Predictive Analytics for Breast Cancer Diagnosis Using PCA-Enhanced Logistic Regression

Prepared for:

UNT ADTA 5230

Prepared by:

Bryanna Williams, Biniam Abebe, Guttikonde Bharat, and Maybel Hernandez

April 26, 2024

1. Abstract (Make sure to fill this in at the end)

The purpose of this research is to identify 3-4 features that have a heavy relationship with receiving a malignant diagnosis.

***Keywords****: Machine Learning, Supervised Machine Learning, Credit Card Fraud, Fraud Detection, Decision Tree, Random Forest, XG-Boost, Kernelized Support Vector Machine, Logistic Regression, Neural Networks*

Contents

[Introduction 5](#_Toc165558572)

[Literature Review 6](#_Toc165558573)

[Data Description 7](#_Toc165558574)

[Exploratory Data Analysis 9](#_Toc165558575)

[V. Metodoloy -- I am working from here Down 10](#_Toc165558576)

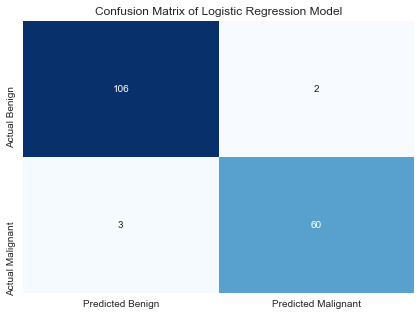
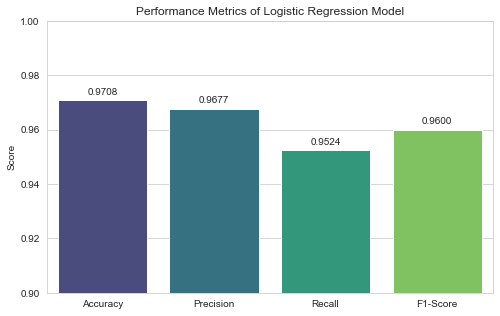
[Data Preprocessing # Normalization 10](#_Toc165558577)

[PCA Implementation # Dimensionality Reduction using PCA 10](#_Toc165558578)

[Model Development Initializing and training the logistic regression model 11](#_Toc165558579)

[Performance Metrics 11](#_Toc165558580)

[Results/Discussion 12](#_Toc165558581)

[Model Performance   12](#_Toc165558582)

[References 13](#_Toc165558583)

[Bibliography 13](#_Toc165558584)

## Introduction

We will be analyzing Breast Cancer Wisconsin (Diagnostic) Data Set from the UC Irvine Machine Learning Repository to conduct our research. The aim of this research is to determine which attributes tested trend to have the most significant connection with a positive diagnosis. These findings could contribute significantly to breast cancer research by allowing clinicians to identify which factors have a heavier impact, therefore guiding research and study practices more heavily towards identifying the cause of those features and how to mitigate them. Accurately detecting a positive diagnosis as early as possible is crucial for getting patients the needed treatment and ultimately leading to a drastic reduction in Breast Cancer mortality rate. Knowing what features to focus research and healing efforts on can drastically impact this goal.

## Literature Review

According to the World Cancer Fund International, Breast cancer was the most commonly assigned form of cancer in the year 2020 with over 2.2 million cases that were reported. It was also the culprit of approximately 685,000 deaths in women that same year. With this, it is clear to see that breast cancer poses a significant global health challenge. To improve the prediction and diagnosis of breast cancer, researchers have turned toward machine learning and data mining algorithms in an attempt to better understand this issue and help resolve this urgent problem. This literature review aims to provide an overview of current developments in machine-learning techniques for the detection and prediction of breast cancer, based on knowledge from two relevant scientific publications.

In their study to predict the diagnosis of breast cancer, Naji et al. (2001) tested five different machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN) to better understand their performance. Their study was based on the Wisconsin Breast Cancer Diagnostic dataset and showed that SVM outperformed other classifiers with 97.2% accuracy. Their research sheds light on the potential machine-learning algorithms have to improve diagnostic accuracy and patient outcomes in breast cancer care.

In a recent study, Fatima et al. (2020), performed an in-depth evaluation of machine-learning methodologies for predicting breast cancer. Their review examined several linear, nonlinear, and ensemble algorithms. Their analysis included algorithms such as Linear Regression, Logistic Regression, Classification and Regression Tree, Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, Boosting, and AdaBoost. The authors aimed to determine the most efficient and precise algorithm for predicting breast cancer through their comprehensive comparative analysis. Their review offers valuable insights into the various machine-learning techniques used in recent breast cancer research, highlighting the importance of algorithm selection in predictive modeling.

In conclusion, machine-learning approaches show promising potential for improving breast cancer prediction and diagnosis. The studies mentioned reinforce the importance of strategically selecting algorithms and evaluating their performance to create reliable predictive models of breast cancer. Going forward, it will be of great benefit for researchers to focus on addressing issues related to limited data sets, unbalanced data, and algorithm refinement to further enhance diagnostic accuracy and in turn, patient outcomes when receiving breast cancer treatment.

## Data Description

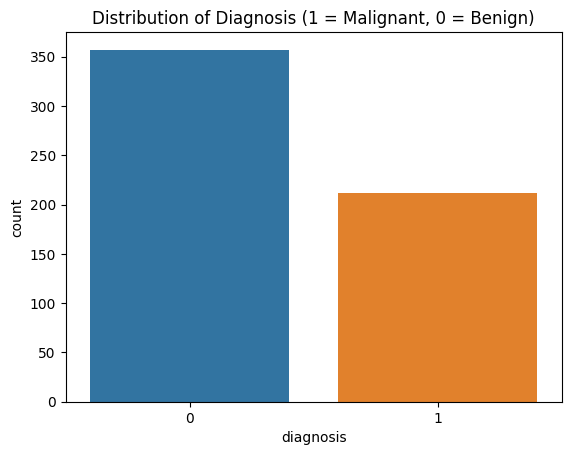
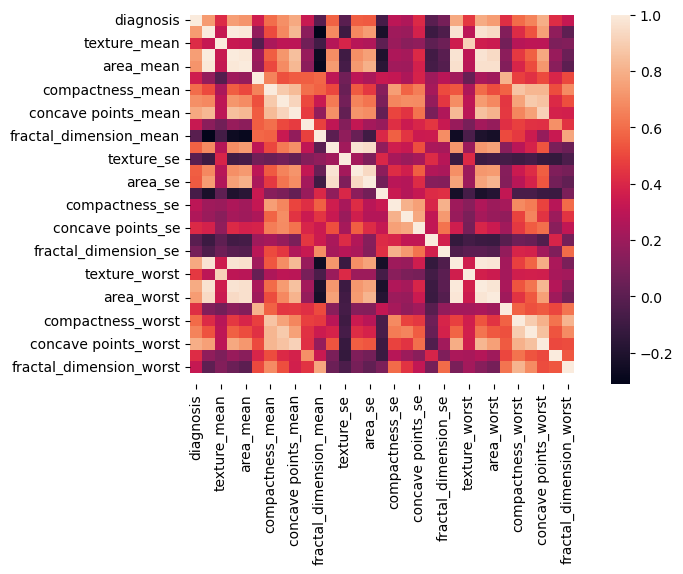
Breast Cancer Wisconsin (Diagnostic) Data Set contains features calculated from the image of the breast mass's fine needle aspirate (FNA) picture. The dataset contains 32 columns and 569 samples. The column description is as follows.

**Column Description:**

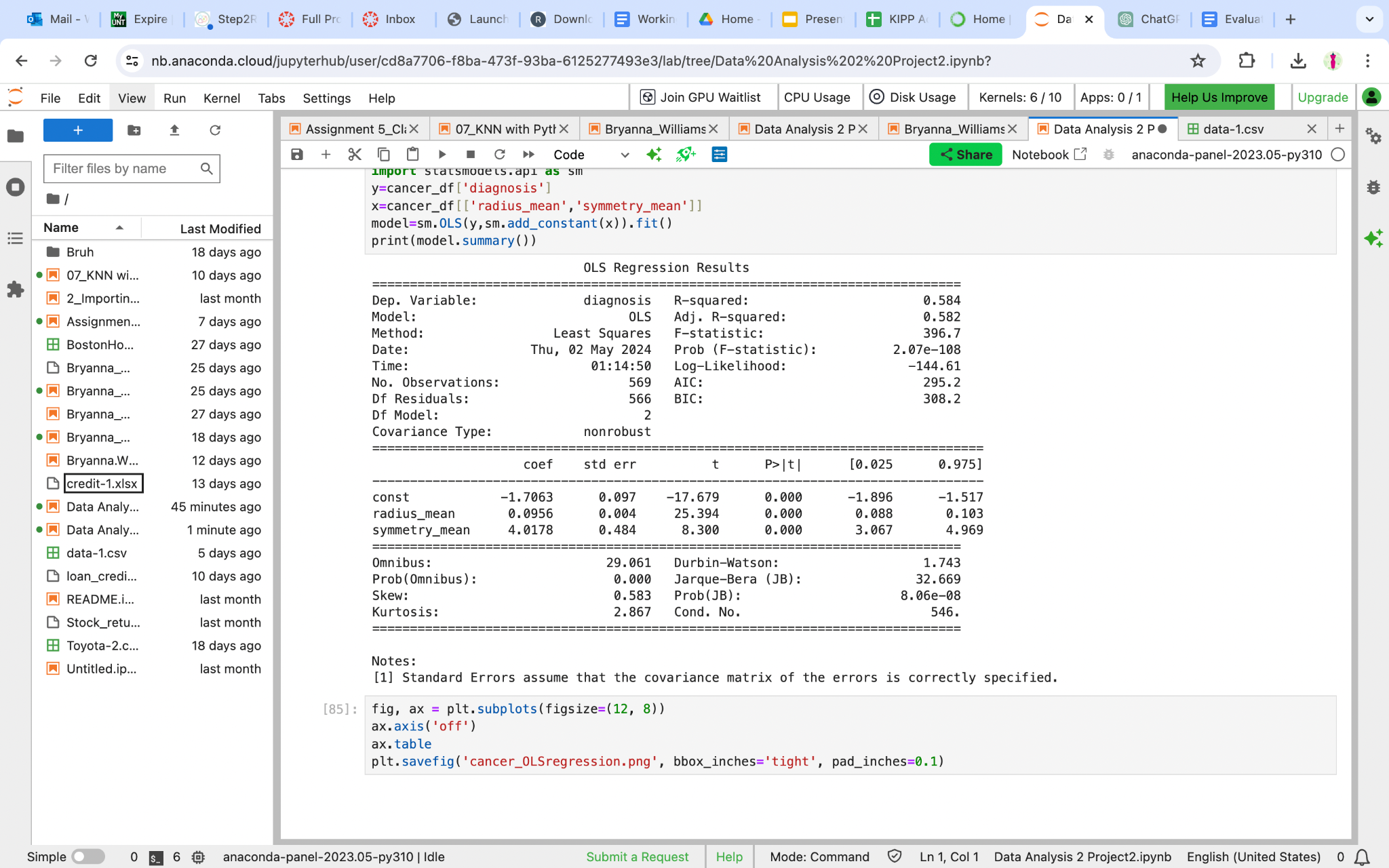
* ID number: Unique identifier for each sample.
* Diagnosis: The diagnosis of breast tissues (M = malignant, B = benign).
* Radius: Mean of distances from the center to points on the perimeter.
* Texture: Standard deviation of gray-scale values.
* Perimeter: Perimeter of the cell nucleus.
* Area: Area of the cell nucleus.
* Smoothness: Local variation in radius lengths.
* Compactness: Perimeter^2 / area - 1.0.
* Concavity: Severity of concave portions of the contour.
* Concave points: Number of concave portions of the contour.
* Symmetry: Symmetry of the cell nucleus.
* Fractal dimension: "Coastline approximation" - 1.

This project will use the Diagnosis column in this dataset as the target variable, which contains the diagnosis of breast tissues as either M (malignant ) or B (benign).

## Exploratory Data Analysis

  
 **Table 1-** This graphic shows the distribution of Malignant compared to Benign cancer diagnoses. The data consisted of \_\_\_\_\_ benign and \_\_\_\_\_ Malignant diagnosis. There were a total of 569 patients from which this data was pulled.

**Table 2-** This heatmap shows correlation matrix between the variables in our data set. Variables in pale orange have the strongest positive correlation with each other such as texture mean and worst and area worst and perimeter



**Table 3 -** Regression analysis y=diagnosis. The two variables used are radius mean and symmetry mean. These two variables can explain about 58.4% of the variation in diagnosis results.

## V. Metodoloy -- I am working from here Down

### Data Preprocessing # Normalization

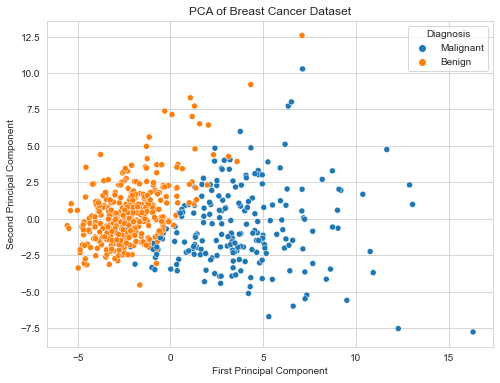
**scaler = StandardScaler()**

**X\_scaled = scaler.fit\_transform(X)**

### PCA Implementation # Dimensionality Reduction using PCA

pca = PCA(n\_components=2) # Reduce dimensions to 2 for visualization or further analysis

X\_pca = pca.fit\_transform(X\_scaled)



### Model Development Initializing and training the logistic regression model

log\_reg = LogisticRegression()

log\_reg.fit(X\_train, y\_train)

### Performance Metrics

accuracy = accuracy\_score(y\_test, y\_pred)

precision = precision\_score(y\_test, y\_pred, pos\_label= 'M')

recall = recall\_score(y\_test, y\_pred, pos\_label= 'M')

f1 = f1\_score(y\_test, y\_pred,pos\_label= 'M')

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

accuracy, precision, recall, f1, conf\_matrix

### Results/Discussion

### Model Performance

Based on findings from the above methods we will be able to determine which features are most strongly connected to a cancer diagnosis. The importance of knowing this is that researchers can gain insight on which factors of detection may need more attention as they have the most significant impact on whether a diagnosis is made.

### References

* Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. <https://doi.org/10.24432/C5DW2B>.
* Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, *8*, 150360–150376. <https://doi.org/10.1109/access.2020.3016715>
* Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Computer Science*, *191*, 487–492. <https://doi.org/10.1016/j.procs.2021.07.062>

### Bibliography