DATA70132 Report

Classification of Vertebral Column Data

1. Unsupervised Method

K-Means is a clustering algorithm that separates unlabelled data points into k clusters, in which points are near each other in terms of Euclidian distance. The algorithm is a hard classifier and depends on the initial assignment of cluster centres (mean position of points in cluster), which can be picked randomly, or through approaches like Forgy's method or Maximin. The process of assigning points to clusters is iterative, since after each assignment, new centroids are calculated, and the points are reassigned until convergence. The objective function of the algorithm is to minimise the within-cluster sum of squares (WCSS):

$$\min_{C} \sum_{j=1}^{k} \sum_{x\, in C_j} |x - \mu_j|^2.$$

The number of clusters (k) needs to be specified, most commonly through an elbow plot, which visualises the k for which the WCSS sharply decreases.

However, the algorithm is affected by the choice of initial centroids and can also perform poorly when clusters are not spherical. Furthermore, the choice of k is not clear-cut task.

2. Supervised Method

Random Forest is an ensemble supervised learning method combining multiple decision trees used for hard classification. Single decision trees are non-parametric models that segment the feature space based on certain conditions to make predictions, but they are prone to overfitting. Random Forests (RF) seek to remedy this by introducing randomness in data selection (bootstrap aggregation) and feature subset selection. Hence, each tree is trained on random sample and at each split, a random selection of features is considered, to diversify trees and minimise overfitting. However, the same data point can end up multiple times in a single sample (bootstrapping with replacement). Nodes in trees are constructed by minimising the probability of misclassification of a data point, based on measures such as Gini Index or Entropy.

The classification (H) of a data point is assigned by a majority vote out of all bootstraps, which is the mode of class outputs for all individual trees:

$$H(x) = \text{mode}\{h(x, \Theta_1), h(x, \Theta_2), \ldots, h(x, \Theta_K)\}.$$

However, RF's decision-making process can be hard to interpret due to the multitude of trees, and can still be affected by class imbalance and overfitting if not properly tuned.

3. EDA

Variable	Description	Type	Missing Values
Pelvic_Incidence	Measure of spinal deformity, sum of pelvic tilt (PT) and sacral slope	Numeric	None
Pelvic_Tilt	Measure of orientation of the pelvis	Numeric	None
Lumbar_Lordosis_Angle	Measure of curve of the lumbar region	Numeric	None
Sacral_Slope	Measure of angle between sacral plate and a horizontal plane	Numeric	None
Pelvic_Radius	Measure of distance from the hip axis to the S1 endplate	Numeric	None
Grade_of_Spondylolisthesis	Measures grade of severity	Numeric	None

The "vertebral_column_data.txt" contains 6 continuous biomechanical indicators used for classification of orthopaedic patients into 2 categories: normal (100 instances) and abnormal (210 instances). None of the variables have missing values.

Figure 1 shows the presence of outliers in all 6 orthopaedic indicators. This can affect the k-means algorithm by influencing the centroids, leading to misleading results. Hence, a log transform with a small constant (for negative values) was performed to reduce the effect of outliers.

Figure 1. Combined Boxplots

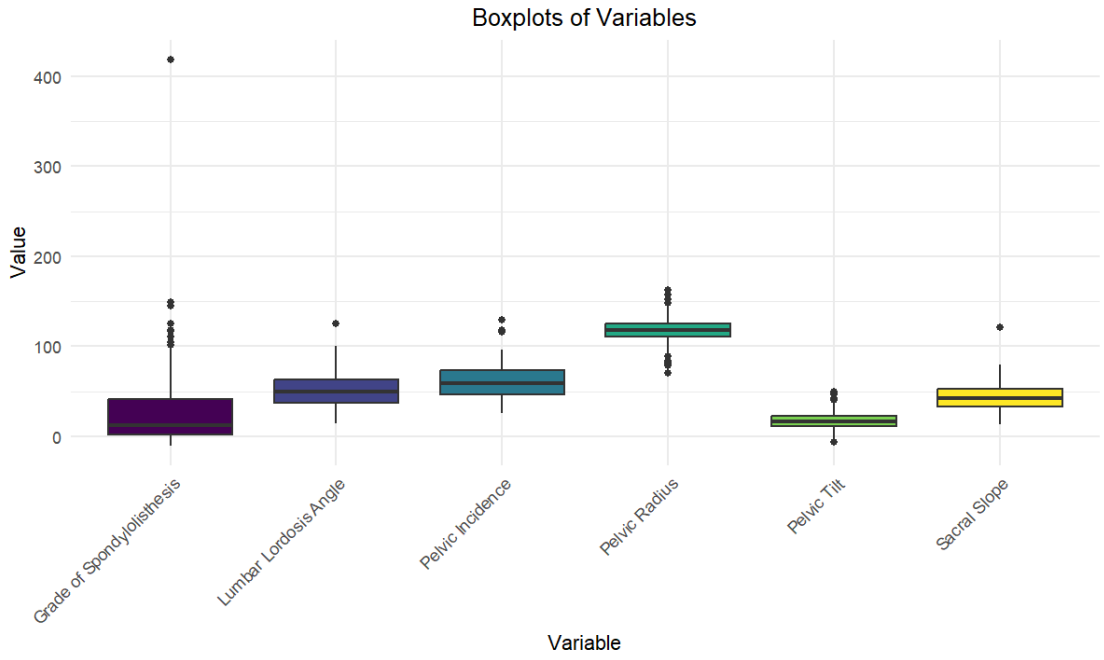


Figure 2. Correlation Matrix

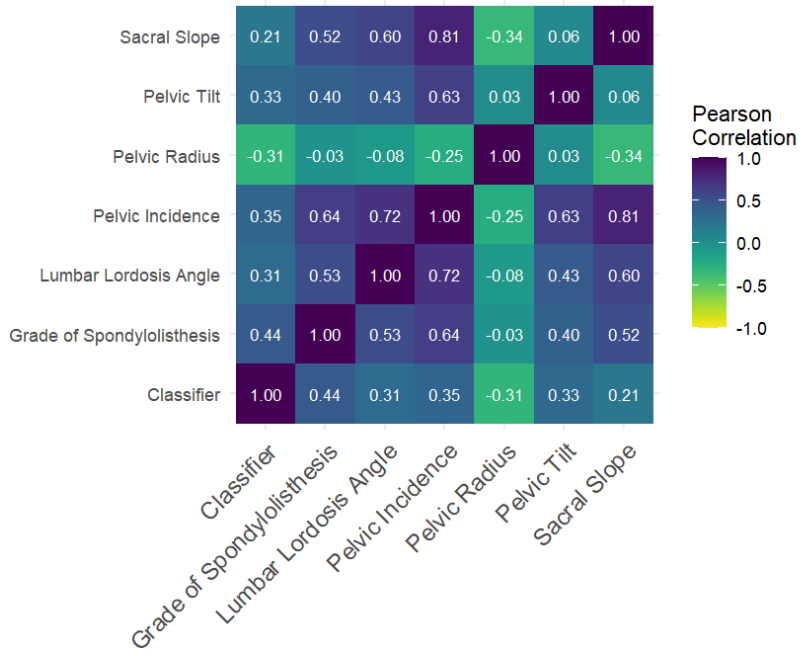


Figure 2. shows presence of problematic multicollinearity between variables, specifically “Sacral_Slope” and “Pelvic_Incidence” (0.81), which could reflect the

inherent relationship between them (“Pelvic_Incidence” is the sum of “Sacral_Slope” and “Pelvic_Tilt”). This suggests that features are providing redundant information which can distort the result of K-Means, hence, Principal Component Analysis (PCA) was performed to transform the data into an uncorrelated feature space. Variables underwent z-score standardization to normalize their scale, crucial for PCA and k-means clustering, which are sensitive to variations in scale among variables.

Figure 3. KDE Plots

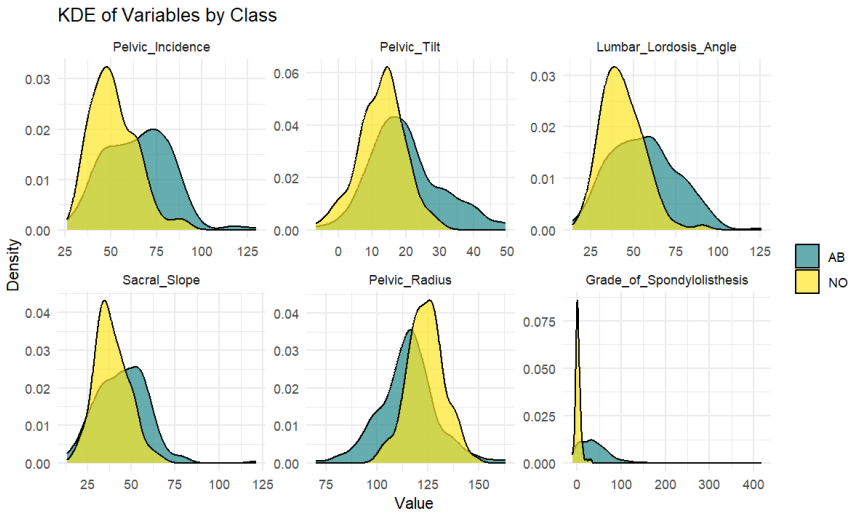


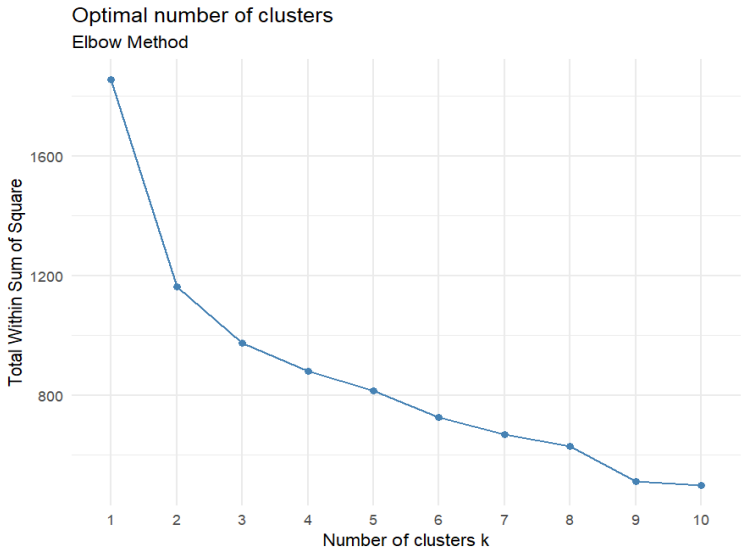
Figure 3. shows the KDE graphs for all features by class. “Grade_of_ Spondylolisthesis” seems to be particularly good for distinguishing between classes, since it shows the least overlap in densities.

RF’s ability to handle multicollinearity and outliers meant no preprocessing in this area was needed. However, due to class imbalance, the algorithm is biased towards the positive class AB (210 instances). To tackle this issue, SMOTE (Synthetic Minority Over-sampling Technique) was employed to achieve class balance without overfitting or reducing the dataset size, which are problems in upsampling and downsampling, respectively. This generates synthetic observations in the cross-validation training sets only to avoid data leakage.

4. Results

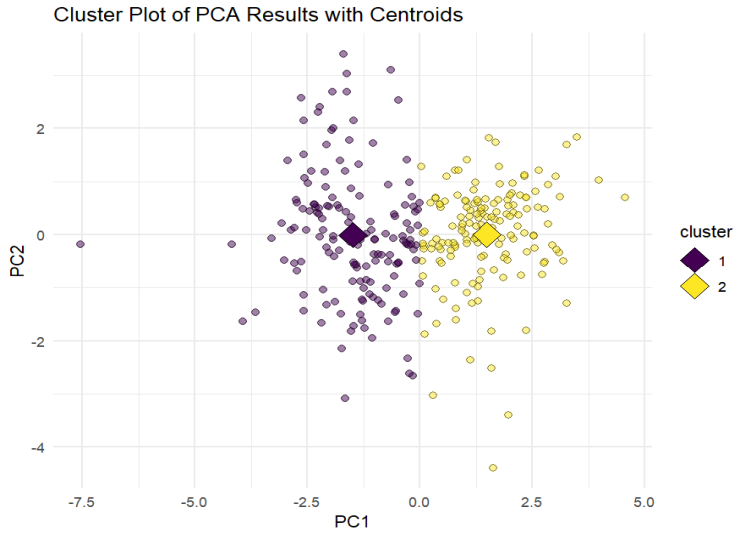
K-Means

Figure 4. Elbow Plot



The elbow plot (Figure 4) shows that the WCSS sharply decreases for 2 clusters. However, there is also a significant amount of decrease from 2 to 3 clusters.

Figure 5. K-Means Results



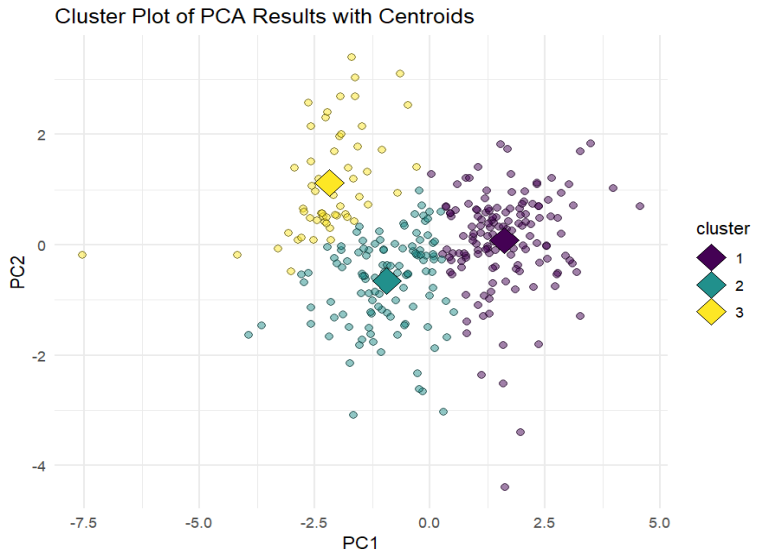
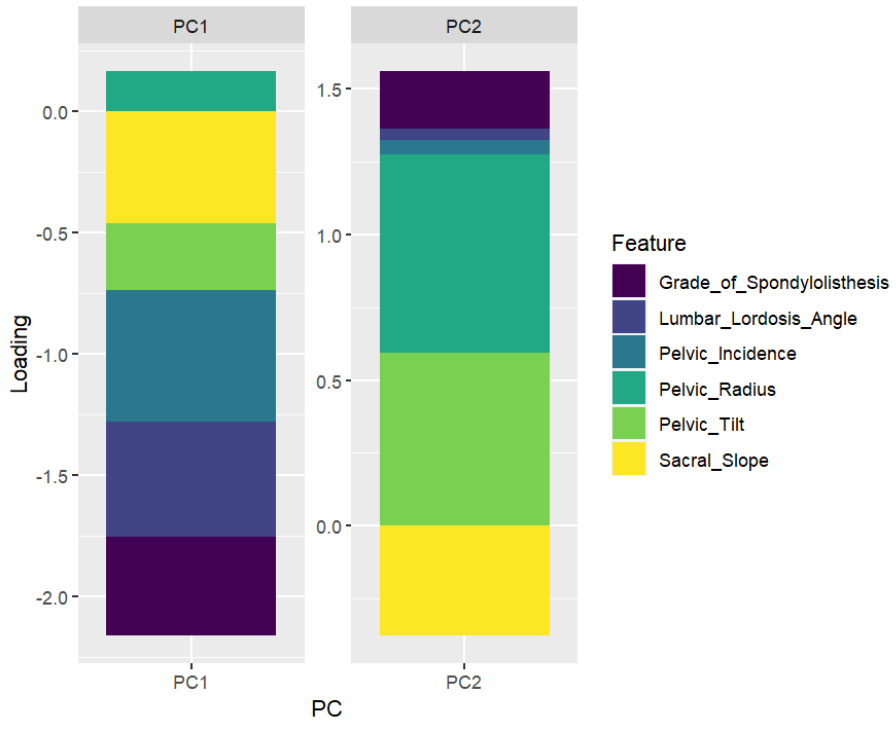


Figure 5 shows the results of the K-Means algorithm through random initialisation for 2 and 3 clusters, respectively, mapped on the feature space of PC1 and PC2. At 4 clusters, only 1 point is assigned to a cluster, demonstrating that 2/3 clusters are better suited to the data. This is also verified by the metadata, which shows that patients can be either classified as abnormal/normal, or with hernia/spondylolisthesis/normal (Barreto and Neto, 2011). When k=2, cluster 1 is more elongated, which is problematic for the performance of the algorithm.

Figure. 6 PC1/PC2 Loadings



PC1 and PC2 explain 72% of the variance of the data. PC1 explains the most variance (57%), showing PCA's effectiveness on this dataset. Figure 6 demonstrates that PC1 is relatively evenly influenced by "Sacral_Slope", "Lumbar_Lordosis_Angle" and "Pelvic_Incidence", while the biggest PC2 contribution comes from "Pelvic_Radius". These results make assessing the predictive value of individual variables complex, which is something that can be addressed through supervised learning.

Random Forest

For the supervised learning RF, the dataset was partitioned into train and test subsets for model evaluation. Furthermore, 10-fold cross-validation and hyperparameter grid tuning was employed to ensure the model generalizes well on unseen information across different data subsets, with Gini Index as a node impurity measure.

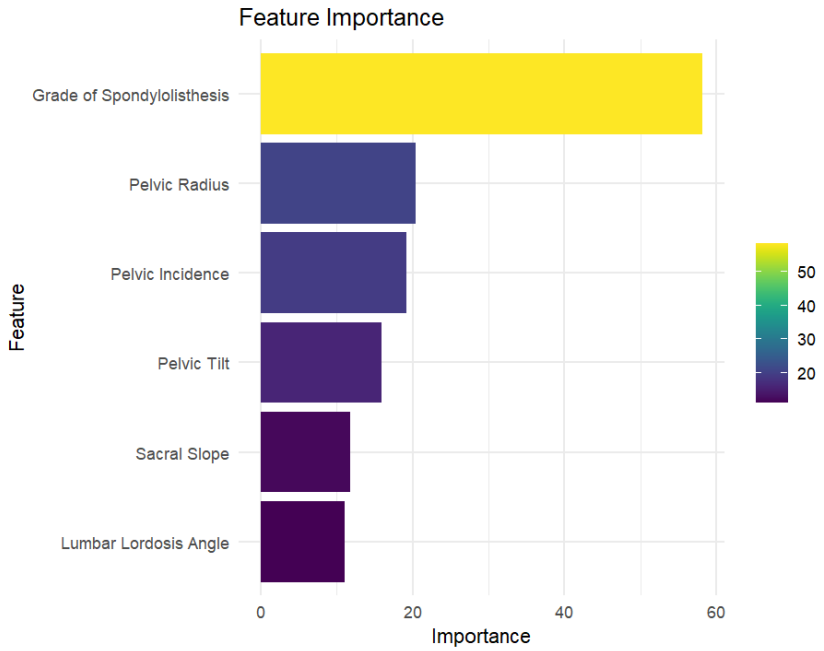
Table 1. Confusion Matrix

		Actual	
		AB	**NO**
Predicted	**AB**	38 (TP)	5 (FP)
	NO	4 (FN)	15 (TP)
Sensitivity	0.90		

Table 1 displays the correct and incorrect predictions for abnormal and normal patients, as well as the sensitivity of the model. Sensitivity is calculated by dividing the true positives by its sum with false negatives and is especially significant in medical classification, where the magnitude of false negative diagnosis has crucial bearing. The RF model has 90% sensitivity, which demonstrates it performs well at correctly identifying abnormal patients.

One of the advantages of RF is that it displays feature importance based on Gini Impurity (Figure 7). "Grade_of_Spondylolisthesis" is the most important predictor for abnormal/normal classification, which confirms the observations from the EDA.

Figure 7. Feature Importance



Unsupervised learning methods can be harder to interpret, as assessing the performance of the algorithm is complex without knowledge of the underlying data structure. However, it can also show hidden patterns in the data, such as the presence of more or different classes, which supervised learning methods cannot perform.

Moreover, both K-Means and RF can serve different purposes for clinicians. For example, RF could be beneficial for integrating classification of CT scan outputs based on previous data. In cases of under-researched diseases, K-Means can be a helpful diagnostic tool for clinicians to identify different groups of patients. Additionally, exploring fuzzy classification methods could provide more nuanced insights, such as patient-specific risk assessment for medical conditions.

Bibliography

Barreto, G. and Neto, A. (2011). Vertebral Column. UCI Machine Learning Repository. https://doi.org/10.24432/C5K89B.