MVA Topics (and Datasets)

Jared Lakhani

February 26, 2024

1 Introduction

According to [7], cancer is one of the leading causes of death, where there were nearly ten million deaths in 2020 (or one in every six deaths). Suffice to say, predicting cancer in its early stages significantly improves the chances of successful treatment and recovery. Therefore, there's a pressing need for reliable cancer prediction tools capable of accurately classifying tumors as either malignant (harmful) or benign (non-harmful) - where to aid this process, classification methods would be utilised.

Preceding classification, one would be inclined to reduce the number of features through dimensionality reduction, for reasons such as preventing overfitting, noise reduction and improving computational efficiency. However, dimensionality reduction in medical data poses a significant challenge, especially for datasets with numerous dimensions containing crucial information [1].

2 Research Questions

Two dimension reduction techniques are to be utilized, namely the popular principle component analysis and cluster analysis. The latter is quite a peculiar means of dimension reduction - where this unsupervised machine learning technique will create clusters to serve as new features as used in [3].

We thus aim to shed light on whether the different dimension reduction techniques used for different classification methods, serve adequately to predict benign or malignant tumors by assessing performance measures (accuracy, sensitivity and specificity)- but more so, we aim to elucidate the differences between all methods used.

- **Train/Test Split Ratio** We aim to see if there is a difference in performance measures by varying the train/test split ratio for both dimension reduction techniques.
- **Number of PCs** By varying the number of principle components utilised in classification and assessing the associated performance scores, we can make a conclusion on whether there is an ideal number of PCs to use.
- **Clustering Techniques** By changing the type of clustering technique (methods in both hierarchical and non-hierarchical clustering) and assessing performance measures, we can assess if there is an ideal technique to be utilized or if they all perform equally well.
- **Number of Clusters** We aim to vary the number of clusters used for classification and examining the performance measures, so to determine if there is an ideal number of clusters to be used.
- **Classification Methods** Different machine learning algorithms will be employed in the classification process - we aim to assess the corresponding similarities and differences between said classification methods (we aim to be reviewing SVM, Logistic Regression and Naive Bayes as possible candidates for classification)

3 Datasets

Two datasets were examined, both having 'benign' and 'malignant' tumor as a response variable - and physical charactersitcs corresponding to the tumors attirbutes/features. The first being the popular Wisconsin Breast Cancer (Original), and the second being a prostate cancer data set as used in [6]

3.1 Breast Cancer Dataset

The Wisconsin Breast Cancer (Original) data set, found in the UC Irvince Machine Learning Repository has a total of 699 instances with 9 attributes. The attributes are derived from a cell nuclei in a digitized image of a fine needle aspirate of a breast mass. More details can be found in [4]. These features are - Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class ('0' for benign and '1' for malignant). Subsequently, there are 16 NA values, and after removing said observations, there are 444 instances with 'benign' tumors and 239 instances with 'malignant' tumors. That is, harmless cells make up the dominant part class (65 %) - and there may emerge the issue of class imbalance. We note from Table 1, that all features are are comparable scaling - hence it is not imperative to standardize the data (although it can be done regardless).

Table 1: First 5 observations of WBC Dataset (Part 1)

Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion
5	1	1	1
5	4	4	5
3	1	1	1
6	8	8	1
4	1	1	3

Table 2: First 5 observations of WBC Dataset (Part 2)

Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli
2	1	3	1
7	10	3	2
2	2	3	1
3	4	3	7
2	1	3	1

Noting the correlation plot in Figure 1, we note only a single high correlation pair, namely Uniformity of Cell Shape and Uniformity of Cell Size having a correlation of 0.9 - as seen in Figure 2. From which we can infer that a tumor with a high Uniformity of Cell Shape will most likely have a high Uniformity of Cell Size and be cancerous.

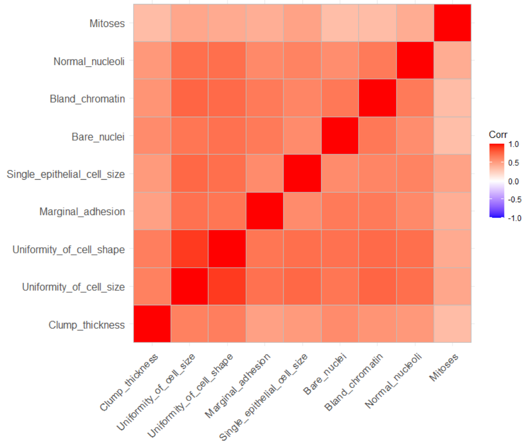


Figure 1: Correlation Plot between 9 Features

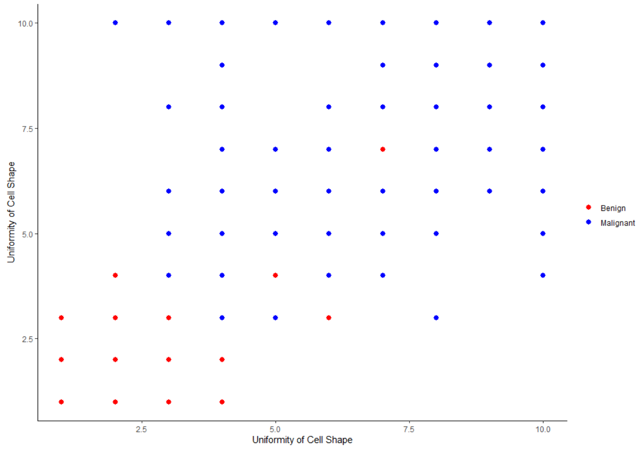
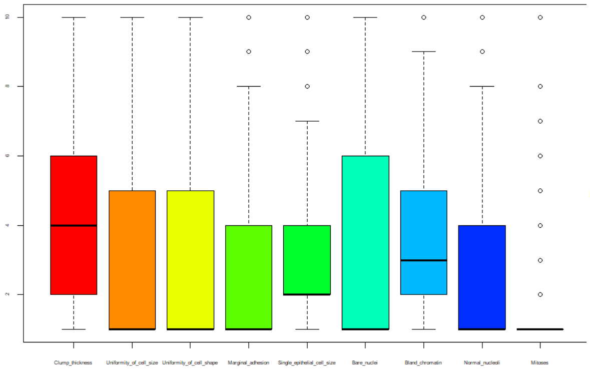


Figure 2: Correlation Plot between Uniformity of Cell Shape and Uniformity of Cell Size

We also note the boxplots in Figure 3. We note the median value for features: Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Bare Nuclei, Normal Nuclei and Mitoses are equal to 1 - which explains the presence of outliers. We note the histogram of the feature Mitoses in Figure 4 - that is - the majority of observations take on a value of 1.



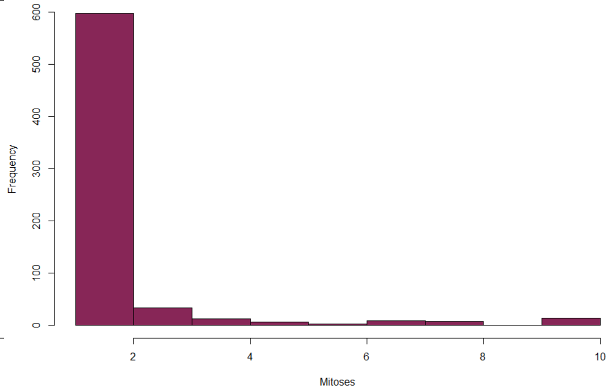


Figure 3: Boxplot of the 9 Features

Figure 4: Histogram of Mitoses

3.2 Prostate Cancer Dataset

The second data set, not as popular as the WBC data, is also obtained from digitized images [2], and can be obtained from the Kaggle data vault in [5], where there are only 100 instances with 8 attributes. The features are Radius, Texture, Perimeter, Area, Smoothness, Compactness, Symmetry, Fractal Dimension and Diagnosis Result ('0' for benign and '1' for malignant). There are no NA values, and 'malignant' tumors are the dominant class (62 %). We note the first 5 observations for the 8 features in Table 3, and posit that scaling the data is necessary as Area and Perimeter seem to dominate the dataset with their prominently high values.

The correlation plot in Figure 5 suggests the features, area and perimeter are highly correlated (98% - which is intuitive in a sense seeing as area is a function of perimeter). We plot their relationship in Figure 6 - where we can infer that if the tumor has a large perimeter, it most likely has a large area and is cancerous.

Table 3: First 5 Observations of Prostate Dataset Features

radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
23	12	151	954	0.143	0.278	0.242	0.079
9	13	133	1, 326	0.143	0.079	0.181	0.057
21	27	130	1, 203	0.125	0.160	0.207	0.060
14	16	78	386	0.070	0.284	0.260	0.097
9	19	135	1, 297	0.141	0.133	0.181	0.059
25	25	83	477	0.128	0.170	0.209	0.076

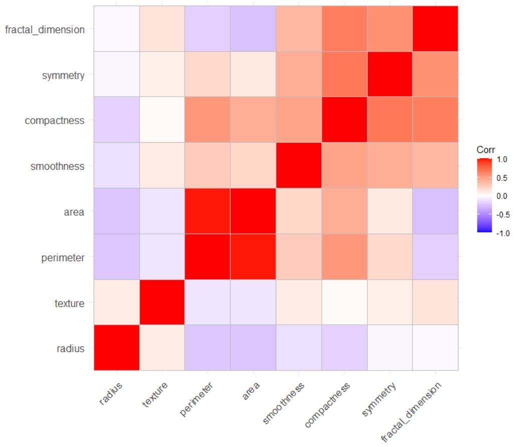


Figure 5: Correlation Plot between all 8 Features

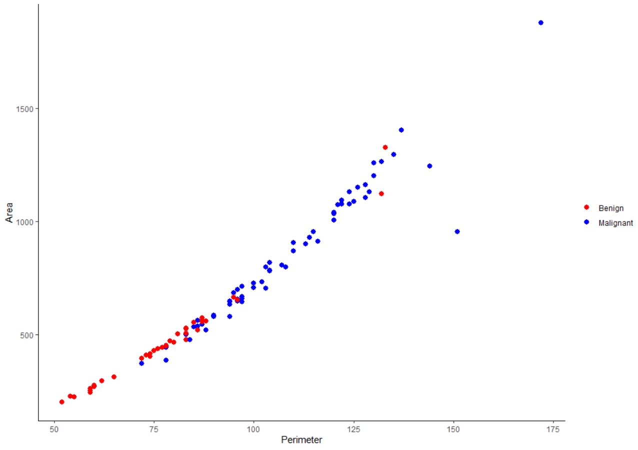


Figure 6: Correlation Plot between Perimeter and Area

From the boxplot of the scaled data in Figure 7, we conclude that there are no drastic outliers to be concerned about.

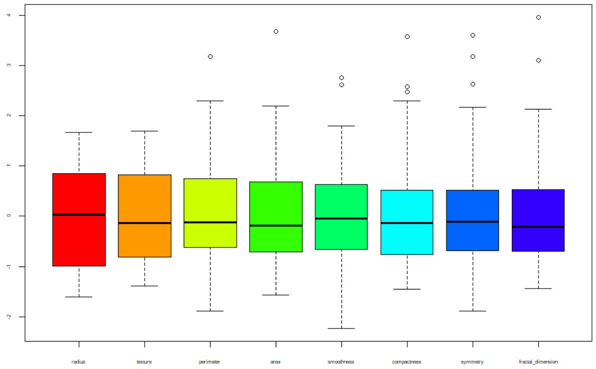


Figure 7: Boxplot of the 8 Features
ate

References

[1] Chitra Desai. Analysis of impact of principal component analysis and feature selection for detection of breast cancer using machine learning algorithms. *Journal Name*, 13:197–221, 01 2023.

[2] Surbhi Gupta, Manoj Kumar Gupta, and Rakesh Kumar. A novel multi-neural ensemble approach for cancer diagnosis. *Applied Artificial Intelligence*, 36(1):2018182, 2022.

[3] Ade Jamal, Annisa Handayani, Ali Akbar Septiandri, Endang Ripmiatin, and Yunus Effendi. Dimensionality reduction using pca and k-means clustering for breast cancer prediction. *Lontar Komput. J. Ilm. Teknol. Inf*, 9(3):192–201, 2018.

[4] Olvi L Mangasarian, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations research*, 43(4):570–577, 1995.

[5] Sajid Saif. Prostate cancer. https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer/data, 2018. [Online; accessed February 25, 2024].

[6] Manoj Kumar Gupta Surbhi Gupta and Rakesh Kumar. A novel multi-neural ensemble approach for cancer diagnosis. *Applied Artificial Intelligence*, 36(1):2018182, 2022.

[7] World Health Organization. Cancer, 2022.