Jedsada Thavornfung

JT43227

PSY 341K Data Sci Foundations II

Final Project Report

Introduction

In this project, I tested 2 algorithms with 2 selected variables and PCA. First, I randomly selected variables to calculate their correlation, which being illustrated in Fig. 1. Based on the correlation, I decided to pick *mean area* and *area error* because they have the correlation of 0.80 for both of my algorithms. For my algorithms, the selected algorithms are K-Nearest Neighbor (KNN) and Gaussian Naive Bayes.

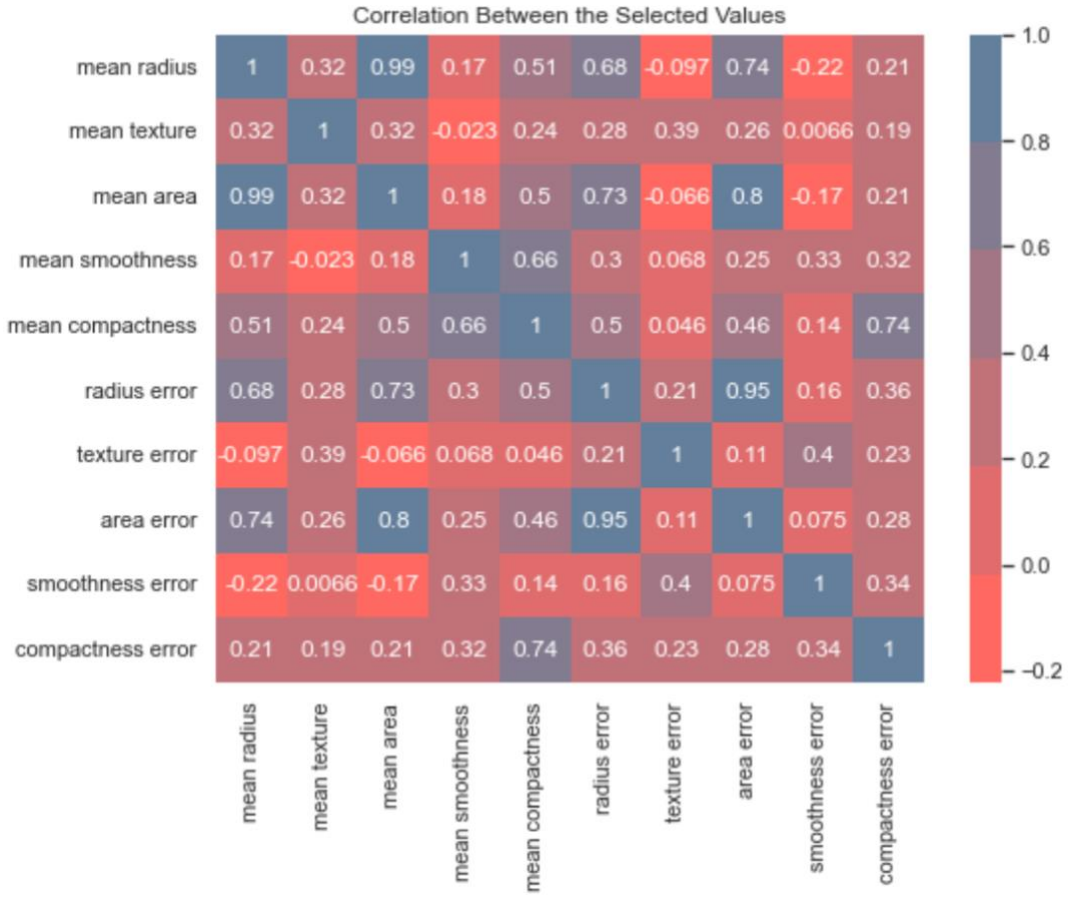

Figure 1. The correlation between the randomly selected variables.

Jedsada Thavornfung

JT43227

PSY 341K Data Sci Foundations II

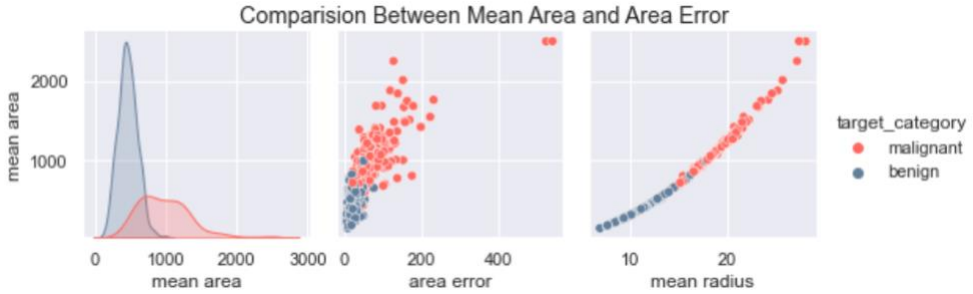

Figure 2. The comparison between **mean area** and **area error**.

Next step, I look at the PCA values to reduce the multiples dimension (columns) into 2 dimensions to make them more interpretable and easier to look at.

Principal Component Analysis (PCA)

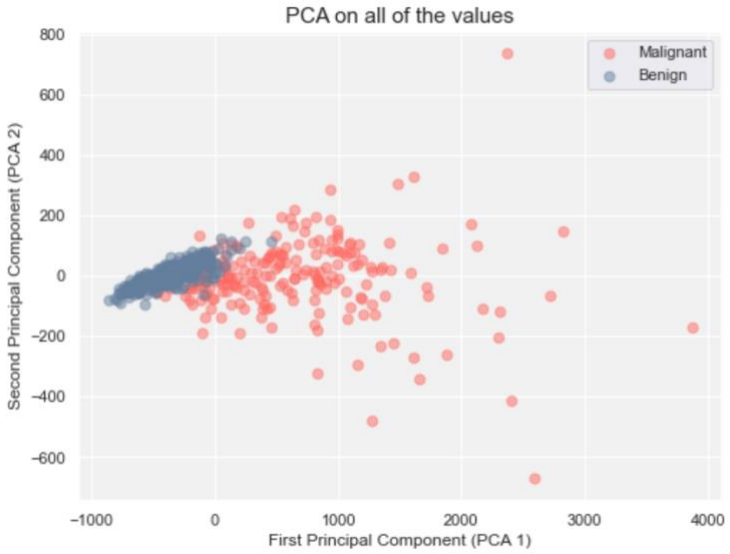

Figure 3. Performing principal component analysis for all of the data, which was reduced from 30 dimensions to 2 dimensions: PCA 1 and PCA 2.

The PCA was performed to reduce from 30 dimensions to 2 dimensions to compared along with other algorithms.

Jedsada Thavornfung

JT43227

PSY 341K Data Sci Foundations II

K-Nearest Neighbor (KNN)

In KNN algorithms, I performed KNN with 2 selected columns: mean area and area error. Then, I performed KNN for the prepared PCA 1 and PCA 2 in the previous section. In this project, I chose to k = 3.

A. B.

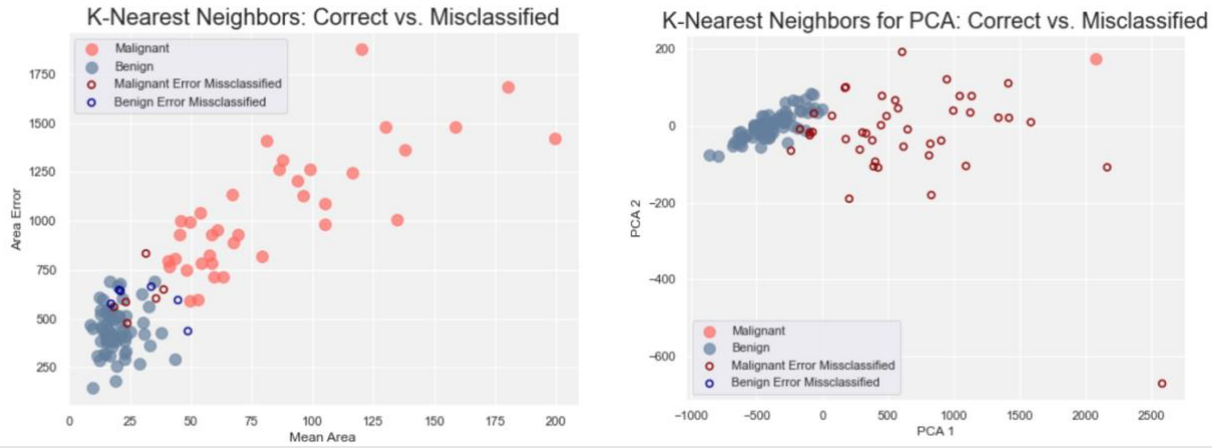

Figure 4. The first figure (A) illustrated KNN between mean area and area error, which illustrated correct classified malignant, correct classified benign, misclassified malignant, and misclassified benign. The second figure (B) illustrated KNN between PCA 1 and PCA 2 with the same information as figure A.

As a result, KNN algorithm correctly predicted the type of the target: benign and malignant 89.47% of the total data for mean area and area error. However, KMM algorithm only correctly predicted the type of the target: benign and malignant 63.16% of the total data for PCA 1 and PCA 2. The reason that KNN algorithm for mean area and area error (Fig. 4A) has better accuracy than PCA 1&2 could be because there were less variable to considered for Fig. 4A when performing the algorithm. Also, mean area and area error had a high correlation, which could influence the KNN classifier to perform more accurately. It appears that KNN algorithm did not perform well at capturing overall data for PCA values in which mostly misclassified malignant. However, KNN algorithm appeared to capture the benign class very accurately. I hypothesized that this happen due to the incorrect Choice of K; the choice of the parameter K, which determines the number of nearest neighbors used to classify a data point, can also affect the accuracy of KNN clustering. In this project, I decided to use k = 3, but there might be another k-value that fit better for PCA 1 and 2.

Jedsada Thavornfung

JT43227

PSY 341K Data Sci Foundations II

Gaussian Naive Bayes (GNB)

Another algorithm that was selected in this project is Gaussian Naive Bayes (GNB). Similar to KNN, I performed two GNB for the selected variables (mean area and area error) and PCA 1&2.

A. B.

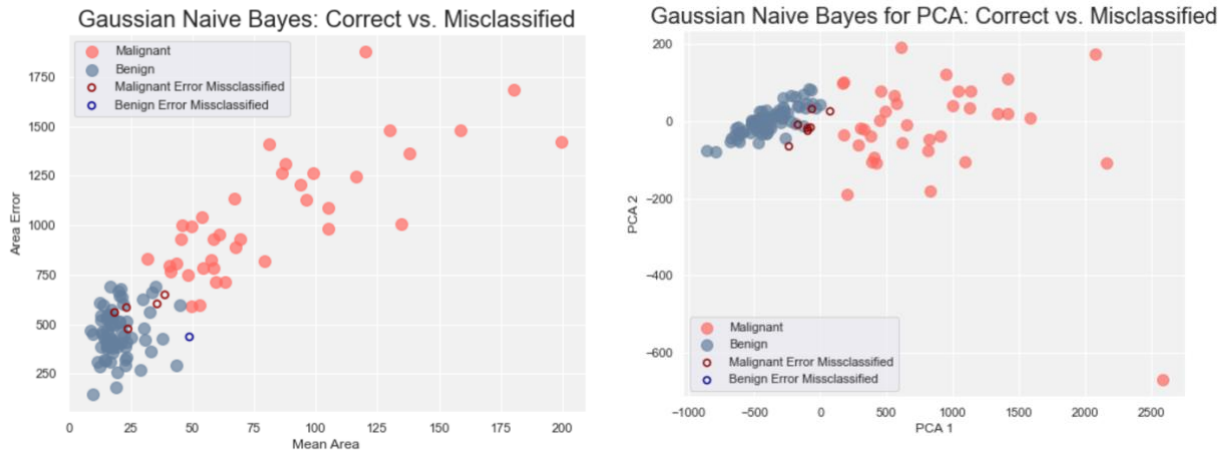

Figure 5. The first figure (A) illustrated GNB between mean area and area error, which illustrated correct classified malignant, correct classified benign, misclassified malignant, and misclassified benign. The second figure (B) illustrated GNB between PCA 1 and PCA 2 with the same information as figure A.

As a result, GNB algorithm correctly predicted the type of the target: benign and malignant 94.74% of the total data for mean area and area error. However, GNB algorithm correctly predicted the type of the target: benign and malignant 93.86% of the total data for PCA 1 and PCA 2, which is slightly slower than the Fig. 5A. Thus, GNB algorithm could classified and predicted the target better than KNN algorithm. The reason that GNB algorithm performed better than KNN algorithm because of the model complexity, GNB is a simpler and more computationally efficient model than KNN. KNN needs to store all the training data and compute distances between data points, which can be time-consuming for large datasets. In contrast, GNB only needs to compute the mean and variance of the features for each class, which can be done quickly.

Jedsada Thavornfung

JT43227

PSY 341K Data Sci Foundations II

Overall Result

	Algorithm 1	**Algorithm 2**
2 good variables by eye (Mean Area x Area Error)	K-Nearest Neighbors: 89.47% Accuracy	Gaussian Naive Bayes: 94.74% Accuracy
Best two components via PCA	63.16% Accuracy	93.86% Accuracy