Keiuntae Smith
DSC680
Applied Data Science
21 Sep 2022

# Project 1: Milestone 3

**Topic:**

The topic of choice for my first project is to predict breast cancer through machine learning tactics by using breast cancer data.

**Business Problem:**

In my lifetime, I have had significant loss due to cancer. My mother, aunt, and wife all lost their battles with breast cancer. Breast cancer is one of the most common cancers among women worldwide and it represents the majority of new cancer cases and cancer-related deaths according to global statistics. This makes it a significant public health problem in today's society. Some of the risk factors for breast cancer are age, personal and family history of breast cancer, genetic factors, and childbearing history (Goel, 2018). An early diagnosis of breast cancer can vastly improve the prognosis and chance of survival because it can foster timely clinical treatment for patients in efforts.

**Datasets:**

I obtained my data through the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (Dua & Graff, 2017).

The dataset is a Wisconsin breast cancer Diagnostic data set in which the 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. The attributes of dataset are as follows:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

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)

**Methods and Analysis:**

I did an initial EDA on the dataset to explore what the dataset is comprised of so that I can analyze it and make predictions using machine learning. To begin this process, I used 'df.shape' to get the number of rows and columns within the dataset which were 569 columns and 32 rows. This let me know that there was data on 569 patients. Next, I wanted to see how many patients were diagnosed with breast cancer by examining the diagnosis column to distinguish between malignant and benign cancerous cells.

***See Appendix, Figure 1: Bar Chart of Diagnosed Patients***

I then assessed the data types of each column and noticed that the 'id' column was an "int64" and the 'diagnosis' was a dtype of "object". To rectify this, I chose to eliminate the 'id' column due to irrelevance and encoded the 'diagnosis' column to integer type for analytic purposes. After cleaning the data to my satisfaction, I then wanted to see the correlation between the columns by creating a heatmap using seaborn.

***See Appendix, Figure 2: Correlation Heatmap of Columns***

The heatmap displays the highest correlation features to a cancer diagnosis are:

| | |
|---|---|
| Radius_mean | 0.73 |
| Perimeter_mean | 0.74 |
| Area_mean | 0.70 |
| Concave points_mean | 0.77 |
| Radius_worst | 0.77 |
| Perimeter_worst | 0.78 |
| Area_worst | 0.73 |
| Concave points_worst | 0.79 |

I also wanted to do a bar chart of the 30 features to show correlation between diagnosis using the corrwith function.

***See Appendix, Figure 3: Bar Chart of Features***

Using the Correlation bar chart and correlation heatmap, I wanted to take a closer look at the features with the highest correlation by setting the threshold to 0.60 and create a special Threshold Heatmap.

***See Appendix, Figure 4: Threshold Heatmap***

Due to the fact that my topic of breast prediction is a classification problem, I decided to attempt several different classification algorithms in machine learning. These algorithms include:

1. Logistic Regression: estimates the probability of an event occurring based on a given dataset of independent variables (IBM, n.d.).

2. Nearest Neighbor: approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

3. Support Vector Machines: creates a line or a hyperplane which separates the data into classes.

4. Decision Tree Algorithm: supervised learning algorithm, which is utilized for both classification and regression tasks (IBM, n.d.).

5. Random Forest Classification: consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest produces a class prediction and the class with the most votes becomes the model's prediction (Yiu, 2019).

To begin the modeling process, I had to split the data: 75% for training and 25% for testing. Once split, I scaled the data using StandardScalar(). This is a necessary step prior to modeling because different scales of the data features affect the modeling of a dataset adversely. Finally, it was time to fit the data to the models of choice in order to see which one performed the best.

**Conclusions:**

After applying and fitting the models, I chose to use the classification accuracy method to evaluate the performance of the models implemented. Classification accuracy is the ratio of number of correct predictions to the total number of input samples. The results yielded the following:

| | |
|---|---|
| LogisticRegression | 0.9790209790209791 |
| KNeighborsClassifier | 0.958041958041958 |
| SVC | 0.972027972027972 |
| DecisionTreeClassifier | 0.9300699300699301 |
| RandomForestClassifier | 0.965034965034965 |

To my surprise, the Logistic Regression algorithm performed the best of the 5 algorithms chosen. I also chose to display these results using a seaborn bar plot for visualization purposes.

***See Appendix, Figure 5: Horizontal Bar Chart of Algorithm Results***

**Ethical Considerations:**

One of the main ethical concerns for completing an analysis of this caliber is of course the possibility of patient information getting out into the public. It is important to safeguard their information to shield them from nefarious actors whom might want to use their information for other motives.

**Challenges/Issues**:

One of the concerns I am afraid of encountering is overfitting. This is one of the most common problems faced by Machine Learning engineers and data scientists. Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. This can ultimately affect the performance of the model negatively.

Keiuntae Smith
DSC680
Applied Data Science
21 Sep 2022

**Future Uses/Additional Applications**

I believe that this method is a great opportunity to predict several other cancers as well as other medical diagnosis. In particular, this could definitely be implemented in predicting childhood autism through the use of an in-depth screening dataset.

## References

Dua, D., & Graff, C. (2017). *{UCI} Machine Learning Repository*. Retrieved from Breast Cancer Wisconsin (Diagnostic) Data Set: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Goel, V. (2018, September 29). *Building a Simple Machine Learning Model on Breast Cancer Data*. Retrieved from Towards Data Science: https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

IBM. (n.d.). *What is a Decision Tree?* Retrieved from IBM: https://www.ibm.com/topics/decision-trees

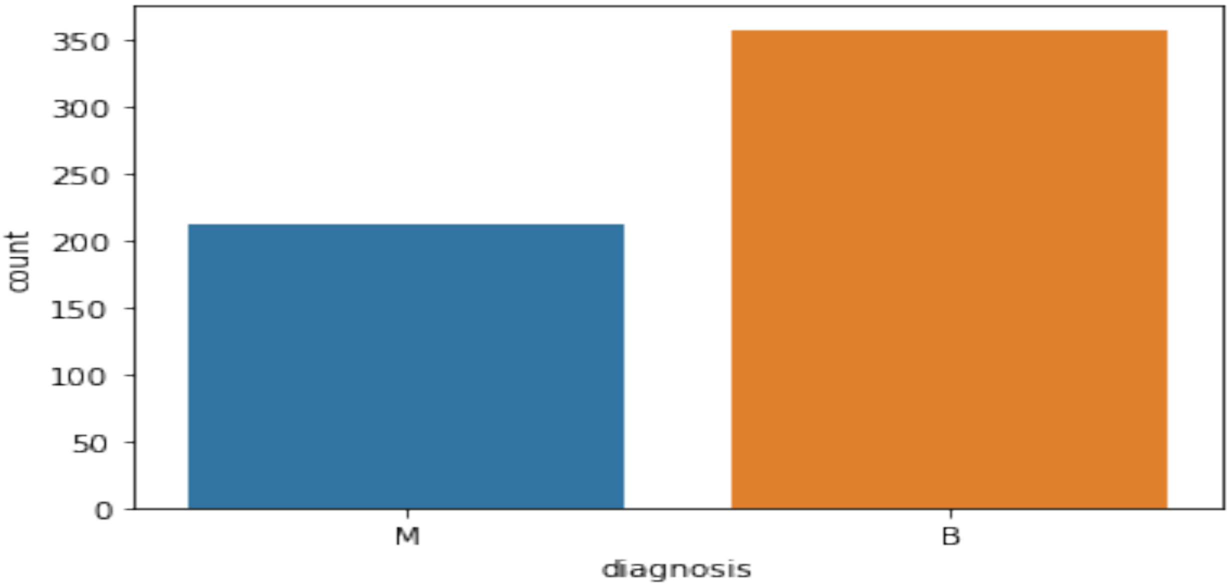
IBM. (n.d.). *What is logistic regression?* Retrieved from IBM: https://www.ibm.com/topics/logistic-regression

Yiu, T. (2019, June 12). *Understanding Random Forest*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

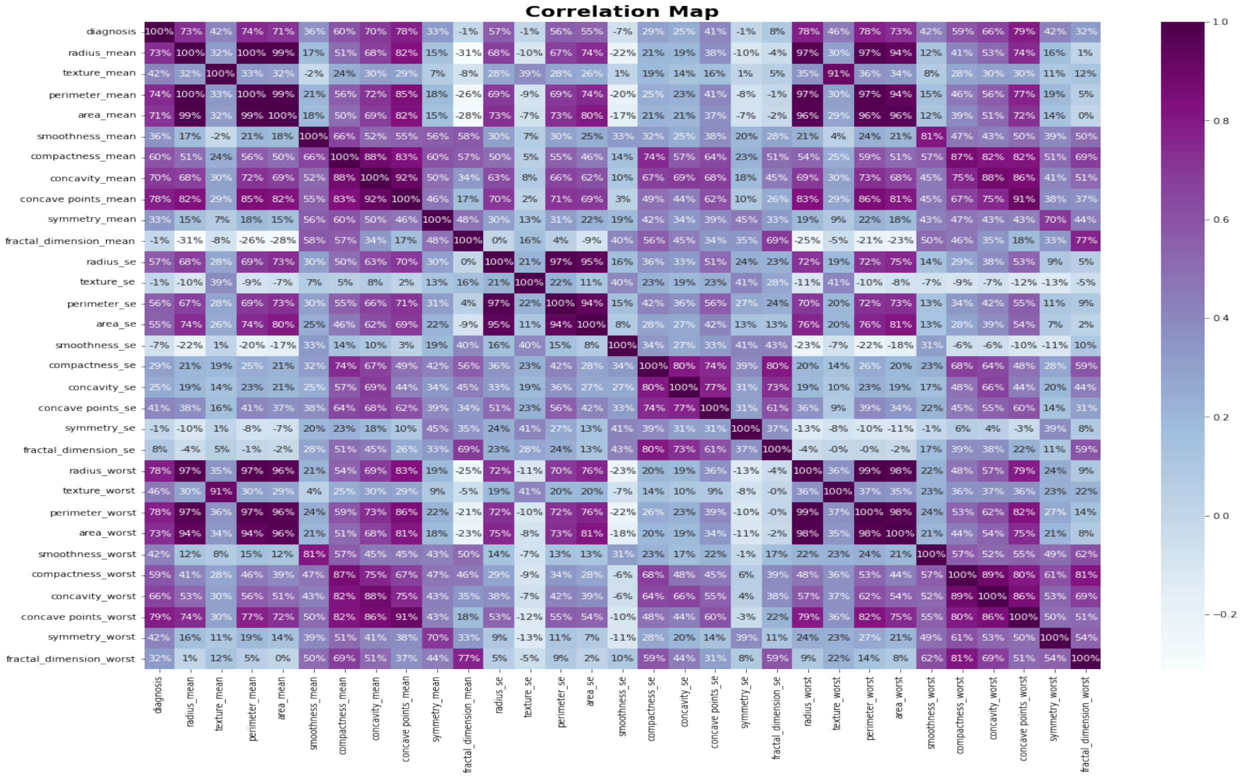
Keiuntae Smith
DSC680
Applied Data Science
21 Sep 2022

# Appendix

### Figure 1: Bar Chart of Diagnosed Cancer Patients

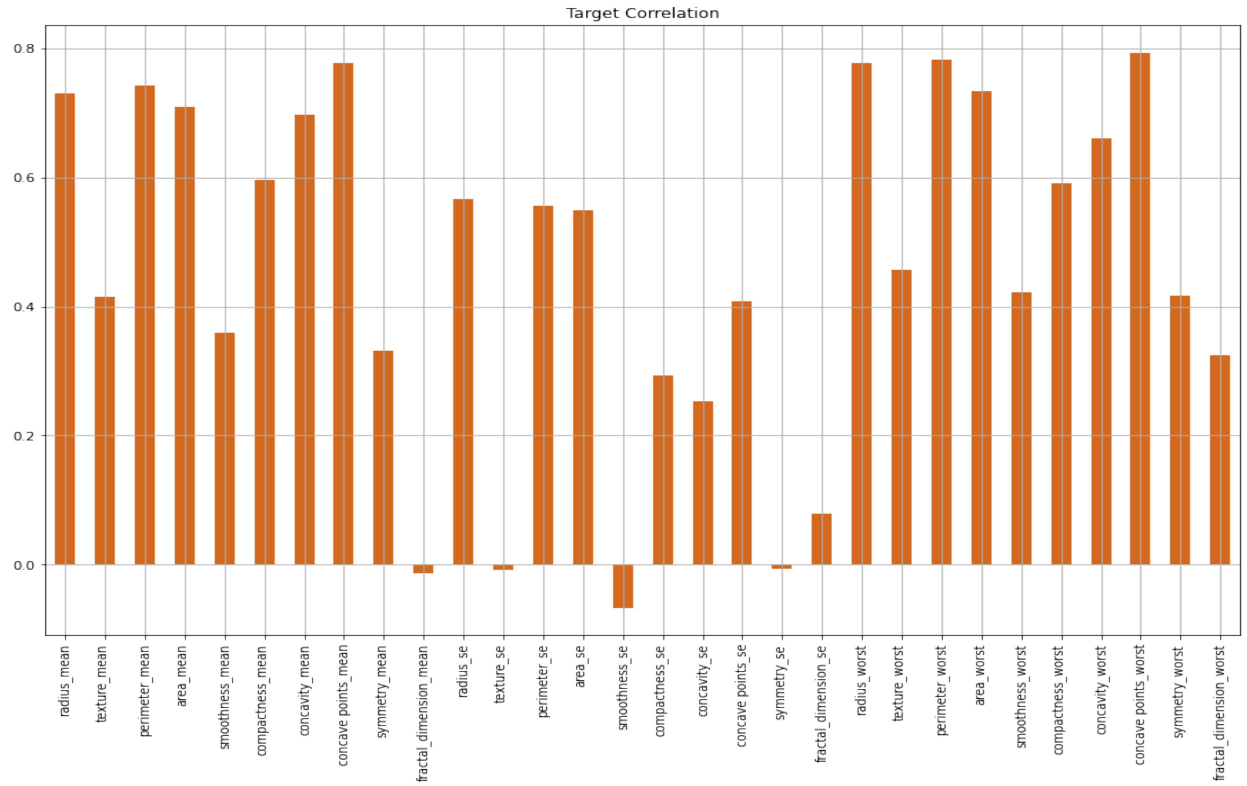


### Figure 2: Correlation Heatmap of Columns

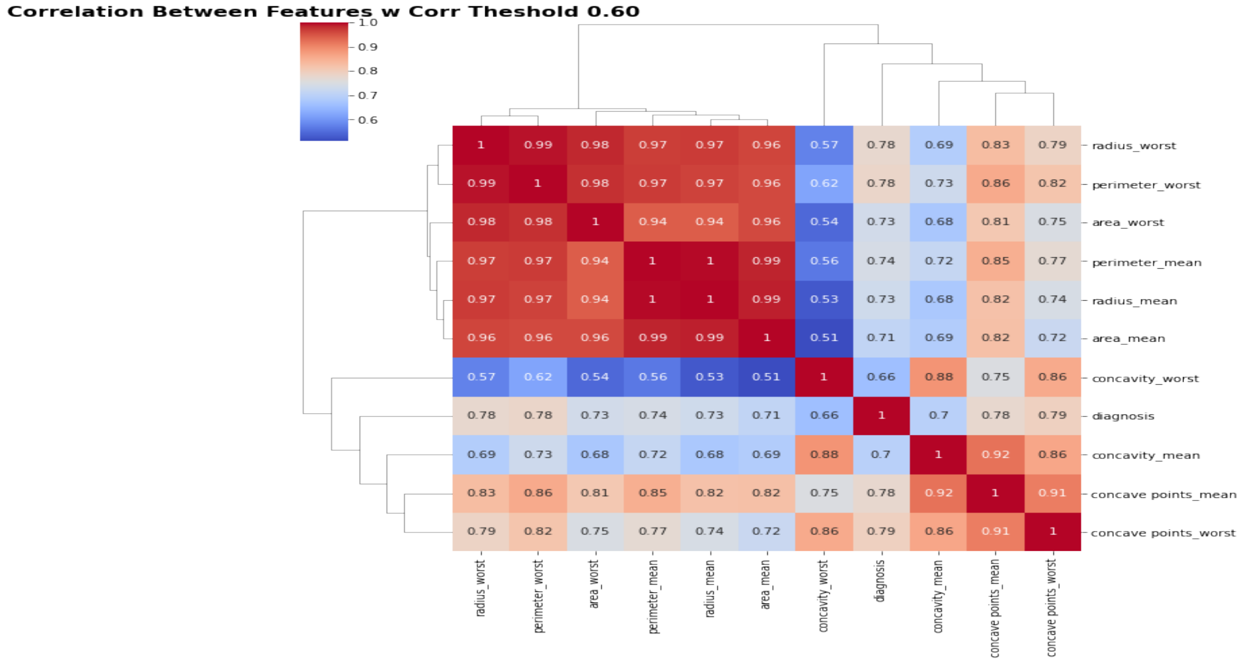


Keiuntae Smith
DSC680
Applied Data Science
21 Sep 2022

## Figure 3: Bar Chart of Features



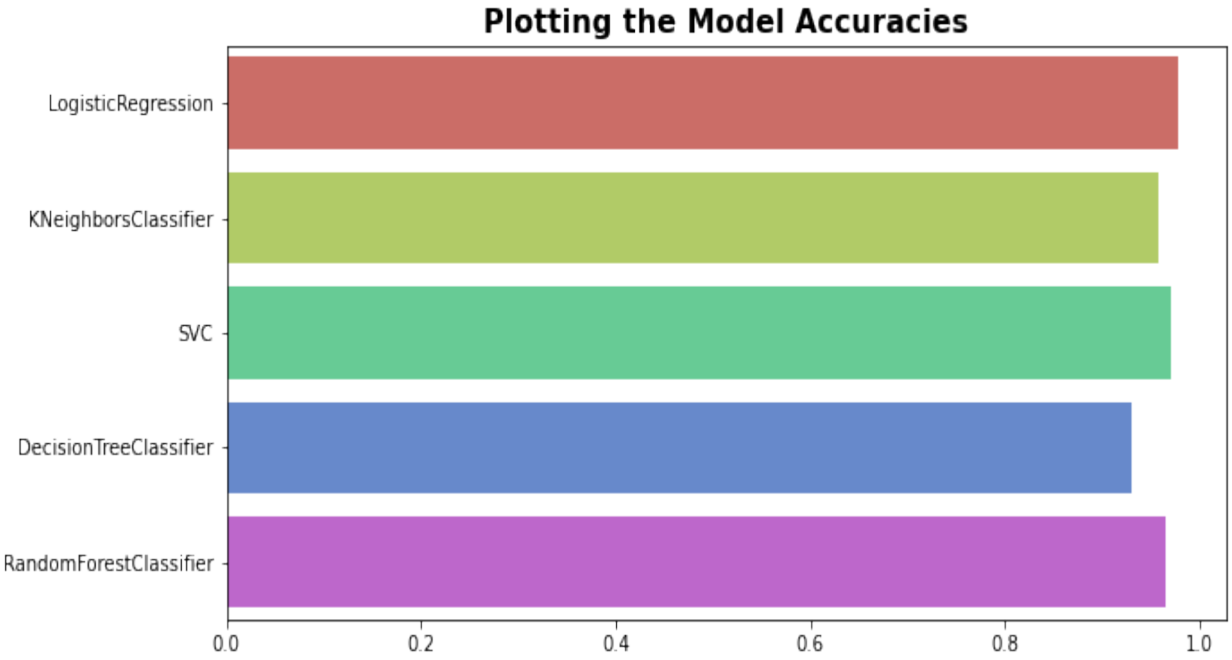
## Figure 4: Threshold Heatmap



| radius_worst | perimeter_worst | area_worst | perimeter_mean | radius_mean | area_mean | concavity_worst | diagnosis | concavity_mean | concave points_mean | concave points_worst | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.99 | 0.98 | 0.97 | 0.97 | 0.96 | 0.57 | 0.78 | 0.69 | 0.83 | 0.79 | radius_worst |
| 0.99 | 1 | 0.98 | 0.97 | 0.97 | 0.96 | 0.62 | 0.78 | 0.73 | 0.86 | 0.82 | perimeter_worst |
| 0.98 | 0.98 | 1 | 0.94 | 0.94 | 0.96 | 0.54 | 0.73 | 0.68 | 0.81 | 0.75 | area_worst |
| 0.97 | 0.97 | 0.94 | 1 | 1 | 0.99 | 0.56 | 0.74 | 0.72 | 0.85 | 0.77 | perimeter_mean |
| 0.97 | 0.97 | 0.94 | 1 | 1 | 0.99 | 0.53 | 0.73 | 0.68 | 0.82 | 0.74 | radius_mean |
| 0.96 | 0.96 | 0.96 | 0.99 | 0.99 | 1 | 0.51 | 0.71 | 0.69 | 0.82 | 0.72 | area_mean |
| 0.57 | 0.62 | 0.54 | 0.56 | 0.53 | 0.51 | 1 | 0.66 | 0.88 | 0.75 | 0.86 | concavity_worst |
| 0.78 | 0.78 | 0.73 | 0.74 | 0.73 | 0.71 | 0.66 | 1 | 0.7 | 0.78 | 0.79 | diagnosis |
| 0.69 | 0.73 | 0.68 | 0.72 | 0.68 | 0.69 | 0.88 | 0.7 | 1 | 0.92 | 0.86 | concavity_mean |
| 0.83 | 0.86 | 0.81 | 0.85 | 0.82 | 0.82 | 0.75 | 0.78 | 0.92 | 1 | 0.91 | concave points_mean |
| 0.79 | 0.82 | 0.75 | 0.77 | 0.74 | 0.72 | 0.86 | 0.79 | 0.86 | 0.91 | 1 | concave points_worst |

Keiuntae Smith
DSC680
Applied Data Science
21 Sep 2022

**Figure 5: Horizontal Bar Chart of Algorithm Results**


Plotting the Model Accuracies
LogisticRegression
KNeighborsClassifier
SVC
DecisionTreeClassifier
RandomForestClassifier