MVA Assignment: Proposal

Jared Lakhani

04 March 2024

1 Introduction

According to the World Health Organization, cancer is one of the leading causes of death, where there were nearly ten million deaths in 2020 (or one in every six deaths) [23]. 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.

Now, most medical medical data sets are comprised of a plethora of features and observations - resulting in the predicting capabilities of machine learning algorithms being hampered [20]. Furthermore, a large feature set makes it difficult to visualize the data, as well as determining which feature has an impact on classification [21]. Redundant features, or ones that are derived from poor quality input would also impede the model's predictive quality [3]. Thus, the use of all the collective features would result in the curse of dimensionality - computational complexity, over-fitting and a hindering of classification performance [8]. Being such, the need for dimensionality reduction arises.

This study simply aims to ask whether there is an ideal dimension reduction technique for tumor prediction, and furthermore if there are ideal parameters needed to be used for each technique. Additionally, assessments will be made on the classification algorithms employed. We posit that model building on the entire feature set (without dimension reduction) will result in superior predictive quality - and propose that model building with a reduced feature set results in an adequate predictive performance.

2 Data Description and Analysis

The study aims to assess predictive quality of the built models on prostate tumors - where the proposed prostate tumor data set can be obtained from the Kaggle data vault in [17]. This data set was also utilised by [4] - a study which investigated deep learning techniques used in all types of cancer research. The features are calculated from a digitized image of a fine needle aspirate of a prostate tumor mass - more specifically, the cell nuclei present in the image [4]. In this case, the instances/observations are digitized images of tumors (both benign and malignant) - although it is uncertain through which means these images where obtained (how and from which organization). The data consists of 100 instances with 8 attributes/features namely Radius, Texture, Perimeter, Area, Smoothness, Compactness, Symmetry, Fractal Dimension - to be used as predictor variables for model building. And a binary response variable, namely Diagnosis Result ('0' for benign and '1' for malignant). There are no NA values present in the data, and 'malignant' tumors are the dominant class (62 %) as seen in Figure 1.

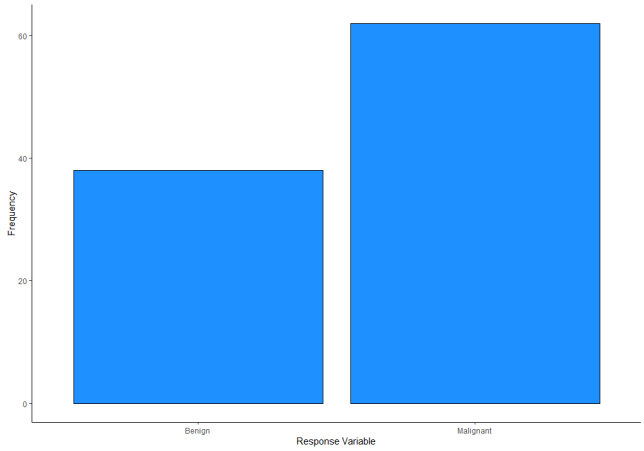
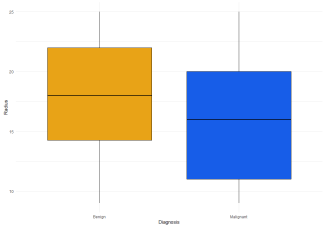
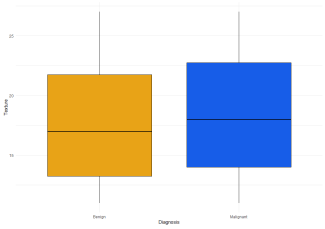


Figure 1: Histogram of Benign and Malignant Tumors in Data Set

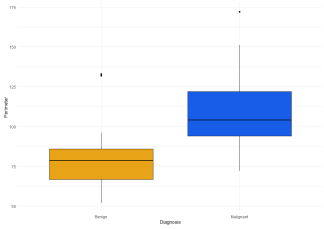
Boxplots for each feature are given in Figures 2a - 3d. Furthermore, we posit that there are no severe outliers to be concerned about.



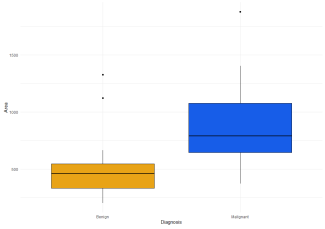
(a) Boxplot for Radius



(b) Boxplot for Texture

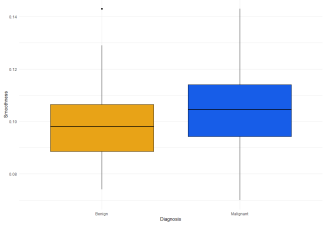


(c) Boxplot for Perimeter

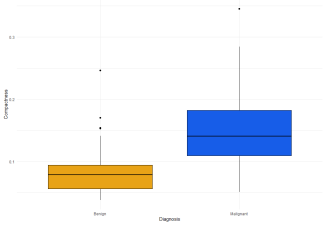


(d) Boxplot for Area

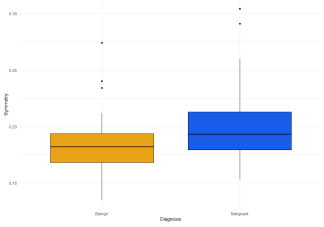
Figure 2: Boxplots for First 4 Features



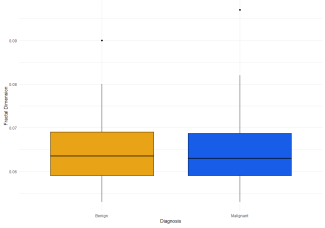
(a) Boxplot for Smoothness



(b) Boxplot for Compactness



(c) Boxplot for Symmetry



(d) Boxplot for Fractal Dimension

Figure 3: Boxplots for Latter 4 Features

3 Methodology

3.1 Dimension Reduction

There are two approaches to dimension reduction: the first being feature selection where only certain features are selected (and the rest disgarded), and the second being compression which aims to create new features from existing ones. Moreover, feature selection in the medical field may result in a loss of crucial information (if not adequately understood through exploratory data analysis [3])- hence we will only be undergoing compression.

3.1.1 Linear Principle Component Analysis

The technique of PCA forms the bedrock for dimensionality reduction. Invented by Karl Pearson in 1901, PCA aims to find a new coordinate system such that each new dimension is orthogonal to the next and are ranked according to the amount of variation explained [1] - which has been found useful in noise reduction and retaining important information. The principle components are actually the eigenvectors of the covariance matrix of the scaled data (or correlation matrix) - where the eigenvalues represent the the variation explained of the corresponding eigenvector.

3.1.2 Kernel Principle Component Analysis

Kernel PCA (Principal Component Analysis) is a nonlinear dimensionality reduction technique that extends traditional PCA to capture nonlinear relationships in data [12]. While traditional PCA assumes linear relationships between variables, Kernel PCA allows for more flexible modeling by implicitly mapping the input data into a high-dimensional feature space using a nonlinear mapping function, known as the kernel function [6]. The key idea behind Kernel PCA is to project the input data into a higher-dimensional space, where nonlinear relationships become linear or easier to separate. This is achieved by computing pairwise similarity measures, or kernel functions, between data points in the original space. Commonly used kernel functions include the radial basis function (RBF) kernel, polynomial kernel, and sigmoid kernel. Once the data is mapped into the high-dimensional feature space, PCA is performed to find the principal components, which are the directions of maximum variance in the transformed space [18].

3.1.3 Principle Curves

The concept of principal curves was introduced by Hastie and Stuetzle in 1989 as a means to uncover the "central tendency" of data points in a high-dimensional space. They defined principal curves as curves that pass through the "middle" of the

data cloud, minimizing the sum of squared perpendicular distances from the data points to the curve [5]. This formulation ensures that the curve effectively captures the essential features of the dataset while discarding noise and outliers. Unlike linear techniques such as PCA, which seek linear projections that maximize variance, principal curves aim to capture the underlying structure of the data through smooth, nonlinear paths. Similarly to PCA however, the transformed points are generated by the projections of the original points onto the principal curve (or principle component in the case of PCA) and are measured by their arc lengths.

3.1.4 Cluster Analysis

Clustering is a method used in unsupervised machine learning to identify inherent structures within data by grouping similar observations based on certain criteria. These criteria can vary depending on the context and nature of the data. Generally, similarity between observations is determined by the distance or dissimilarity between data points in a multidimensional space [24]. Other criteria include clustering based on probability, modality or density. Probability-based clustering assigns points to clusters based on the probability distribution that is most likely to have generated them. For instance, a point's value might determine which distribution it is more likely to belong to. Modality clustering identifies clusters based on the number of peaks in the density of observations. If there are multiple peaks in the density plot of a variable, it suggests the existence of multiple clusters. Density clustering identifies clusters based on regions of high density within a dataset. Points within these high-density regions are grouped together to form clusters.

Clustering can be applied to observations, variables, or both. When clustering observations, the goal is to group similar data points together based on their feature values. On the other hand, clustering variables involves grouping features that exhibit similar patterns across observations. This can be useful for identifying redundant or highly correlated variables in datasets [2]. Two-way clustering, also known as biclustering, simultaneously clusters both observations and variables. This technique is particularly beneficial when exploring datasets where the relationship between observations and variables is complex or when patterns of interest are present in subsets of both observations and variables [11]. Two-way clustering is popular in gene expression analysis in bioinformatics.

3.1.4.1 Hierarchical Clustering Hierarchical clustering is a method of unsupervised machine learning used to organize data into a hierarchical structure of clusters. Unlike other clustering methods, hierarchical clustering creates a tree-like hierarchy of clusters, known as a dendrogram, where clusters at each level of the tree are formed by merging or splitting existing clusters based on similarity [13].

There are two main types of hierarchical clustering: agglomerative and divisive. In agglomerative hierarchical clustering, each data point starts as its own cluster, and pairs of clusters are iteratively merged together based on a chosen similarity measure until all points belong to a single cluster. Divisive hierarchical clustering, on the other hand, starts with all data points in one cluster and recursively splits them into smaller clusters until each point is in its own cluster.

3.1.4.2 Non-Hierarchical or Partitioning Clustering Non-hierarchical clustering, also known as partitioning clustering, refers to a class of clustering algorithms that directly divide data points into a pre-defined number of clusters without forming a hierarchical structure. Unlike hierarchical clustering, which creates nested clusters, partitioning methods assign each data point to exactly one cluster [15].

One of the most popular partitioning clustering algorithms is K-means clustering. In K-means, the algorithm aims to partition the data into K clusters, where K is specified by the user. The algorithm iteratively assigns each data point to the nearest cluster centroid and updates the centroids based on the mean of the points assigned to each cluster. This process continues until convergence, typically when the assignments of data points to clusters no longer change significantly [9]. We aim to utilize K-means clustering, along with other partitioning methods.

Another widely used partitioning clustering algorithm is Gaussian Mixture Models (GMM). GMM represents the distribution of data as a mixture of several Gaussian distributions, each corresponding to a cluster. The algorithm estimates the parameters of these Gaussian distributions, including mean and covariance, to maximize the likelihood of observing the data [25]. Data points are then assigned to clusters based on their probability of belonging to each Gaussian distribution.

3.2 Classification Methods

Classification methods are systematic approaches used to construct classifiers from input datasets. These classifiers are built based on a learning target function that maps each feature set to predetermined class labels [7]. The process of classification involves two main steps.

Firstly, a classification algorithm constructs the classifier by analyzing a training set composed of database tuples and their corresponding class labels. This phase, often referred to as supervised learning, entails providing the class label for

each training tuple. During this step, the algorithm learns the relationships between the features and the class labels, enabling it to make accurate predictions.

In the second phase, the trained classifier is utilized for classification. Given a new set of features, the classifier assigns it to one of the predefined classes based on the learned patterns from the training data. This step allows the classifier to generalize its predictions to unseen data.

3.2.1 Support Vector Machines

Support Vector Machine (SVM) works by finding the optimal hyperplane that best separates data points into different classes while maximizing the margin between classes. SVM is effective in high-dimensional spaces and is versatile due to its ability to handle linear and nonlinear data through the use of kernel functions. It aims to find the decision boundary that maximizes the margin, making it robust to outliers [19].

3.2.2 Logistic Regression

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability of a binary outcome based on one or more predictor variables. Despite its name, logistic regression is a classification algorithm rather than a regression one. It models the relationship between the predictor variables and the probability of the binary outcome using the logistic function.

In logistic regression, the logistic function, also known as the sigmoid function, transforms the linear combination of predictor variables into a probability score between 0 and 1. This probability score represents the likelihood of the binary outcome belonging to a particular class [10]. The logistic regression model is trained using optimization techniques such as maximum likelihood estimation or gradient descent, where the model parameters are adjusted to minimize the difference between predicted probabilities and actual class labels in the training data [14].

3.2.3 Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the "naive" assumption of feature independence. It calculates the probability of a class given a set of features by multiplying the probabilities of each feature occurring in that class and normalizing by the probability of the features occurring together [16]. It requires a small amount of training data to estimate the necessary parameters, making it computationally efficient [22]. Naive Bayes comes in several variants, including Gaussian, Multinomial, and Bernoulli, each suited to different types of data. Gaussian Naive Bayes assumes that continuous features follow a Gaussian distribution, while Multinomial Naive Bayes is designed for discrete features often encountered in text classification tasks. Bernoulli Naive Bayes is suitable for binary feature vectors.

3.3 Prognosis Model

A cancer prediction is a classification or prognosis of whether a given tumor is benign or malignant. Two core frameworks of dimension reduction to be utilized for the model building are principle component analysis (including Kernl PCA and principal curves) 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 the new feature as used in [7]. Figure 4 aids in illustrating the prognosis model - where the specified data set (with the full feature set) is split into both a training and testing set. Each phase consists of dimension reduction which reduces the feature space, where these features are then used to generate the model in the training phase by using classification algorithms: SVM, Logistic Regression and Naive Bayes. After which, the generated model is used to classify whether a given tumor is benign or malignant in the testing phase.

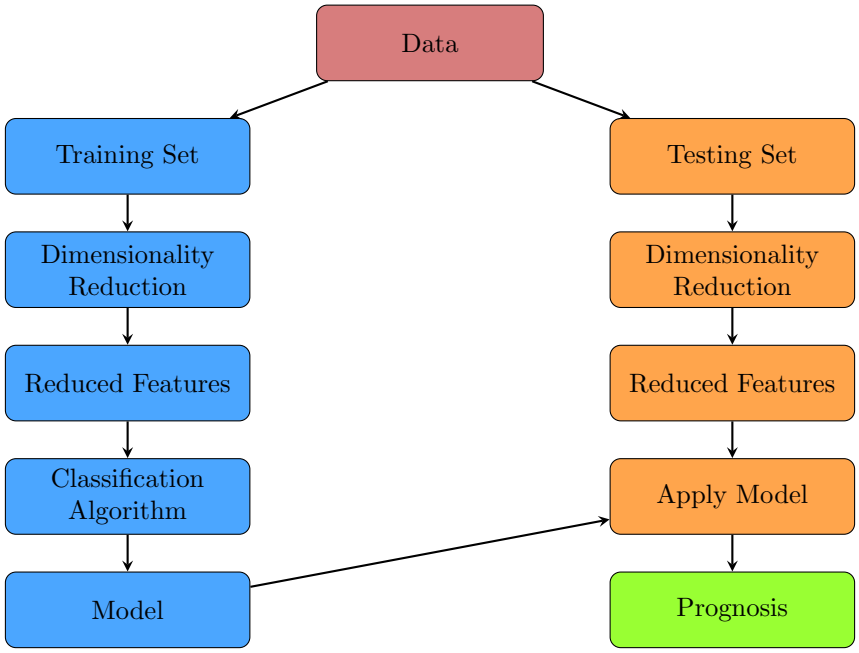


Figure 4: Flow Diagram of Prognosis Model

3.4 Visualization of Methodology

3.4.1 PCA Example (Using SVM and 80/20 Split)

We aim to shed light on the inner workings of the prognosis model by illustrating what occurs when we use PCA as the means of dimensionality reduction. PCA also serves to simplify the data by linearly altering the data and creating a new coordinate system with the largest retained variance. Figure 5 shows the first two principal coordinates after dimension reduction on the training set - we note the tumor type is not entirely separable so we plot the first three principal components (of the training data) in Figure 6.

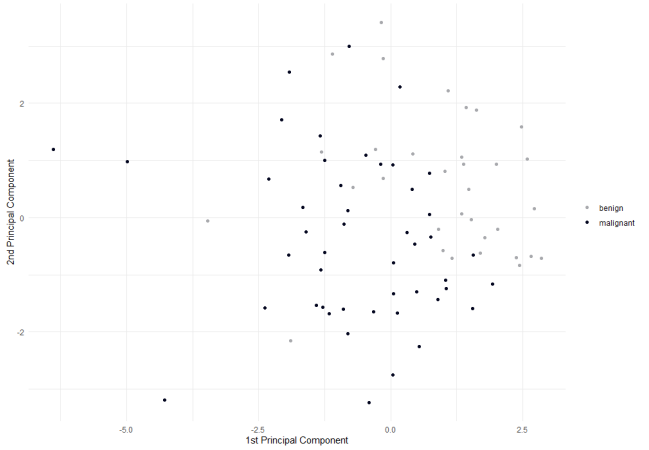


Figure 5: First 2 Principal Components for Training Data (80/20 split)

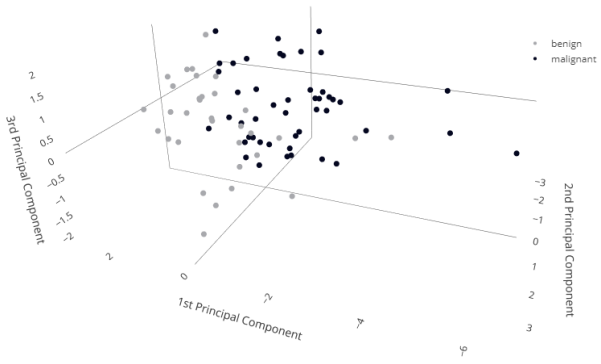


Figure 6: First 3 Principal Components for Training Data (80/20 split)

Using the first two principal components for dimension reduction in both the training phase (to build the classifier with SVM) and testing phase, we apply the model to the testing set with reduced features to make predictions. We compare these predictions to the response variable in the testing set - where each comparison is given as a true positive (TN), true negative (TN), false positive (FP) and false negative (FN) as seen in Figure 7. The same methodology applied to the first three principal components can be seen in Figure 8.

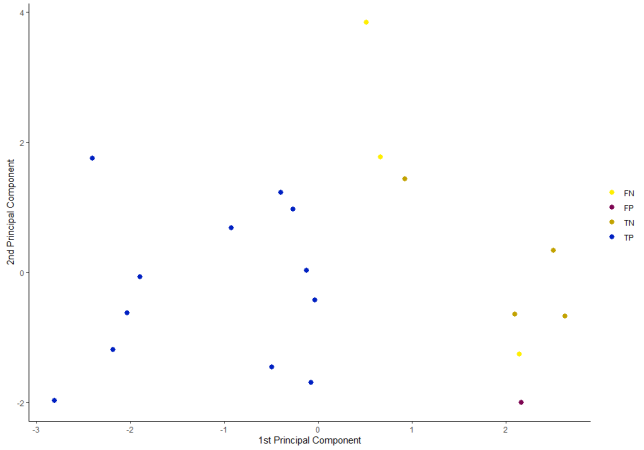


Figure 7: Model Predictions using 2 Principal Components (SVM and 80/20 split)

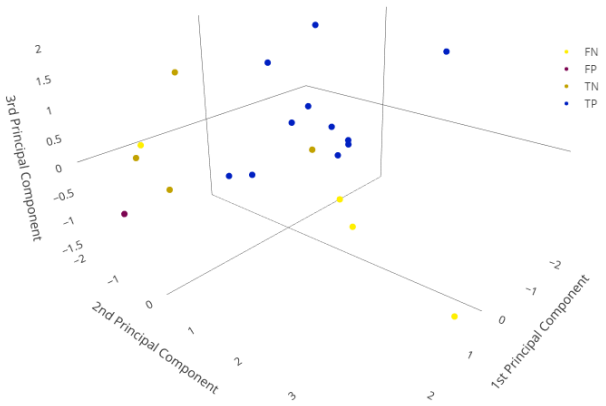


Figure 8: Model Predictions using 3 Principal Components (SVM and 80/20 split)

3.4.2 Clustering Example (Using 2 Clusters, SVM and 80/20 split)

Similarly, we visualize the prognosis methodology utilizing cluster analysis as the means of dimension reduction - where the new feature is the cluster to which the particular observation belongs. In this example, observations will be grouped into 2 clusters (whereby the new feature created will be the cluster to which the observation belongs- either cluster 1 or 2). Figure 9 shows clustering undergone on the training set, where we see most of the malignant tumors have been grouped to the first cluster.

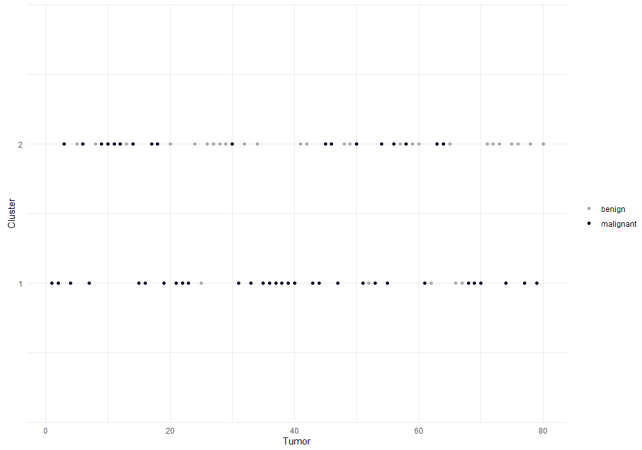


Figure 9: 2 Clusters on Training Data (SVM and 80/20 split)

Furthermore, after building the classifier using 2 clusters, we make predictions on the response variable of the testing set as seen in Figure 10.

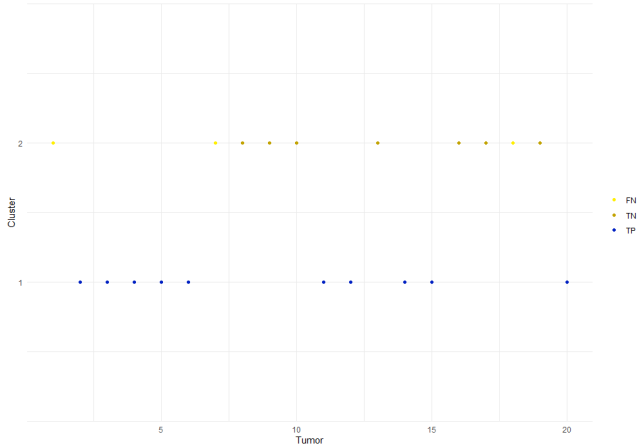


Figure 10: Model Predictions using 2 Clusters (SVM and 80/20 split)

References

[1] Andriy Burkov. *The hundred-page machine learning book*, volume 1. Andriy Burkov Quebec City, QC, Canada, 2019.

[2] Marie Chavent, Vanessa Kuentz, Benoˆıt Liquet, and L Saracco. Clustofvar: An r package for the clustering of variables. *arXiv preprint arXiv:1112.0295*, 2011.

[3] 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.

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

[5] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[6] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern recognition*, 40(3):863–874, 2007.

[7] 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.

[8] Shubhangi N Katole and Swapnili P Karmore. A new approach of microarray data dimension reduction for medical applications. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, pages 409–413. IEEE, 2015.

[9] Trupti M Kodinariya, Prashant R Makwana, et al. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[10] Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[11] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 1(1):24–45, 2004.

[12] Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. *Advances in neural information processing systems*, 11, 1998.

[13] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[14] Todd G Nick and Kathleen M Campbell. Logistic regression. *Topics in biostatistics*, pages 273–301, 2007.

[15] Alan P Reynolds, Graeme Richards, Beatriz de la Iglesia, and Victor J Rayward-Smith. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5:475–504, 2006.

[16] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

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

[18] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[19] Shan Suthaharan and Shan Suthaharan. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235, 2016.

[20] Eva Tuba, Ivana Strumberger, Timea Bezdan, Nebojsa Bacanin, and Milan Tuba. Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. *Procedia Computer Science*, 162:307–315, 2019.

[21] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer, 2005.

[22] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714, 2010.

[23] World Health Organization. Cancer, 2022.

[24] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[25] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.