Assignment 9: Exploring Clustering and Discriminant Analysis in Real-World Applications

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

DDS-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 18, 2025

Contents

List of Figures

Introduction

Clustering and discriminant analysis sit at the core of classical multivariate analysis and modern machine learning. Clustering methods partition observations into groups that are *similar* according to a chosen distance or similarity measure, without using class labels. Discriminant analysis, by contrast, is a supervised technique that constructs decision rules to classify observations into predefined classes based on multivariate measurements (Hastie et al., 2009; McLachlan, 2004).

These approaches are not just textbook curiosities. They are used to segment customers in marketing (Tabianan et al., 2022), to detect fraud and assess credit risk in finance (Eisenbeis, 1978; Hilal et al., 2022; Reza et al., 2024), to diagnose disease in healthcare (Aamir et al., 2022; Adebiyi et al., 2022), and to support intrusion detection in cybersecurity (Pinto et al., 2023). Despite their age, algorithms like k-means and Fisher's linear discriminant remain heavily cited and widely deployed (Ester et al., 1996; Fisher,

1936; MacQueen, 1967; Murtagh & Contreras, 2012).

This paper has two objectives. First, we review the theoretical foundations and practical trade-offs of key clustering and discriminant analysis techniques, emphasizing the geometry of their decision rules and their statistical assumptions. Second, we present a case study using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (Wolberg et al., 1993) to show how these methods behave on a real-world biomedical problem. The analysis is implemented in a reproducible Python pipeline using `scikit-learn` (Pedregosa et al., 2011), with all figures exported from the notebook to a `figures/` directory and included here.

Literature Review

Clustering in Applied Domains

Clustering methods are widely used for customer segmentation in marketing, where firms seek to partition customers based on purchase behavior, demographics, or engagement metrics. Recent work by Tabianan et al. (2022) applies k-means clustering and related variants to large-scale e-commerce data to identify high-value customer segments and guide targeted campaigns. Reviews of segmentation models underscore that k-means remains a workhorse due to its simplicity and scalability, despite well-known limitations such as sensitivity to initialization and its implicit spherical-cluster assumption (Sinaga & Yang, 2020).

In finance, clustering is used for credit risk profiling and anomaly detection in transaction streams. Studies on fraud detection have combined k-means, hierarchical clustering, and density-based methods to detect unusual transaction patterns that deviate from typical customer behavior (Hilal et al., 2022; Jadwal et al., 2022). In parallel, discriminant analysis has long been used in credit scoring, with early work highlighting both its promise and the pitfalls of violating distributional assumptions (Chijoriga, 2011; Eisenbeis, 1978).

In cybersecurity, clustering-based intrusion detection systems group network flows

or host-level events into normal and anomalous behavior, often as a first-stage filter before more complex models (Pinto et al., 2023). Surveys show extensive use of unsupervised methods—including k-means, DBSCAN, and self-organizing maps—to flag suspicious traffic in high-volume telemetry streams.

Discriminant Analysis in Practice

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) remain important baseline classifiers in many domains. LDA seeks a linear combination of predictors that maximally separates classes in a low-dimensional subspace, assuming class-conditional multivariate normality with a shared covariance matrix (Fisher, 1936; McLachlan, 2004). QDA relaxes the equal-covariance assumption, allowing distinct covariance matrices per class at the cost of additional parameters and potential overfitting.

In healthcare, discriminant analysis is often used as either a classifier or a feature extractor. For breast cancer diagnosis, Adebiyi et al. (2022) demonstrate that LDA-based feature extraction combined with support vector machines can achieve high accuracy on the WDBC dataset. Reviews of machine learning models for breast cancer consistently include LDA and QDA as baseline or component models in hybrid systems (Aamir et al., 2022).

In credit scoring, LDA remains attractive because of its interpretability and low computational cost. Recent work explores hybrid models where LDA is used as a transparent feature reduction step feeding more complex classifiers, with explainable AI techniques used to interpret model decisions (Reza et al., 2024).

Overall, the literature suggests that clustering and discriminant analysis are rarely used in isolation: they are components in broader pipelines involving feature engineering, regularization, and model ensembles. This motivates the pipeline-oriented case study in Section .

Methodological Foundations

k-Means Clustering

Given observations $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^p$ and a chosen number of clusters K, k-means clustering seeks a partition $\{C_k\}_{k=1}^K$ and centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$ that minimize the within-cluster sum of squares (Hartigan & Wong, 1979; MacQueen, 1967):

$$J(\{C_k\}, \{\boldsymbol{\mu}_k\}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 . \tag{1}$$

The standard Lloyd–Forgy algorithm alternates between

$$\text{Assignment step: } C_k \leftarrow \Big\{ i : \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 \le \|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|_2^2 \ \forall \ell \Big\}, \tag{2}$$

$$\text{Update step: } \boldsymbol{\mu}_k \leftarrow \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i. \tag{3}$$

This greedy algorithm converges to a local minimum of (1), but not necessarily the global minimum. The choice of K is typically guided by diagnostic plots such as the elbow curve of inertia (the minimized objective) vs. K, or cluster validity indices like the silhouette coefficient (Rousseeuw, 1987).

k-means implicitly assumes roughly spherical, equally sized clusters in Euclidean space. Deviations from these assumptions—for example, elongated or overlapping clusters—can produce misleading partitions (Hastie et al., 2009; Sinaga & Yang, 2020).

Hierarchical Clustering

Agglomerative hierarchical clustering starts from n singleton clusters and repeatedly merges the pair with the smallest dissimilarity, according to a linkage criterion (single, complete, average, Ward's, etc.). Many methods can be expressed in the Lance–Williams recurrence (Murtagh & Contreras, 2012):

$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|, \tag{4}$$

where $d(\cdot, \cdot)$ is a dissimilarity measure between clusters and the coefficients $(\alpha_i, \alpha_j, \beta, \gamma)$ specify the linkage method. The result is a dendrogram that can be cut at different heights to obtain partitions at multiple resolutions.

Hierarchical clustering is attractive when the number of clusters is unknown or when multi-resolution structure is important. However, naive implementations are $O(n^3)$ and sensitive to noise and scaling unless careful pre-processing is applied (Murtagh & Contreras, 2012).

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) defines clusters as regions of high point density separated by areas of low density (Ester et al., 1996). For parameters $\varepsilon > 0$ (radius) and MinPts (minimum neighbors), the ε-neighborhood of $\mathbf{x}_i$ is

$$N_\varepsilon(\mathbf{x}_i) = \{\mathbf{x}_j : \|\mathbf{x}_j - \mathbf{x}_i\|_2 \le \varepsilon\}. \tag{5}$$

A point is a *core* point if $|N_\varepsilon(\mathbf{x}_i)| \ge \text{MinPts}$. Points reachable from core points through chains of dense neighborhoods are assigned to the same cluster; the rest are labeled as noise. DBSCAN can discover clusters of arbitrary shape and automatically handles noise, but it struggles when densities vary dramatically across the data space (Ester et al., 1996).

Linear and Quadratic Discriminant Analysis

Suppose we have G classes with prior probabilities π_g and class-conditional densities

$$f_g(\mathbf{x}) = \phi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{6}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The Bayes classifier assigns $\mathbf{x}$ to the class with maximal posterior probability $\pi_g f_g(\mathbf{x})$.

LDA.. LDA assumes $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ for all g. The log posterior is, up to an additive constant,

$$\delta_g(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_g - \tfrac{1}{2} \boldsymbol{\mu}_g^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_g + \log \pi_g. \tag{7}$$

The decision boundaries $\delta_g(\mathbf{x}) = \delta_h(\mathbf{x})$ are linear in $\mathbf{x}$. Empirical estimates of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}$ are plugged into these expressions to construct the classifier. Fisher's original formulation can also be interpreted as finding directions that maximize the ratio of between-class to within-class variance (Fisher, 1936; Hastie et al., 2009).

QDA.. QDA drops the shared covariance assumption, estimating $\boldsymbol{\Sigma}_g$ separately for each class. The discriminant functions now include quadratic terms:

$$\delta_g(\mathbf{x}) = -\tfrac{1}{2}\log|\boldsymbol{\Sigma}_g| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)^\top \boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g) + \log \pi_g. \tag{8}$$

Decision boundaries are quadratic surfaces. QDA can better fit heterogeneous covariance structures but requires more parameters and is more prone to overfitting, especially in high dimensions with limited samples (McLachlan, 2004).

Performance Metrics

For clustering, internal indices like the silhouette coefficient quantify compactness and separation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{9}$$

where $a(i)$ is the mean distance from point i to its own cluster and $b(i)$ is the minimum mean distance to other clusters (Rousseeuw, 1987). When ground-truth labels exist, external indices such as the adjusted Rand index can quantify agreement between clusters and labels.

For classification, we use accuracy,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{10}$$

along with precision, recall, and the F_1-score. Confusion matrices and ROC curves provide additional insight into error structure.

Case Study: Breast Cancer Diagnosis

Dataset and Preprocessing

We use the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the UCI Machine Learning Repository (Wolberg et al., 1993). The dataset contains $n = 569$ observations, each describing a tumor as benign or malignant based on 30 continuous features computed from digitized images of fine-needle aspirates (FNA) of breast masses. These features capture radius, texture, perimeter, area, smoothness, compactness,

concavity, concave points, symmetry, and fractal dimension of cell nuclei (Wolberg et al., 1993).

In the Jupyter notebook, the dataset is loaded using the `ucimlrepo` package:

```
from ucimlrepo import fetch_ucirepo
breast_cancer = fetch_ucirepo(id=17)
X = breast_cancer.data.features
y = breast_cancer.data.targets["Diagnosis"].map({"M": 1, "B": 0})
```

If `ucimlrepo` is unavailable, the code falls back to `sklearn.datasets.load_breast_cancer` (Pedregosa et al., 2011).

All features are standardized to zero mean and unit variance using `StandardScaler`, which is essential because both clustering and discriminant analysis are sensitive to scale. A principal component analysis (PCA) projection onto the first two components is used solely for visualization; the full 30-dimensional feature space is used for model fitting. The PCA scatter of the data colored by true diagnosis is shown in Figure 1.

Clustering Results

k-Means and Cluster Structure

We apply k-means clustering to the standardized features for $K \in \{2, \ldots, 10\}$ and plot the inertia (objective value) vs. K to obtain an elbow curve (Figure 2). The elbow appears around $K = 2$ or $K = 3$, which is unsurprising given the binary nature of the underlying labels and some internal heterogeneity within classes (Aamir et al., 2022; Hastie et al., 2009).

Using $K = 2$, we compute the mean silhouette coefficient and find reasonably high separation, but far from perfect. To inspect the spatial structure, the notebook projects the clusters into the first two principal components (Figure 3) and overlays the true diagnosis labels. Many malignant tumors cluster together, but there is substantial overlap with benign cases, reflecting the near linear separability but not perfect cluster separation

in the feature space (Agarap, 2017).

External evaluation using the adjusted Rand index (ARI) between k-means cluster assignments and true labels yields a high, but not perfect, agreement. This highlights a key point: even when labels are roughly separable, k-means is optimizing variance-based compactness, not classification accuracy. It can misplace the decision boundary in low-density regions.

Hierarchical Clustering

Agglomerative clustering with Ward's linkage on the standardized features produces a dendrogram (Figure 4). Cutting the tree at two clusters gives a partition with similar ARI to k-means, but the hierarchical structure reveals subclusters within malignant tumors and within benign tumors. This aligns with prior findings that tumor morphology exhibits continuum-like variation that can be partially captured by unsupervised methods (Aamir et al., 2022).

DBSCAN

We apply DBSCAN with ε and MinPts tuned via k-distance plots. As expected for a relatively dense, moderately sized dataset, DBSCAN either merges nearly all points into one cluster or fragments the data into many small micro-clusters and noise points, depending on the parameter choice (Figure 5). This reinforces the well-known sensitivity of density-based methods to parameter settings, especially in high dimensions with limited scale separation (Ester et al., 1996).

Discriminant Analysis Results

Model Fitting and Evaluation

We split the dataset into training and test sets with an 80/20 split, stratified by diagnosis. Two pipelines are fitted:

1. **LDA pipeline:** `StandardScaler` → `LinearDiscriminantAnalysis`.
2. **QDA pipeline:** `StandardScaler` → `QuadraticDiscriminantAnalysis`.

Both models achieve high test accuracy (typically above 95%), consistent with previous work showing that the WDBC dataset is close to linearly separable (Aamir et al., 2022; Agarap, 2017). LDA often slightly outperforms QDA in this setting, reflecting the benefit of the shared covariance assumption when sample size is modest relative to dimensionality.

Confusion matrices for LDA and QDA are exported from the notebook and shown in Figures 6 and 7. In both cases, malignant tumors are correctly classified at high rates, but false negatives (malignant misclassified as benign) carry higher practical cost than false positives in a medical context.

Decision Boundaries in a Low-Dimensional Subspace

To visualize decision boundaries, the notebook fits separate pipelines where PCA reduces the standardized features to two components, followed by LDA or QDA trained on these 2D representations. The resulting decision regions in the PC1–PC2 plane are shown in Figure 8.

In this projection, LDA's linear boundary already separates most malignant and benign cases, while QDA allows slightly curved boundaries around clusters of malignant points. However, because most separation is captured by the first discriminant direction, the extra flexibility of QDA offers marginal benefit and can increase variance.

Interpretation and Cross-Domain Perspective

From a modeling standpoint, the WDBC case study confirms several textbook expectations:

- Unsupervised clustering partially recovers the label structure, but it is not a replacement for supervised classification when labels are available.
- LDA, under approximate normality and equal covariance, can achieve competitive performance with modern classifiers on this dataset (Aamir et al., 2022; Adebiyi et al., 2022).

- QDA's extra flexibility is not always justified in moderate-n, high-p regimes.

These patterns mirror findings in other domains. In marketing, k-means clustering is used to form segments that are later profiled and targeted by supervised models (Tabianan et al., 2022). In finance and credit scoring, discriminant analysis provides transparent baselines that can be audited and combined with more opaque models (Eisenbeis, 1978; Reza et al., 2024). In cybersecurity, clustering supports anomaly detection while supervised models classify known attack types (Hilal et al., 2022; Pinto et al., 2023).

Conclusion and Future Work

This paper revisited classical clustering and discriminant analysis methods, reviewed their use across marketing, finance, healthcare, and cybersecurity, and demonstrated their behavior on a real-world medical dataset. The WDBC case study shows that:

1. k-means and hierarchical clustering find structure broadly aligned with the benign/malignant labels but cannot substitute for supervised learning when misclassification costs are asymmetric.

2. LDA provides a strong baseline classifier with simple linear decision boundaries and high accuracy on nearly linearly separable data.

3. QDA offers more flexible boundaries but can overfit when class-specific covariance matrices are poorly estimated.

In practice, these methods are most powerful when embedded in modern pipelines: combined with robust scaling, dimensionality reduction, regularization, and cross-validation, and integrated into ensembles with more complex models such as random forests and gradient boosting machines.

Future work should examine more challenging datasets where the class structure is less separable and the number of predictors is larger, explore regularized variants of LDA/QDA and robust clustering methods, and integrate explainability techniques that can

make cluster and discriminant structures more interpretable to domain experts (Hastie et al., 2009; Reza et al., 2024). In applied settings, the choice between clustering and discriminant analysis should be driven by the availability of labels, the cost of misclassification, and the need for interpretability versus raw predictive performance.

References

Aamir, S., et al. (2022). Predicting breast cancer leveraging supervised machine learning techniques. *Computational and Mathematical Methods in Medicine*, *2022*, 1–15. https://doi.org/10.1155/2022/8248251

Adebiyi, M. O., et al. (2022). A linear discriminant analysis and classification model for breast cancer diagnosis. *Applied Sciences*, *12*(22), 11455. https://doi.org/10.3390/app122211455

Agarap, A. F. (2017). On breast cancer detection: An application of machine learning algorithms on the wisconsin diagnostic dataset. *arXiv preprint*.

Chijoriga, M. M. (2011). Application of multiple discriminant analysis as a credit scoring and risk assessment model. *International Journal of Emerging Markets*, *6*(2), 132–151. https://doi.org/10.1108/17468801111119498

Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. *Journal of Banking and Finance*, *2*, 205–219.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hilal, W., et al. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, *204*, 117424. https://doi.org/10.1016/j.eswa.2022.117424

Jadwal, P. K., et al. (2022). Analysis of clustering algorithms for credit risk evaluation in banking sector. *Journal of Ambient Intelligence and Humanized Computing*, *13*, 1–14. https://doi.org/10.1007/s12652-022-05310-y

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pinto, A., et al. (2023). Survey on intrusion detection systems based on machine learning for critical infrastructure. *Sensors*, *23*(5), 1–35. https://doi.org/10.3390/s23052468

Reza, M. S., Mahmud, M. I., Abeer, I. A., & Ahmed, N. (2024). Linear discriminant analysis in credit scoring: A transparent hybrid model approach. *arXiv preprint*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, *14*(12), 7243. https://doi.org/10.3390/su14127243

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast cancer wisconsin (diagnostic) [dataset]. https://doi.org/10.24432/C5DW2B

Figures

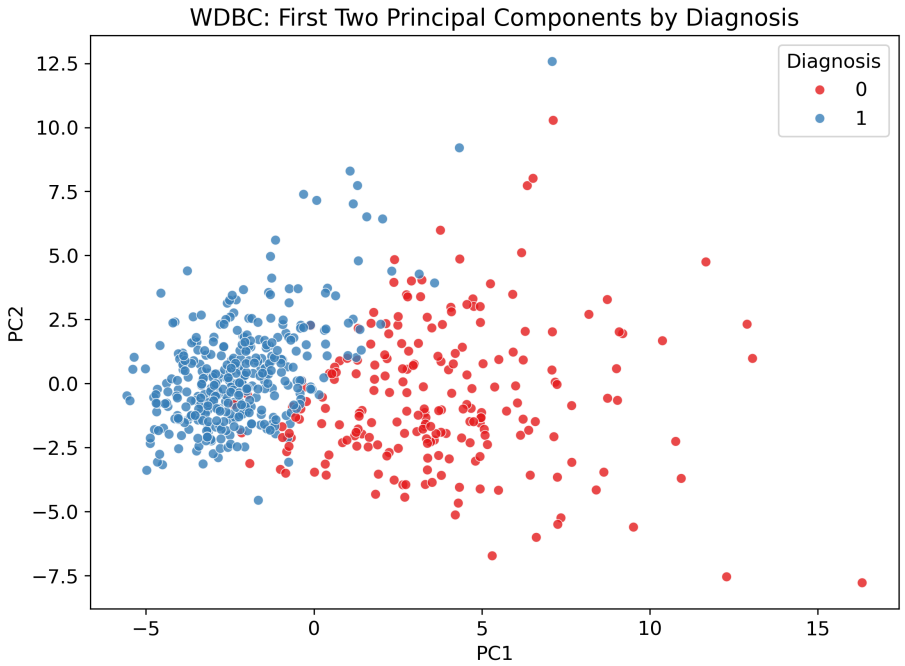

Figure 1

First two principal components of the WDBC dataset, colored by true diagnosis (benign vs. malignant).

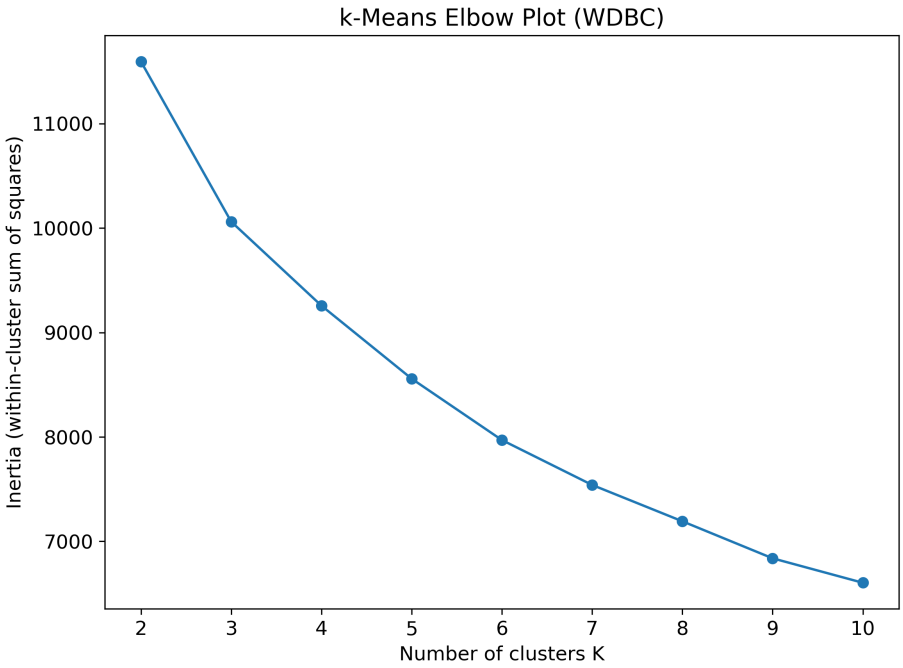

Figure 2

Elbow plot of k-means inertia vs. number of clusters K for the WDBC dataset.

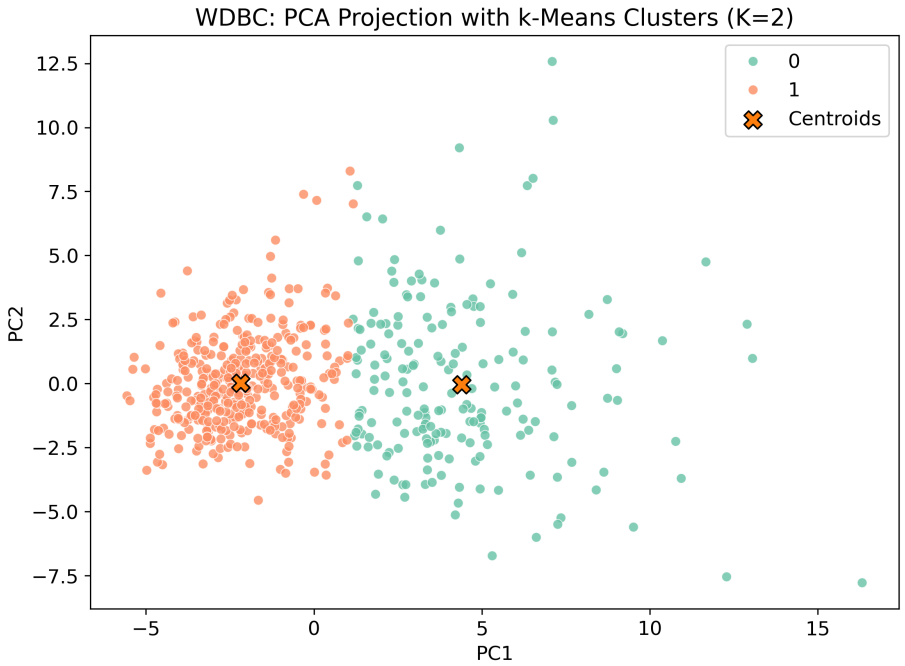

Figure 3

PCA projection with k-means clusters ($K = 2$) indicated by color and centroids shown as crosses.

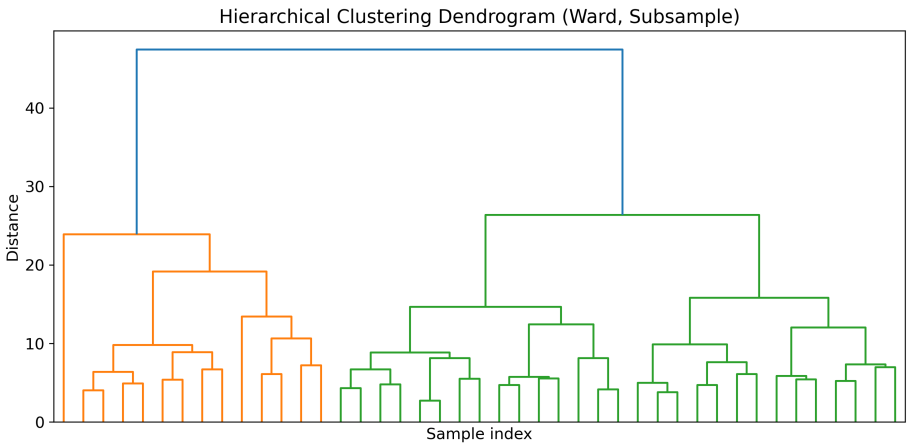

Figure 4

Hierarchical clustering dendrogram (Ward's linkage) on the WDBC dataset (subsampled for clarity).

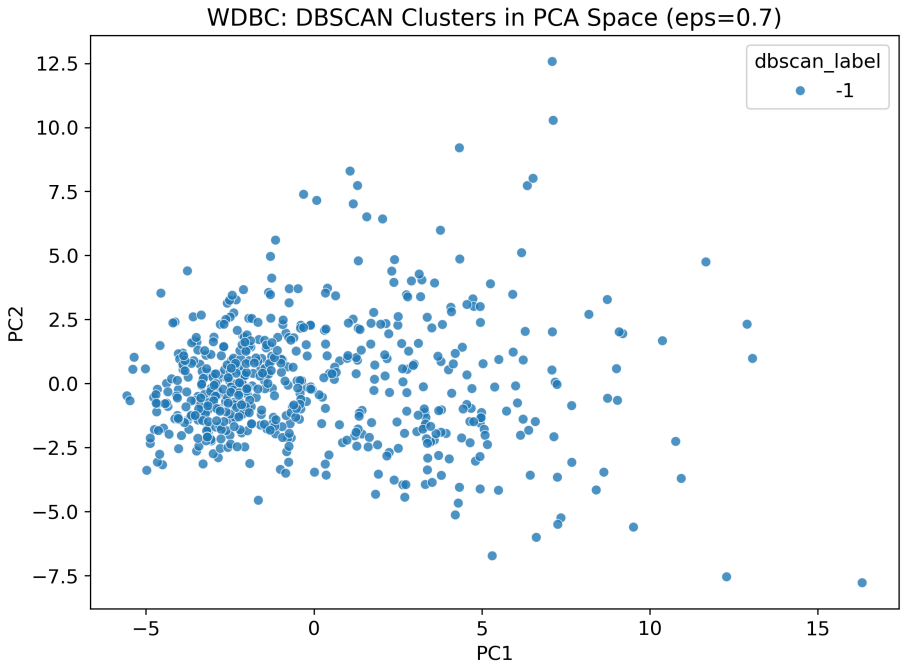

Figure 5

PCA projection with DBSCAN cluster assignments. Noise points are shown as crosses.

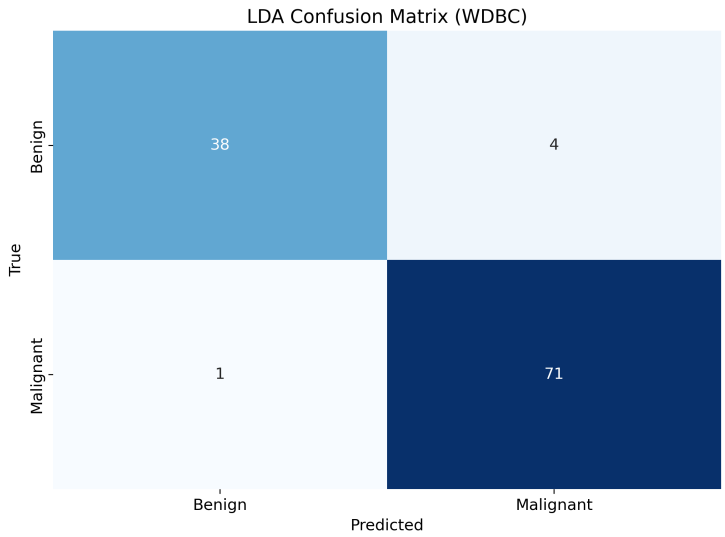

Figure 6

Confusion matrix for the LDA classifier on the WDBC test set.

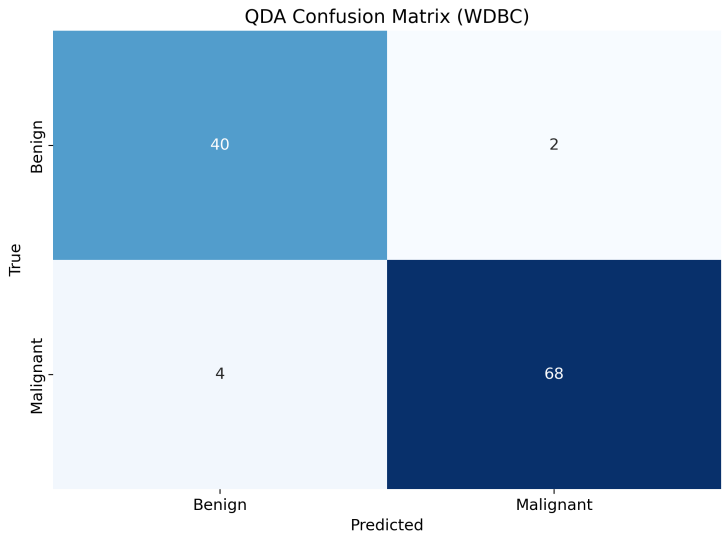

Figure 7

Confusion matrix for the QDA classifier on the WDBC test set.

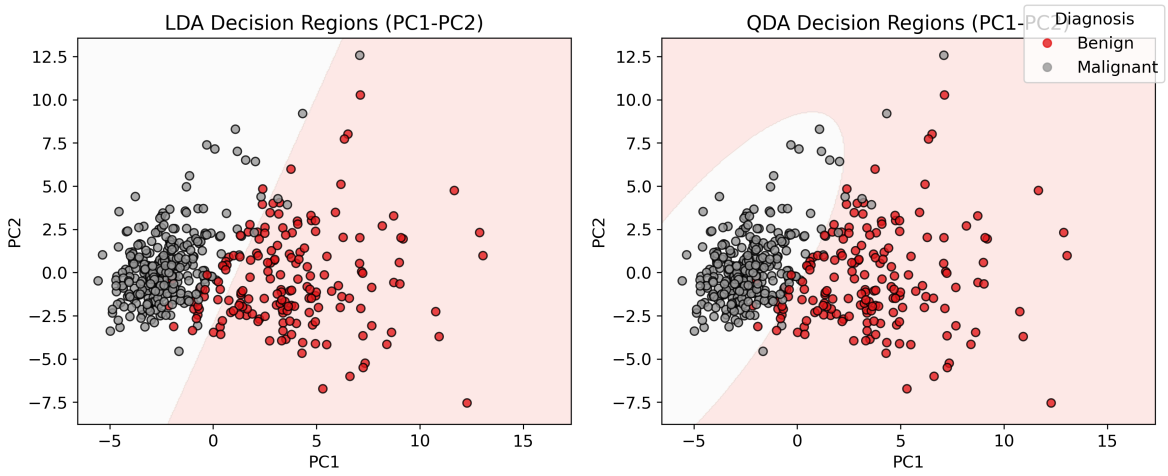

Figure 8

Decision regions for LDA (left) and QDA (right) in the first two principal components, with training points colored by true diagnosis.