Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022

Milestone 2

What is the topic I chose?

The topic that I chose for this project is breast cancer. Breast cancer is often an aggressive form of cancer in which cells begin to grow out of control starting in either one or both breasts. It is also important to understand that not all cases of breast lumps are on occasion not cancerous otherwise known as benign. Tumors that are shown to grow at a rate faster than normal or show abnormalities after a biopsy are considered malignant, or cancerous in nature. Breast cancer is the most common form of cancer found amongst women in the world and has affected over 2.1 million people in 2015 alone. Some of the key challenges found in diagnosing breast cancer are whether a tumor is considered malignant (meaning that it is cancerous in nature caused by inherited genetic mutations to certain genes such as the BRCA1 and BRCA2), density of fatty tissue within the chest and breast areas, a family history of breast or ovarian cancer, or previous treatment using radiation therapy. There are also cases in which men have also been diagnosed with breast cancer, but not nearly as common as other forms of cancer that affect men more prevalently.

What types of model or models do you plan to use and why?

The datasets that I plan on using for this project are the following:

- https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
- https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022

- https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset

I chose these three datasets as they are all related to each other, and I did find an issue with finding data using the Centers for Disease Control statistics board. The issue that I found with them was that many of the referenced data has either been moved, purged, or not backlinked to which I could find the data sets for reference under CDC guidelines. The data found instead is relevant, as breast cancer datasets are prevalent compared to other cancer sets, and breast cancer accounts for nearly 25 percent of all cancer cases. I find these datasets to be relevant to the topic at hand, and though I think the datasets might be similar, they should be different enough to give a broad understanding of what the data means.

How do you plan to evaluate your results?

Evaluation of results will have to be a process of taking the data into account, and removing all the duplicate data, or data that is too similar to each other in which could produce a significant duplication of data. Evaluation of the data results should provide unique results each time. The datasets should provide a meaningful response to the number of women and men who are diagnosed with malignant or benign forms of cancer, give insight on the radius, texture, and other statistical measures.

What do you hope to learn?

Based on the datasets, I am hoping to learn the effect of race, age, and range of women and men who are diagnosed with breast cancer and their survival rate, the chances for metastasizing cells, and for a potential early warning detection of breast cancer, though I also do not believe that this scope is within the project requirements.

Assess any risks with your proposal.

The largest risk of any proposal is to find out that the data is not either current, or relevant to the topic at hand. I believe that the topic that I have chosen is a worthwhile endeavor to explore, even if the data is not as current as it could be. I will continue to look for other data sets that might fit in the scope of the project, while keeping on track with producing clear and meaningful results from the current data sets. Data is the biggest driver for success when it comes to this project, and I hope that once all the data is cleaned and prepared that it will produce meaningful results.

Identify a contingency plan if your original project plan does not work out.

My other options which still be within the realm of cancer, but the other main forms of cancer that are out there that have quite a large data set for them would be considered prostate cancer. I plan on collecting the data sets for these particular forms of cancer as well, just as a backup plan under the contingency that if my breast cancer research does not wind up producing fruitful nor positive results for this project, I would be able to pivot quickly towards the other data sets without losing too much time. The data for prostate cancer seems to be less as prevalent in terms of data sets than breast cancer, although prostate cancer is the most common form of cancer for men except for skin cancers, which seem to be based on cursory research, indiscriminate of either males or females, and effect both sexes equally.

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022

Milestone 3

Preliminary Analysis

Will I be able to answer the questions I want to answer with the data I have?

With the data set that I have collected I believe that I can answer all the questions that are proposed from milestone two. I do not believe that I will run into any issues being able to answer any of the questions provided above.

What visualizations are especially useful for explaining my data?

The following are going to be work well to explain my data:

- Bar graphs
- Histogram Graphs
- Spectral clustering
- Box Plot Graphs

Do I need to adjust the data and/or driving questions?

To my knowledge and understanding at this point, there is no need to adjust the data or driving questions to accommodate for my original questions. There will be a need to combine and manipulate the data into a better organized CSV type file but that will not change the trajectory of the other milestones.

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022

Do I need to adjust my model/evaluation choices?

No, my model will fit and will work for the questions and proposal presented in Milestone 2.

Are my original expectations still reasonable?

Yes, my original expectations are still reasonable. I do not consider there to be any issues that arise from my original expectations.

Milestone 4

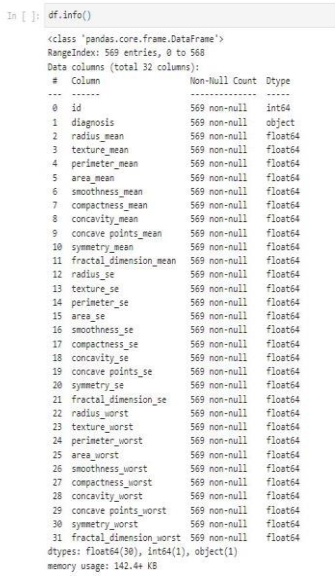
- Explain your process for prepping the data
- Build and evaluate at least one model
- Interpret your results
- Begin to formulate a conclusion/recommendations

Please submit Milestone 4 in Blackboard under the group submission link.

This should be submitted through the group assignment submission regardless if it is an independent project or multi-person group.

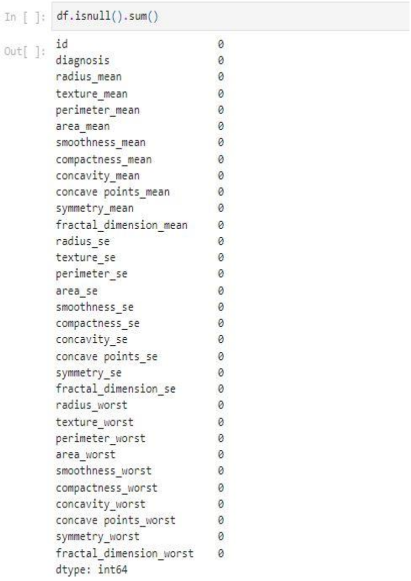
My first attempt at cleaning the data was to first understand the data that I had received from the data set. I began breaking down the data by looking at its shape, columns, and describing the data using the describe function from Python.

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022



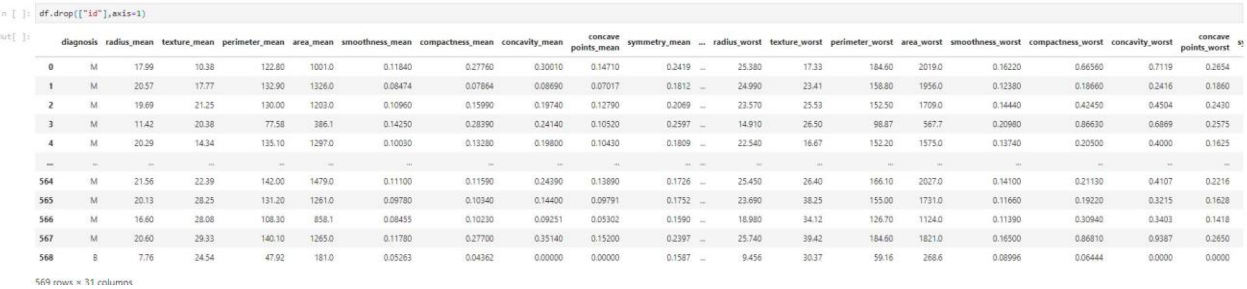
```
In [ ]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   id                       569 non-null    int64  
 1   diagnosis                569 non-null    object 
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), int64(1), object(1)
memory usage: 142.4+ KB
```



```
In [ ]: df.isnull().sum()

Out[ ]: id                         0
        diagnosis                  0
        radius_mean                0
        texture_mean               0
        perimeter_mean             0
        area_mean                  0
        smoothness_mean            0
        compactness_mean           0
        concavity_mean             0
        concave points_mean        0
        symmetry_mean              0
        fractal_dimension_mean     0
        radius_se                  0
        texture_se                 0
        perimeter_se               0
        area_se                    0
        smoothness_se              0
        compactness_se             0
        concavity_se               0
        concave points_se          0
        symmetry_se                0
        fractal_dimension_se       0
        radius_worst               0
        texture_worst              0
        perimeter_worst            0
        area_worst                 0
        smoothness_worst           0
        compactness_worst          0
        concavity_worst            0
        concave points_worst       0
        symmetry_worst             0
        fractal_dimension_worst    0
        dtype: int64
```



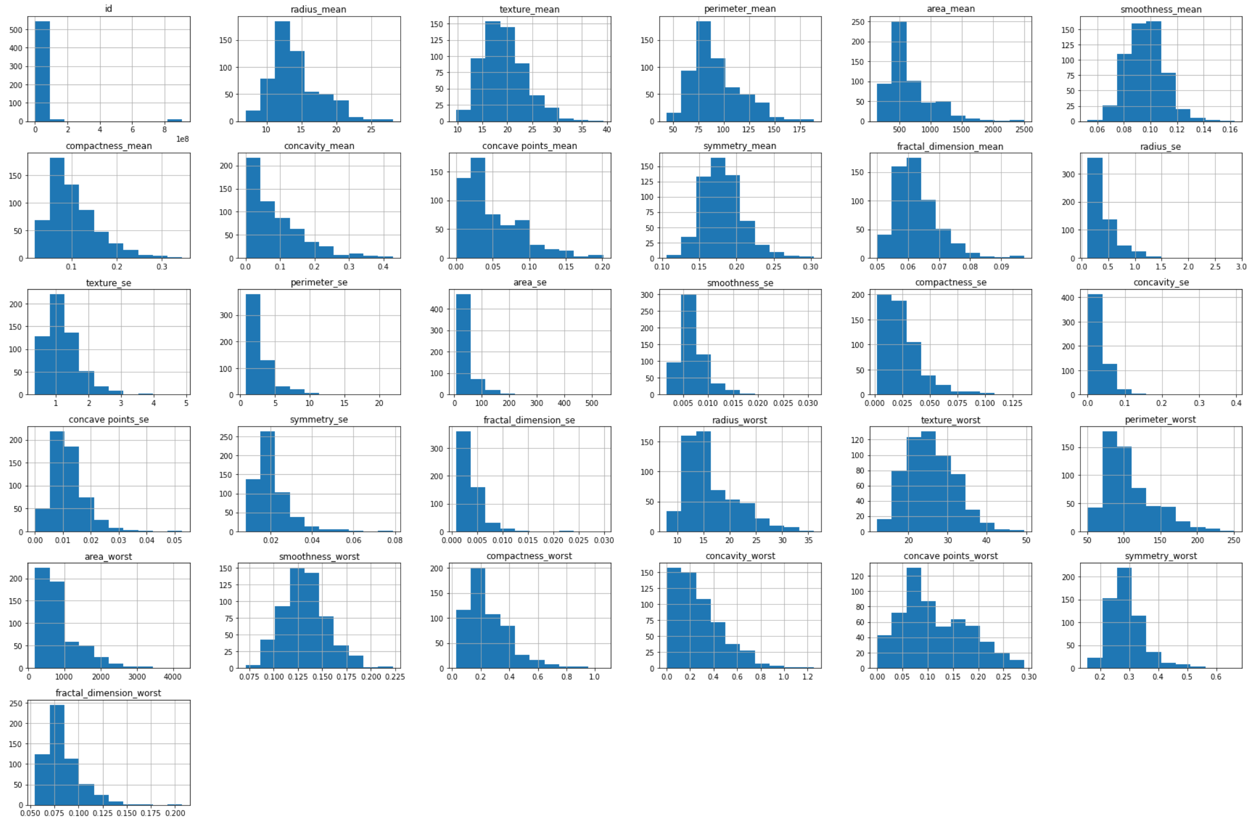
```
In [ ]: df.drop(["id"],axis=1)
```

Out[ ]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	...	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	sy
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	...	25.380	17.33	184.60	2019.0	0.16220	0.66560	0.7119	0.2654	
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	...	24.990	23.41	158.80	1956.0	0.12380	0.18660	0.2416	0.1860	
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	...	23.570	25.53	152.50	1709.0	0.14440	0.42450	0.4504	0.2430	
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	...	14.910	26.50	98.87	567.7	0.20980	0.86630	0.6869	0.2575	
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	...	22.540	16.67	152.20	1575.0	0.13740	0.20500	0.4000	0.1625	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
564	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	...	25.450	26.40	166.10	2027.0	0.14100	0.21130	0.4107	0.2216	
565	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	...	23.690	38.25	155.00	1731.0	0.11660	0.19220	0.3215	0.1628	
566	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	...	18.980	34.12	126.70	1124.0	0.11390	0.30940	0.3403	0.1418	
567	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	...	25.740	39.42	184.60	1821.0	0.16500	0.86810	0.9387	0.2650	
568	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	...	9.456	30.37	59.16	268.6	0.08996	0.06444	0.0000	0.0000	

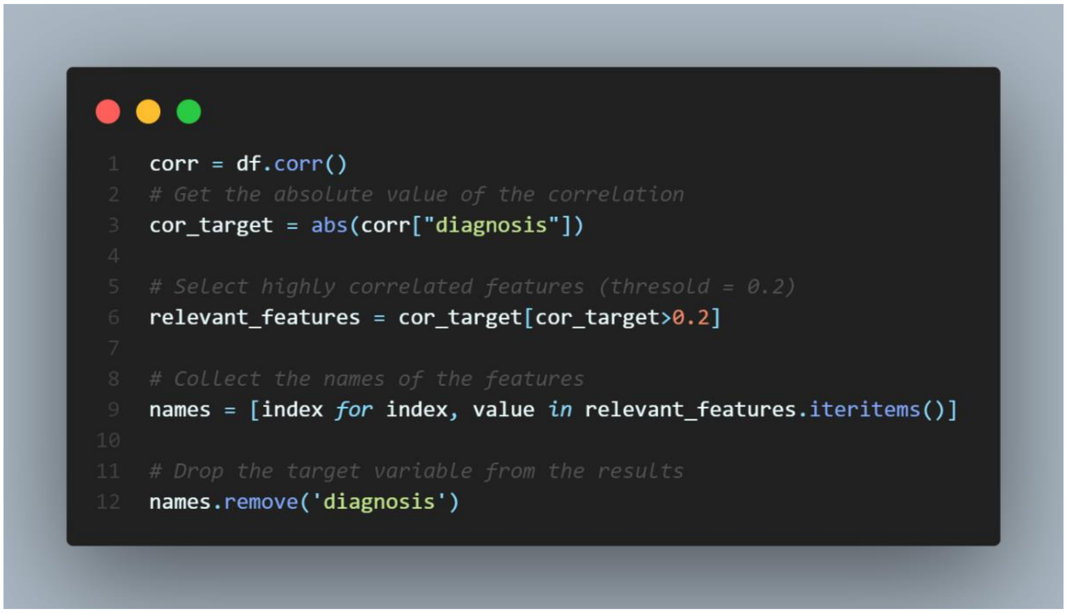
569 rows × 31 columns

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022



After looking at the data I began to scope out some of the changes I would need to make in order to build a model that functions properly within the logistic regression That I was creating. In this process I also looked at the different histograms for each column to better understand my data as well.

Once I had a better understanding of what my data was I began to build a model around my data by getting the absolute value of my correlation after using the dataframe.corr() function on my data frame. The dataframe.corr() function was used to find the pairwise correlation of all the columns in the pandas data frame. This also automatically ignores any NaN and non-numerical data types or columns within my data frame, but since I had already been using a dataset that was numerical in nature, I did not have to rely on this feature. I sorted based on highly correlated features and set the correlation threshold to > 0.2, while also collecting all the names of the features. At this point as well I dropped the diagnosis variable, as it gave an resulting diagnosis.



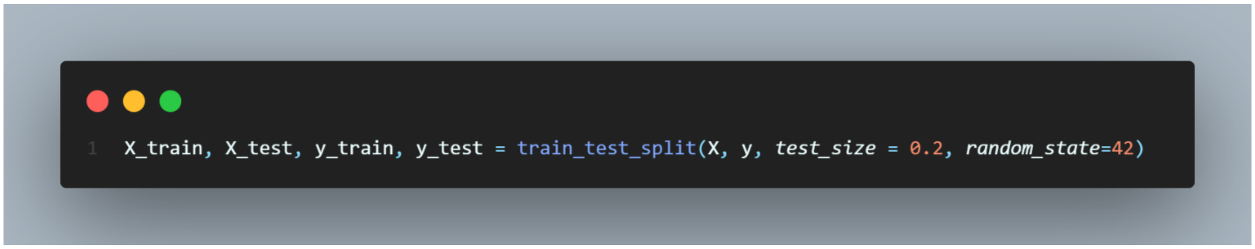
```python
corr = df.corr()
# Get the absolute value of the correlation
cor_target = abs(corr["diagnosis"])

# Select highly correlated features (thresold = 0.2)
relevant_features = cor_target[cor_target>0.2]

# Collect the names of the features
names = [index for index, value in relevant_features.iteritems()]

# Drop the target variable from the results
names.remove('diagnosis')
```

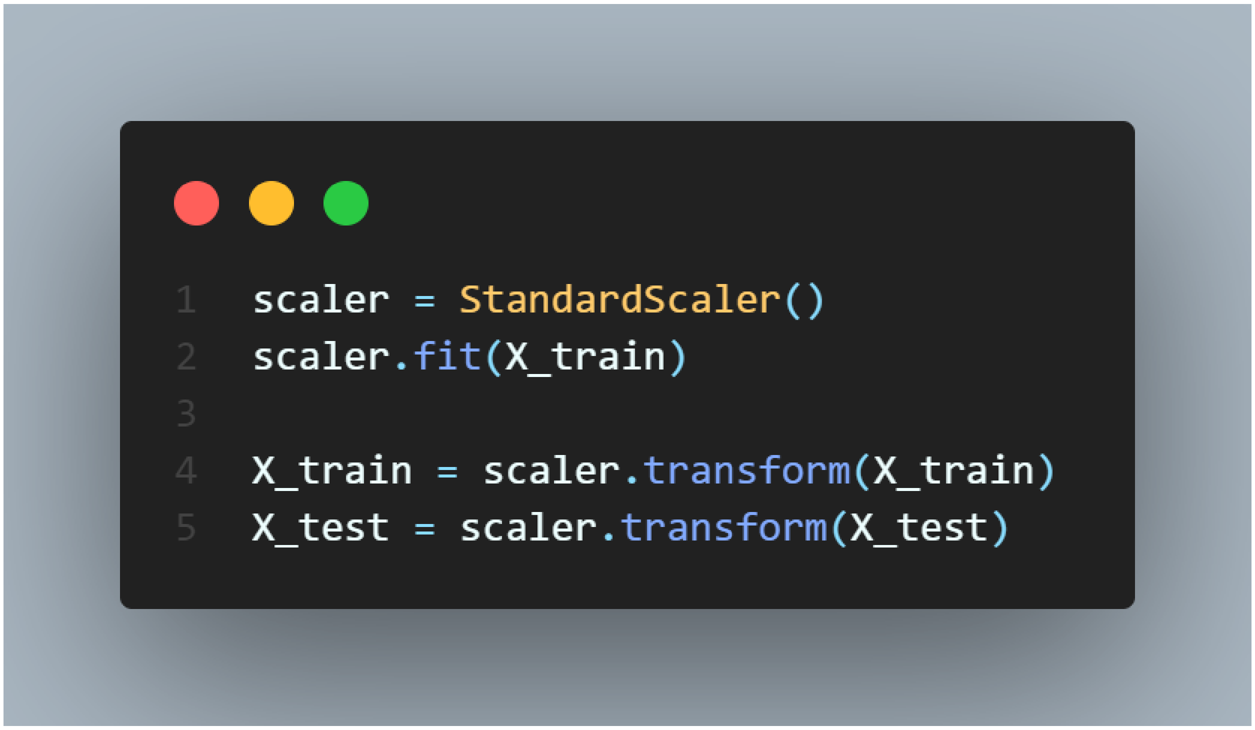
Next, I split my data into training and validation sets.



```python
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=42)
```

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022

At this point I found that I needed to add a StandardScaler() to my data. The idea behind this was that I needed to transform my data in such a way that it’s distribution will have a mean value of 0 and a standard deviation of 1. I added my scaler = StandardScaler() to my training set and test set on the X values.

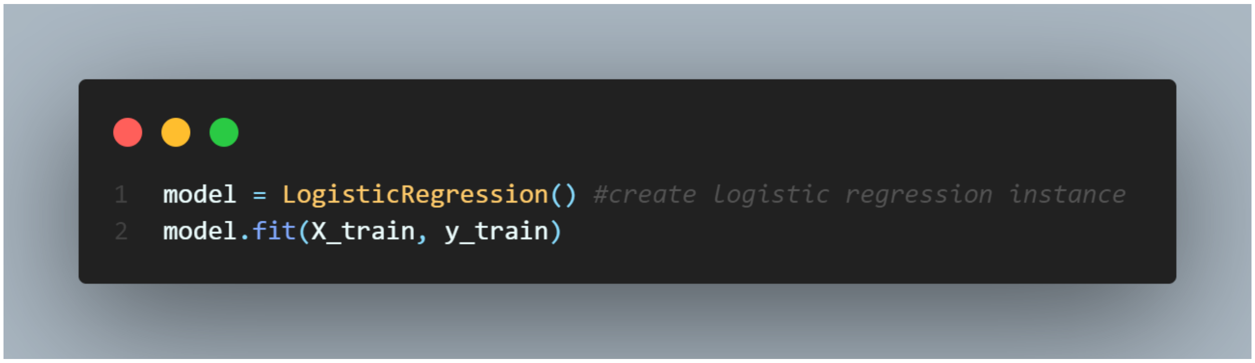


```
scaler = StandardScaler()
scaler.fit(X_train)

X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

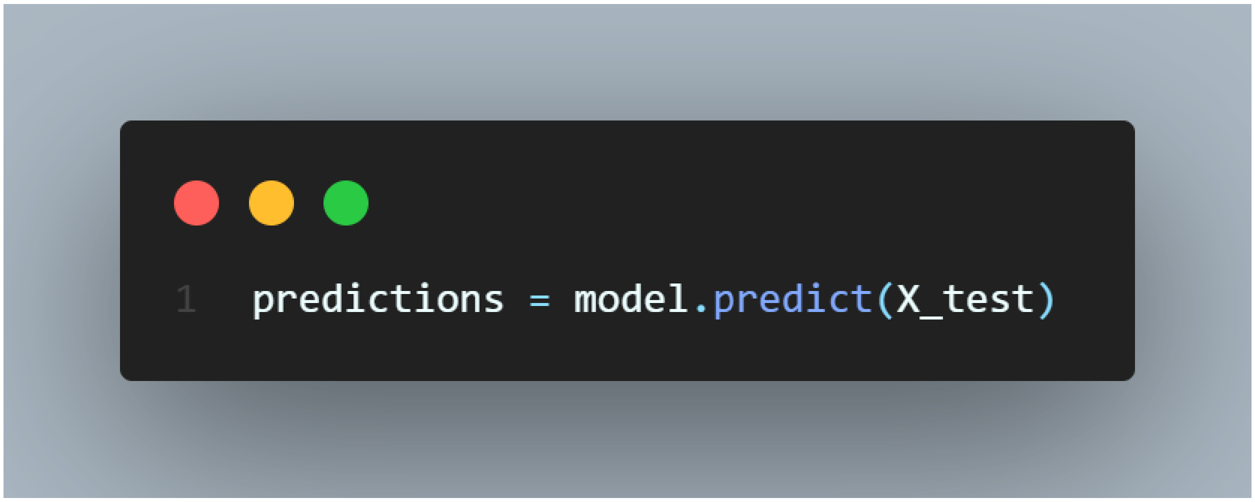
From this point it was a good idea to create a logistical regression instance and fit myt now transformed training sets into my LogisticRegession model. This was chosen as Logistical Regression is a classification algorithm that can be usd to predict the problabity of a caegorical dependent variable and contains the data coded as a 1 as successful or 0 as a failure.

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022



```python
model = LogisticRegression() #create logistic regression instance
model.fit(X_train, y_train)
```

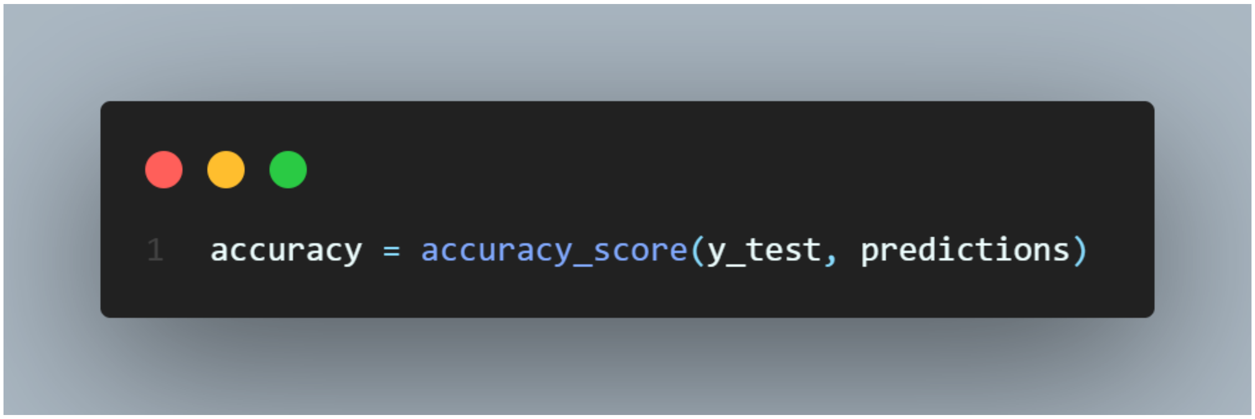
The predict function was used to predict the values based on the previous data's behavior and thus fiting that data to the model. I set my predictions to the X_test model as to preform predictions on each test instance and accept only a single input.



```python
predictions = model.predict(X_test)
```

Once that step was completed, it was time to generate the accuracy score of my model as to determine how accurate my model was against my predictions model. The accuracy score is used as a means to determine the model's performance by measuring the ratio of sum of true positives and true negatives of all the predictions made against the test model. In this case we tested against the y_test model which is meant to be our target data set to predict against.

Joshua Burden
Bellevue University
Milestone 4
Andrew Hua
DSC630 Predictive Analytics
10/30/2022



```python
accuracy = accuracy_score(y_test, predictions)
```

After all that was said and done, my model accuracy: 0.9736842105263158 or 97% accurate.

I want to test this against other models and explore more of my data before I give a more conclusive description of what my data really means. But for the current moment it would appear that my current models predictive quality is that it can predict with an accurate predictive rate of roughly 97% of the test result for cancer diagnosis.