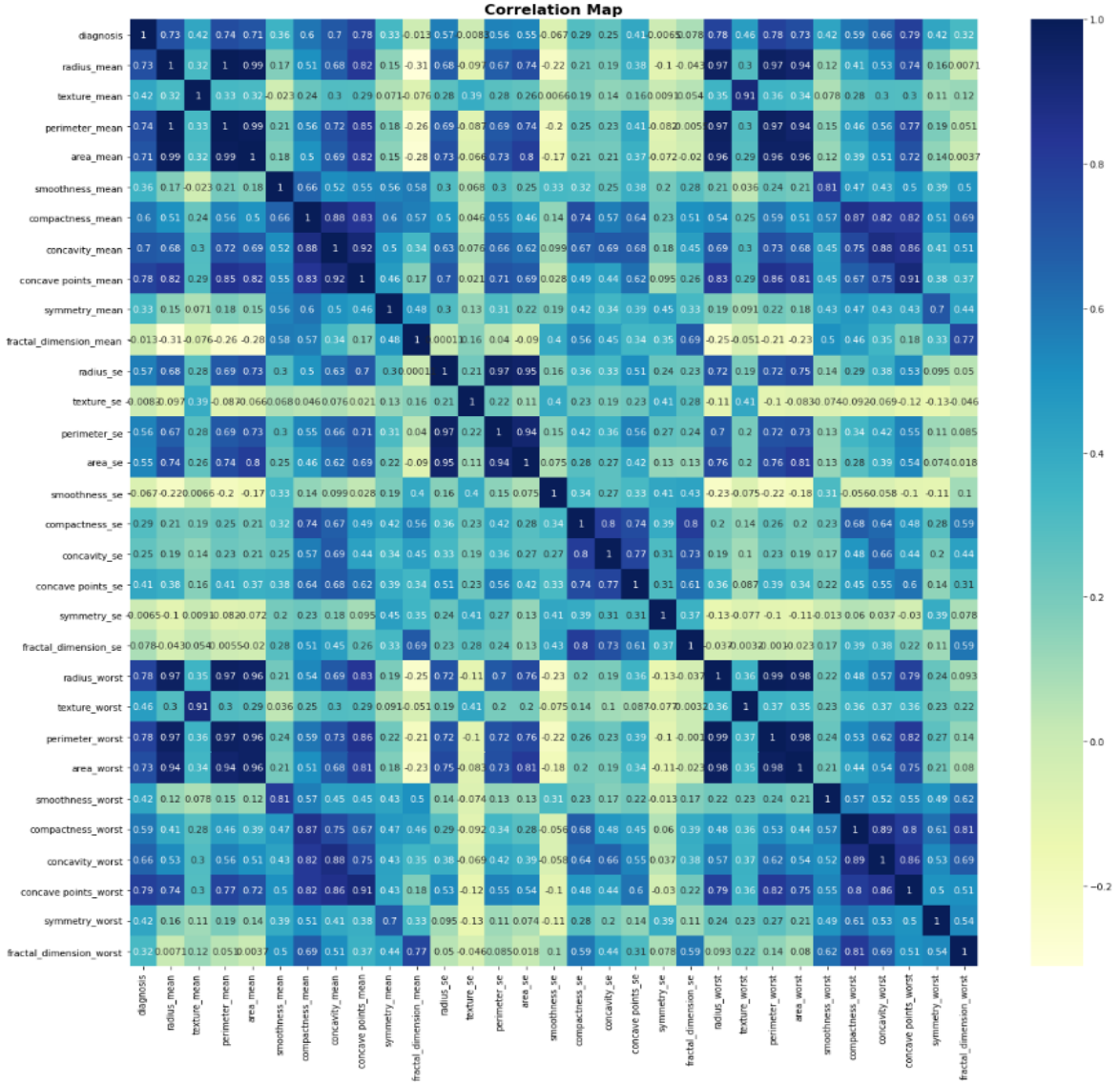

Figure 2: Determining which features out of the 32 are most correlated to "diagnosis"

```
#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')
```

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

We can see how size of the cell nucleus correlates the most when receiving a diagnosis.

Model Selection:

5 different models were made when deciding which model would best produce an accurate prediction. Those models were: Logistic Regression, KNN Classification, Random Forest, Support Vector Machines, and Naïve Bayes. The metric used when building the models was accuracy. How accurately can each model predict whether a diagnosis will be malignant or benign?

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

Figure 3: Comparison Table of Different Models Accuracy

	Model	Score
3	Support Vector Machines	0.982456
0	Logistic Regression	0.976608
1	KNN	0.964912
2	Random Forest	0.959064
4	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.