Professor Lorenzo Pellis
REPORT

VERTEBRAL COLUMN DATA ANALYSIS REPORT

Student ID: 11356488
[Word Count : 1095]

MSc Data Science
The University of Manchester
DATA70132| Statistics and Machine Learning 2

Table of Contents

SECTION 1

1. Unsupervised Classification Method:

1. K-Means Clustering

K-means clustering is a popular unsupervised clustering algorithm used for grouping unlabelled data into 'K' different clusters.

One of the methods choosing the optimal no. of 'K' clusters[1] is:

Elbow method using scree plot – From the plot between within-cluster sum of squares (WCSS) against the no. of clusters, we identify the elbow-point beyond which the rate of decrease of the WCSS slows down. Beyond this adding more clusters are not reducing the distance significantly.

Within-cluster sum of squared distance (minimised):

$$d(x, u) = \min_{C} \sum_{j=1}^{k} \underbrace{\sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \boldsymbol{\mu}_j||_2^2}_{\text{Within } j\text{-th cluster}}.$$

Steps for K-means algorithm:

Step 1: Initialise 'K' centroids[2] randomly or using k-means++ algorithm.

Step 2: Assign data points to nearest clusters.

Step 3: Calculating mean of data points in each cluster and updating K cluster centroids.

Step 4: Repeating Steps 2 and 3 until centroids are stable.4

Step 5: fitting the training data points and predicting cluster labels.

Centroid of 'j'th cluster:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

Advantages:

1. Simple and computationally efficient
2. Clusters formed are easy to interpret.

Limitations:

1. Sensitive to initialisation[3] (K clusters)
2. Assumes circular cluster making it ineffective for non-linear dataset.

SECTION 2

2. Supervised Classification Method:

1. Support Vector Machines(SVM):

Support vector machines is one of the most powerful classification techniques in machine learning. The main objective of SVM[4] is to find an optimal hyperplane or line that segregates data points belonging to distinct classes.

Requirements:

Case 1: When classes are completely separated – SVM aims to maximize the distance between the hyperplane and the nearest data points (known as support vectors) from each class.

Case 2: When classes Overlap - SVM aims to minimizes the width of the margin to accommodate the overlapping regions and reduce misclassifications.

Parameter choices:

- **'C' -** 'C' is a regularization parameter that manages the trade-off between maximizing margin and minimizing misclassification errors. A smaller 'C' means wider margin, higher misclassification error. Larger 'C' leads to narrower margin, penalizes misclassifications.
- **'gamma' –** defines how far the influence f a single training sample reaches. Low 'Y' gives smoother decision boundaries whereas higher 'Y' leads to complex boundaries which captures intricate patterns.
- **'kernel'** – They offer strategies for constructing decision boundaries[4].

Kernel(Listing a few)	**Kernel Function (K)** $K(xr, xs) = \phi(xr) \cdot \phi(xs)$	**Use-Case**
1. **Linear Kernel**	$K(xr, xs) = xr \cdot xs$	Suitable for linearly separable datasets divided by straight line.
2. **Polynomial Kernel**	$K(xr, xs) = (\alpha(xr \cdot xs) + c)\ m$	Effective for dataset with non-linearity between features.
3. **Sigmoid Kernel**	$K(xr, xs) = \tanh(\alpha(xr \cdot xs) + c)$	Used when the data shows a periodic or wave like pattern
4. **RBF Kernel** (Used for the given dataset)	$K(xr, xs) = \exp(-\|\|xr - xs\|\|2\ 2/2\sigma\ 2\ )$	Suitable for datasets that have complex and non-linear decision boundaries (overlapping).

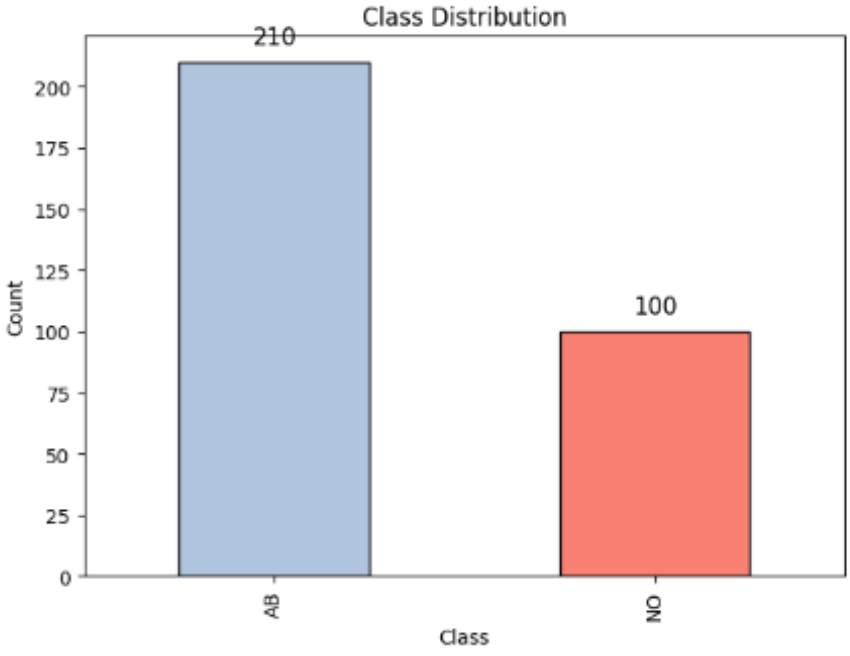
Limitations:

- Training SVM with RBF can be computationally intensive[5].
- Susceptible to noise as RBF kernel fits intricate data reducing generalisation.

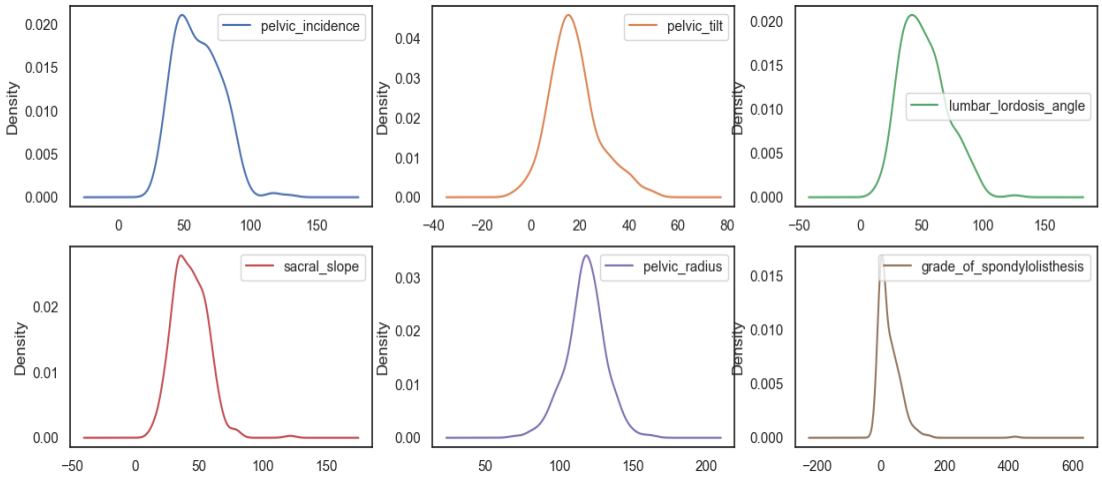
SECTION 3

3. Exploratory Data Analysis (EDA) and Data Preprocessing:

4.1. Univariate Analysis of '**Class'** distribution

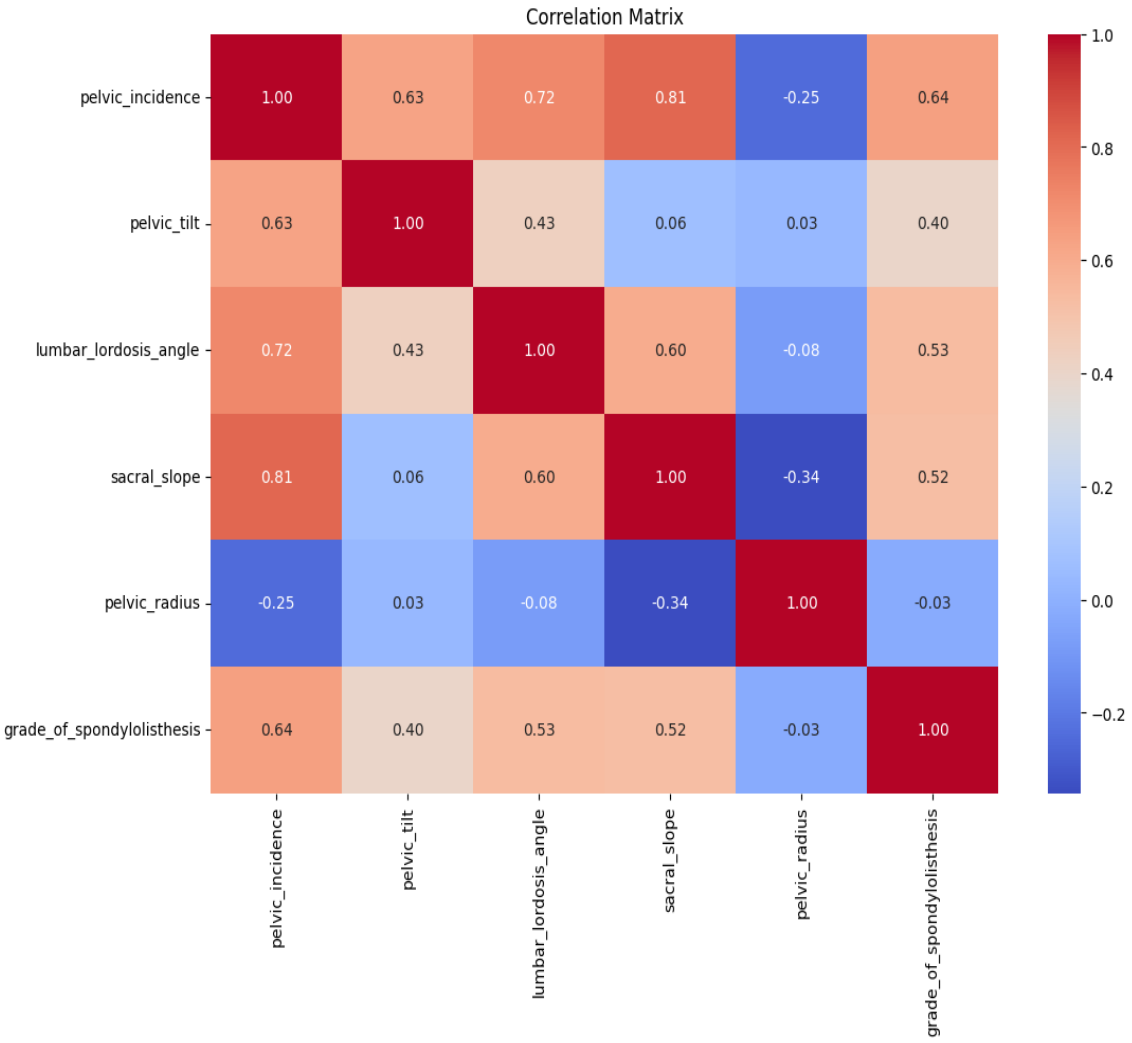


4.2. Univariate Analysis of all features



- All features are normally distributed except 'grade_of_spondylolisthesis which is right-skewed.

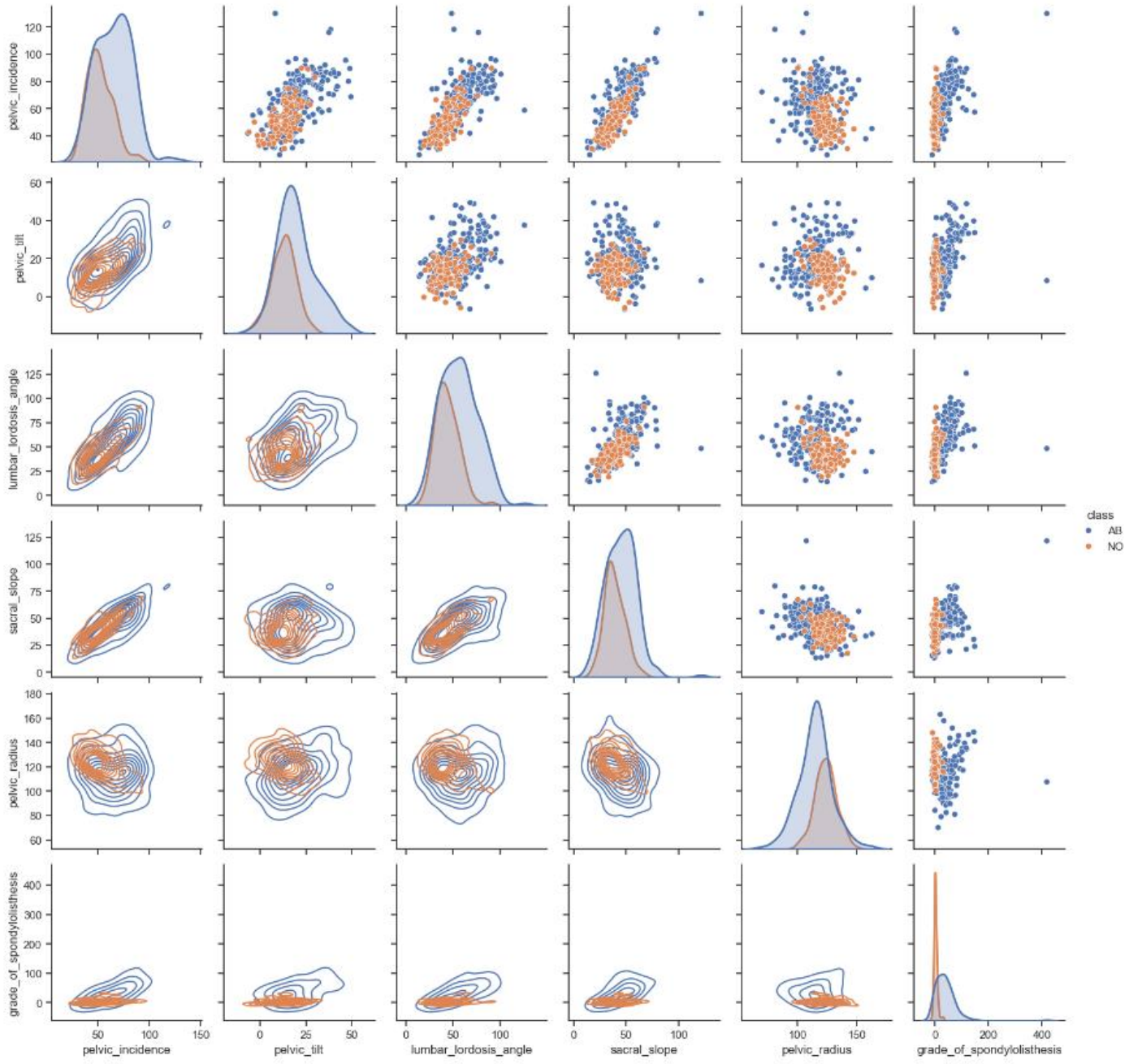
4.4. Multivariate Analysis: Correlation Matrix



- 'pelvic_radius' feature turns out to be less influential with poor correlations with other features.

- ‘pelvic_incidence’ , ‘grade_of_spondylolisthesis’ , lumbor_lordosis_angle are some of the features that are highly important in our problem.

4.3. Pair Plot



- Positive correlations : ‘pelvic_incidence and ‘pelvic_tilt’, ’lumber_lordosis_angle’, ’sacral_scope’, ‘pelvic_r.
- Negative correlations : ‘pelvic_radius’ and ‘pelvic_incidence’.
- ‘grade_of_spondylolisthesis’ is positively skewed
- Since the features are very less, we will train our data on all the features to yield best results.

4.5. Feature Scaling using StandardScaler():

Standard scaling or z-score normalisation technique is used in machine learning to standardise the dataset to ensure that all features are on the same scale and transform them to have '0' mean and '1' standard deviation. This is an important step for algorithms that are sensitive to the scale of the features.

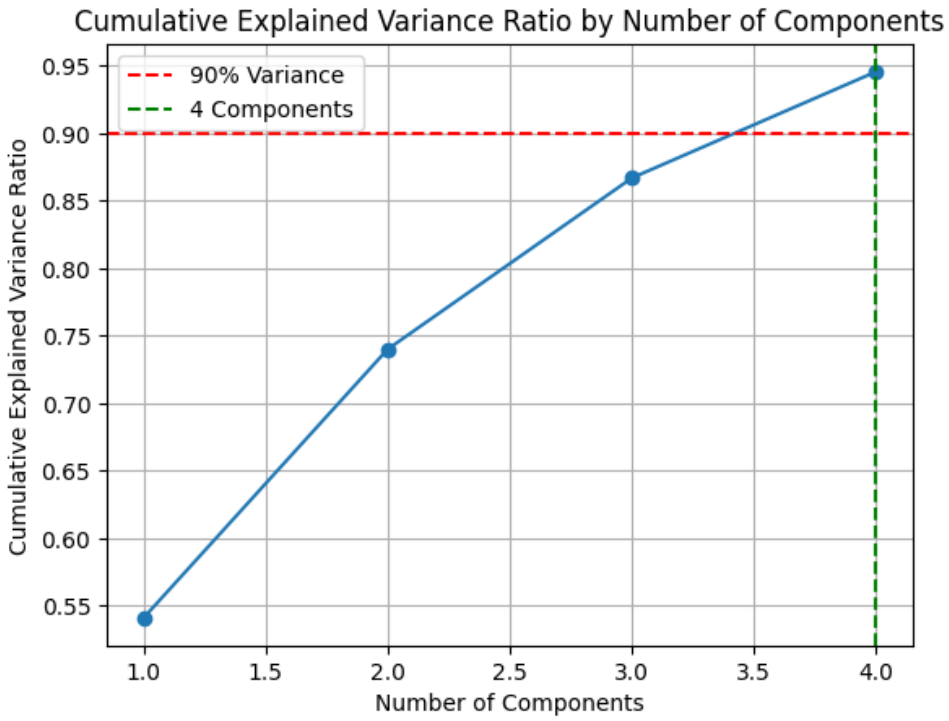
$$z_{ia} = \frac{x_{ia} - \langle x_a \rangle}{\sigma_a}, \quad i \in [n] \quad a \in [p],$$

Where, a = feature, n = data points, x_{ia} = single data point

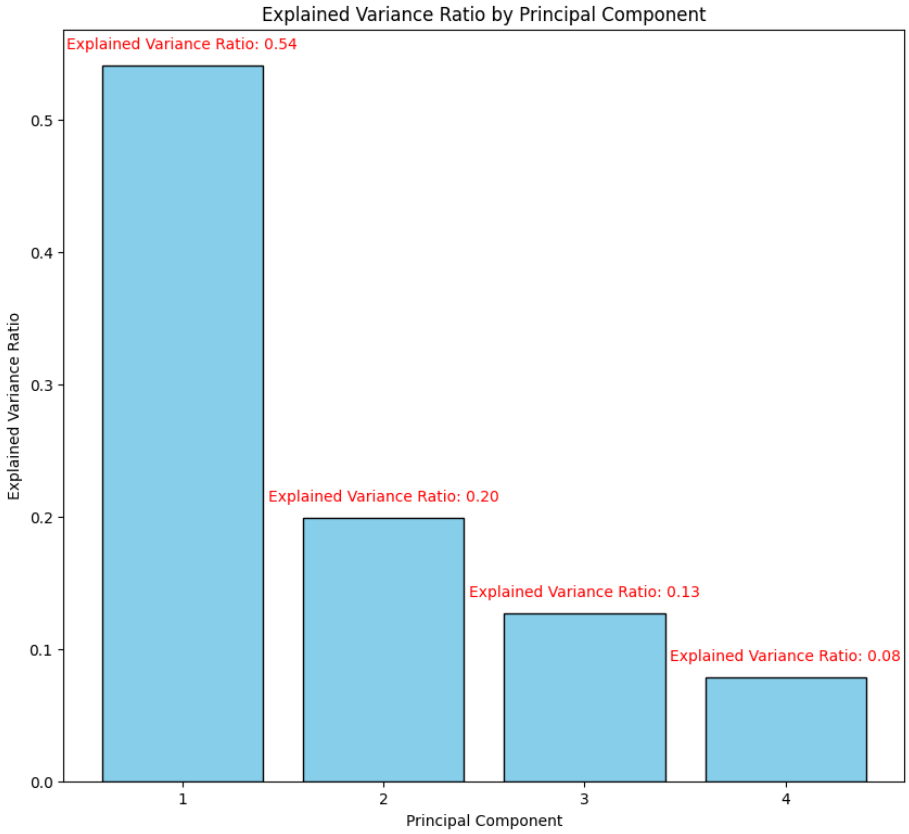
4.6. Dimensionality Reduction using Principal Component Analysis (PCA):

Dimensionality reduction techniques reduce the complexity of our model and avoid overfitting. We apply Principal component analysis to derive information from the existing features and extract lower dimensional feature subspace while maintaining most of the relevant information. Steps:

1. **Standardise** the data and construct covariance matrix.
2. **Instantiate 'PCA' object** from scikit-learn' s 'decomposition' module.
3. **'n_components'** sets threshold for the explained variance ratio, ensuring that transformed data retains atleast 90% of the original variance.
4. **'fir_transform()'** method computes the principal components and projects the data orthogonally.
5. **'reduced_unsup_data'** is our transformed dataset with 4 PCs.



- Fig. depicts the explained variance ratio cumulatively as we add components one-by-one. The red line cuts where **90% of variance** is explained and hence, we reduce the data to 4-dimensional dataset.



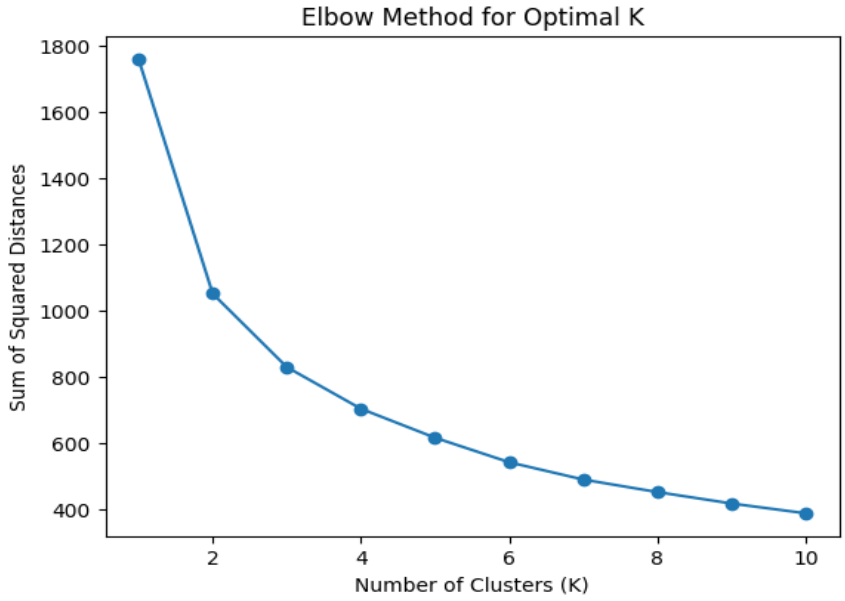
- As seen from fig, the first two principal components (PC1 and PC2) capture majority of the variability, while the remaining two components (PC3 and PC4) contribute less to the overall variance.
- This suggests that the data can be represented in lower dimensions **PC1 and PC2** for model training.

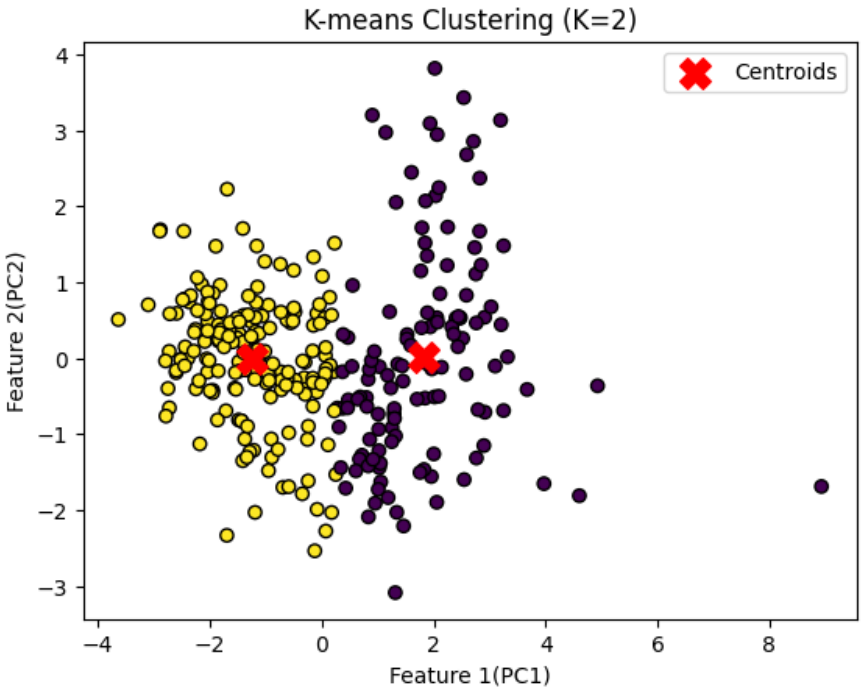
SECTION 4

5. Results and Discussion:

5.1. Unsupervised Classification Results:

Optimal 'K' determined by elbow method (fig) is K=2 as seen from the graph indicating distinct grouping. Visual examination (fig) of K means clustering on PC1 and PC2 revealed some degree of separation between cluster along with slight overlap.

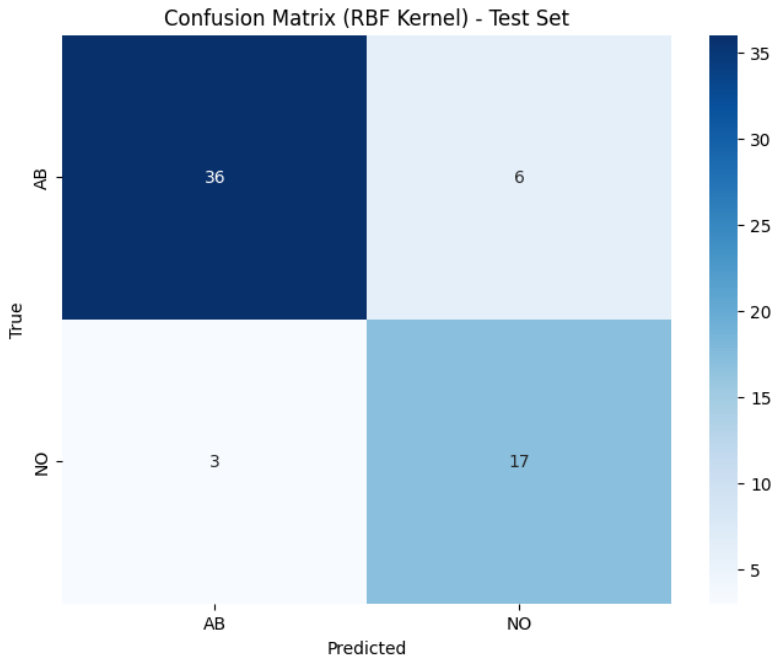




5.2. Supervised Classification Results:

The SVM classifier with the **RBF kernel** achieved the better accuracy when compared with linear kernel as RBF captures the non-linear overlapping decision boundary.

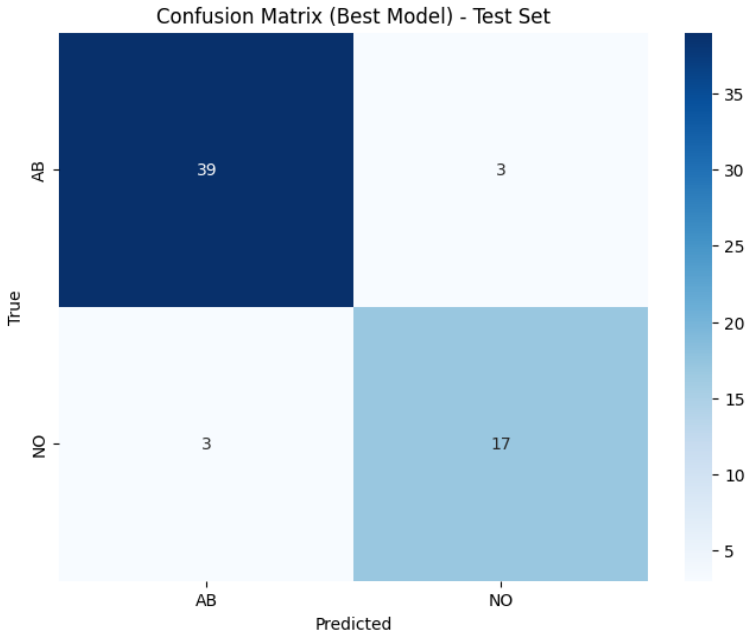
Performance measures	**SVM Training data(80%)**	**SVM Testing data(20%)**
Classification accuracy (%)	86.693	**85.483**



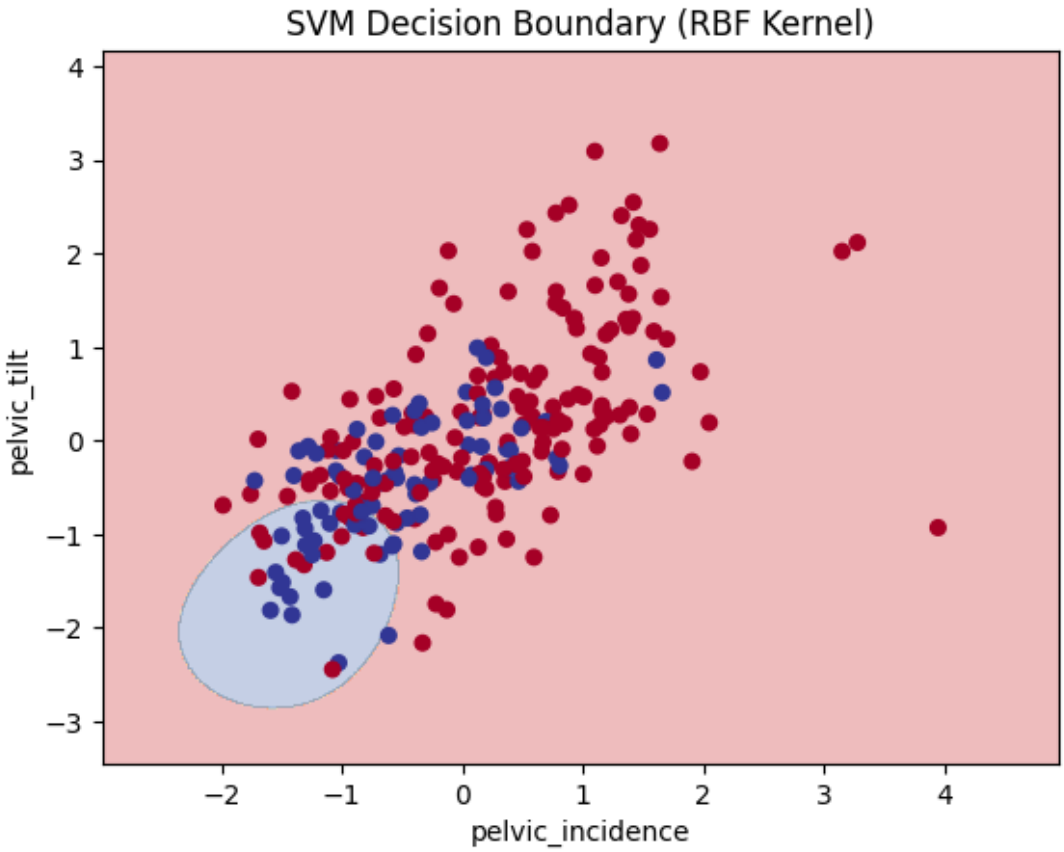
Hyperparameter Tuning and 10-fold cross validation (GridSearchCV()):

```
Best Parameters: {'C': 100, 'gamma': 0.01}
```

Performance measures	**SVM Training data(80%)**	**SVM Training data(20%)**	**SVM Testing data(10-fold cross validation)**
Classification accuracy (%)	86.693	85.483	**90.322**



Visualised **Decision boundary** of predicted class distribution using contour plots highlighting model's ability to separate the two classes:



5.3. Comparative Analysis:

The analysis of the vertebral dataset reveals distinct performances between Unsupervised (K-Means clustering) and Supervised (SVM with RBF kernel). Comparing the K-means classification with ground truth, which has **66%** accuracy, uncovers underlying data structure, and provides insights through PCA revealing clear separation of clusters. On the other hand, SVM with RBF kernel achieves a higher accuracy of **90%** which indicates that supervised technique provides accurate classification.

In conclusion, these two distinct methods complement each other's analysis of the vertebral dataset yielding profound insights into spinal health. Unlabelled K-means's data insights and pattern discovery along with labelled SVM's precise classification leads to a synchronized approach to provide comprehensive understanding of the vertebral dataset, ensuring robust decision-making in medical diagnostics. This holistic view helps us empower medical research and treatment planning.

6. References:

[1] Chabih, O., Sbai, S., Behja, H., Louhdi, M. R. C., & Trousse, B. (2021, June). 'New approach to determine the optimal number of clusters K in unsupervised classification', In *2020 6th IEEE Congress on Information Science and Technology (CiSt)* (pp. 348-352). IEEE.

[2] Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). 'Selection of K in K-means clustering', *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103-119.

[3] Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, *40*(1), 200-210.

[4] Pourebrahim, N. (2019). '*Human Dynamics in the Age of Big Data: A Theory-Data-Driven Approach*', The University of North Carolina at Greensboro.

[5] Liu, Y., & Parhi, K. K. (2016, October). 'Computing RBF kernel for SVM classification using stochastic logic', In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)* (pp. 327-332). IEEE.

SECTION 5

7. Appendix (Code):