COURSEWORK: Classification (10806957)

The vertebral column data consists of six biomechanical indicators which help identify whether an orthopaedic patient classifies as normal or abnormal. By observing their spine shape and orientation data, supervised and unsupervised machine learning methods can be applied to classify the orthopaedic patient's abnormality. Unsupervised classification refers to exploring data structures by classifying them into groups most identical to each other with no response variable (Berry et al., 2019), while supervised classification is labelling unseen data correctly through learning past labelled data (Berry et al., 2019). In this report, these algorithms were applied through Python.

1. **Unsupervised Clustering Method**

The unsupervised classification appropriate for this dataset is *model-based clustering,* particularly Gaussian Model Mixtures (GMM). Considered richer than other clustering methods like k-Means or kernel k-means, the aim of GMM is to probabilistically model the data under the assumption of stacked Gaussian distributions. Statistically, the probability of a datapoint being within a cluster can be shown as:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \; g(\mathbf{x}|\mu_i, \; \mathbf{\Sigma}_i),$$

where x is a D-dimensional continuous value of a vector, Wi is the mixture of weights, and $g(x|\mu i, \sum i)$ is the probability density of x under the i-th Gaussian distribution (Reynolds, 2009).

The probability-based aspect is suitable with the complex nature of the non-linear relationships in the vertebral column dataset, as features like pelvic tilt and lumbar angle can vary in individuals. Therefore, hard assignments through approaches like k-Means may risk misidentification/misdiagnosis through false negatives/positives. Several studies in medicine have preferred this algorithm over others due to this reason (see Gupta et al., 2019; Khanmohammadi & Chou, 2016; Riaz et al.., 2020). However, it can be tricky to determine Gaussian distribution assumptions as this may not always be right in medical data due to presence of outliers and noise (Khanmohammadi & Chou, 2016). Data preprocessing should thus take these crucial factors into account before modelling.

2. **Supervised classification method**

The classification that could be applied in this dataset is Support Vector Machines (SVM). As mentioned, the complex structure of the dataset implies that a more flexible algorithm is preferred – SVM's aim is to find the most optimised hyperplane which separates classes through margin maximisation and error minimisation instead of an unrealistically linear hyperplane. To optimally formulate this algorithm the following is applied:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, m \\ & \xi_i \geq 0, \;\; i = 1, \ldots, m. \end{aligned}$$

, where w is the weight vector, b is the bias, C is the regularisation parameter, ξi the slack variables which allow for noisy data, and $y(i)$ and $x(i)$ features of the train set (Cortes et al., 1995).

In medicine literature, SVM is commonly used and in fact said to perform best based on several studies (Sidney-Gibbons & Sidney-Gibbons., 2019; Kumari & Chitra, 2013; Wolberg & Mangasariant, 1994). However, considering that there is inherent bias in medical datasets - where abnormality is less common than normality in humans, SVMs can be biassed towards that majority thus data handling is crucial (Kumari & Chitra, 2013).

3. Exploratory data analysis and data preprocessing

Based on Figure 1, worth noting are the several outliers noticeable especially in the grade of spondylolisthesis. A split view of abnormal to normal classes in boxplots (Figure 2) noted high deviation of values from the average and median in the abnormal class of spondylolisthesis grade. These may be considerable factors in diagnosing abnormal vertebral condition. The kernel density estimators shown in Figure 3 agree with the distributions shown in histograms.

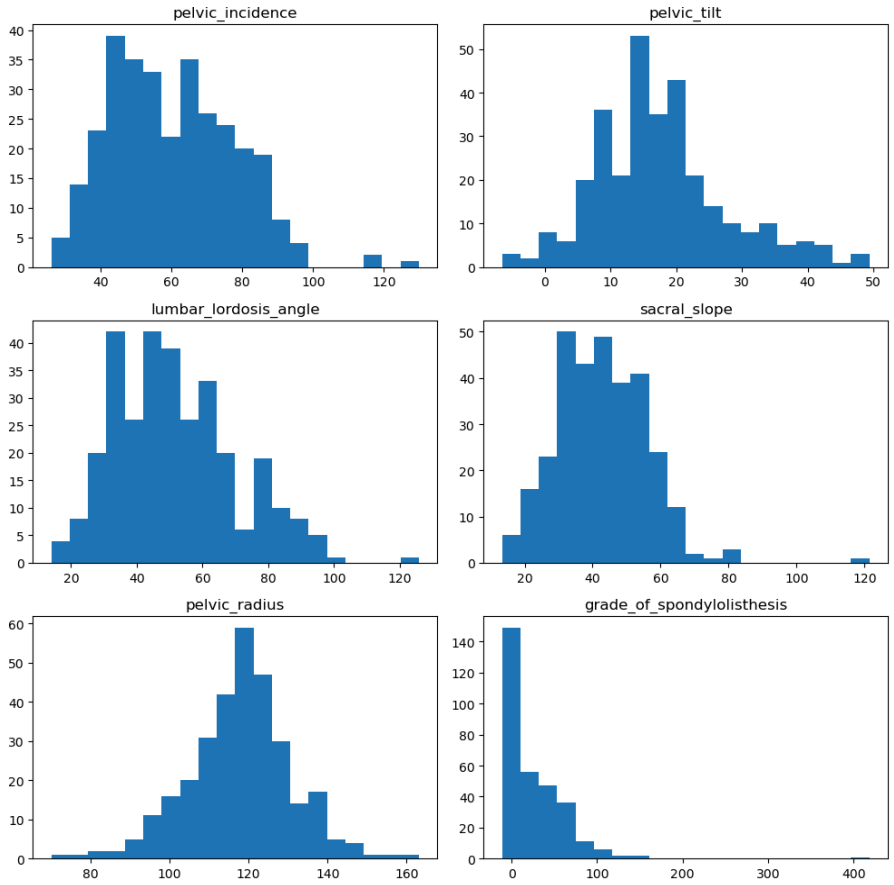

Figure 1: Histograms of features

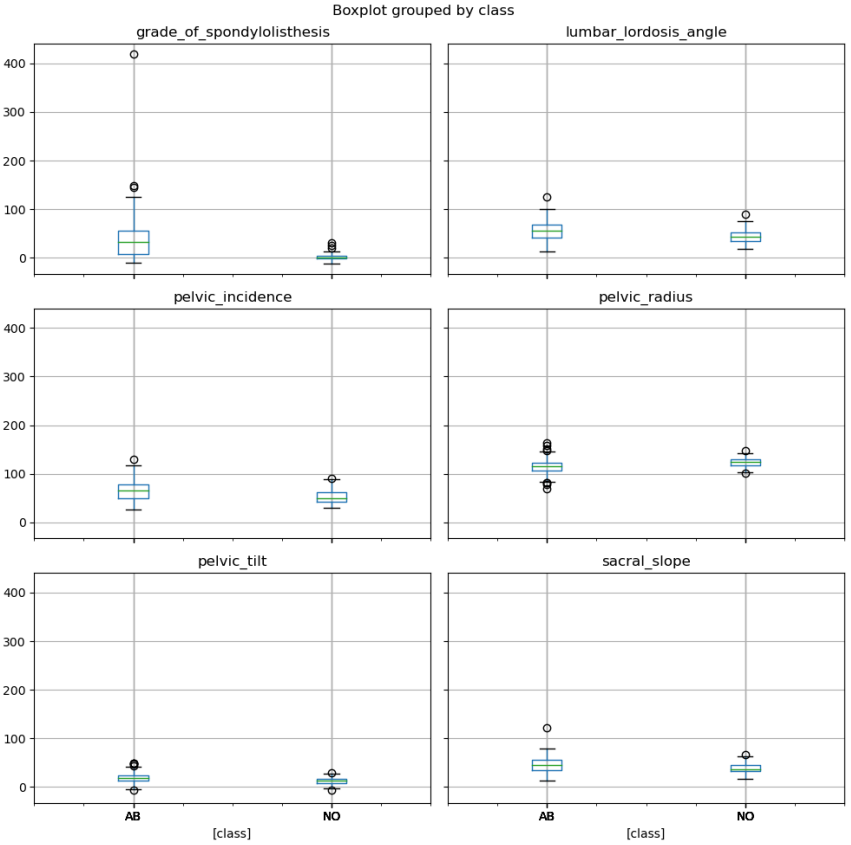

Figure 2: Boxplots of features by class

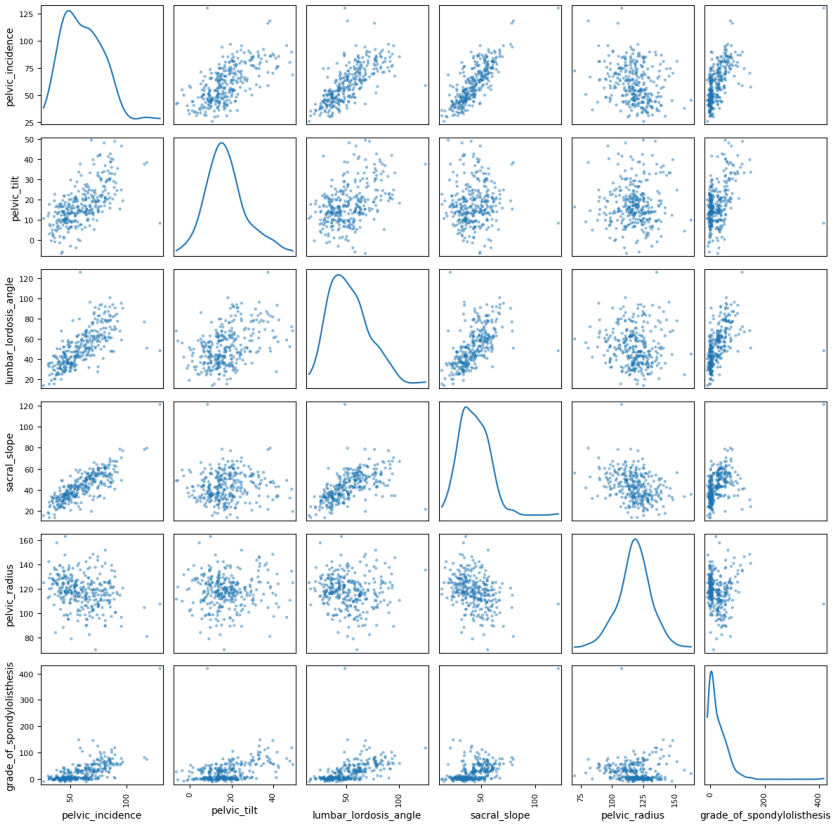

Figure 3: Scatterplot matrix of features

Data preprocessing

- *Outlier handling*: **'***RobustScaler'* was applied to the 'grade_of_spondylolithesis' feature to handle the significant outlier presence. This non-removal of outlier approach is chosen as they may provide important insight into the abnormality of orthopaedic patients.
- *Data normalisation:* For GMM clustering, mean and variance of the numerical features were standardised to 0 and 1 respectively using '*StandardScaler'*.
- *Dimensionality Reduction:* As the dataset is high-dimensional, a Principal Component Analysis was applied to the two most important components to optimise both GMM and SVM model performance.

4. Results and Discussion

- **Model-based clustering (GMM)**

  Gaussian distributions were checked on the features as a preliminary step. Although not precisely spread, this approach remains validly insightful as the distribution comes close, and the Central Limit Theorem agrees that the mean distribution will be normally distributed with sufficient data despite data showing the opposite.

  To determine the most optimal number of clusters, the AIC/BIC criterion were utilised and demonstrated two inherent groups within data. *k = 2* was decided based on this.

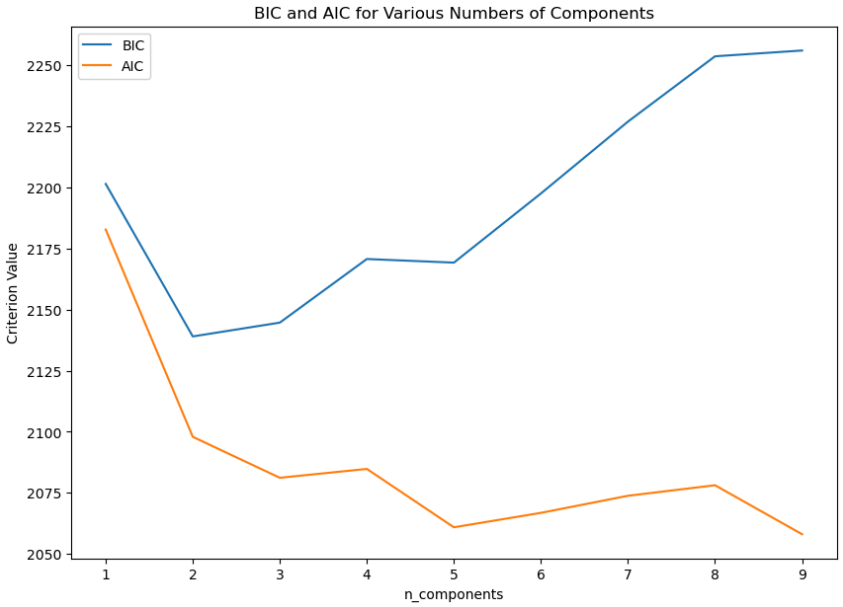

Figure 4: AIC/BIC for Components

The GMM clustering visualisation in Figure 5 indicated a degree of separation between classes in line with the abnormal/normal vertebral columns.

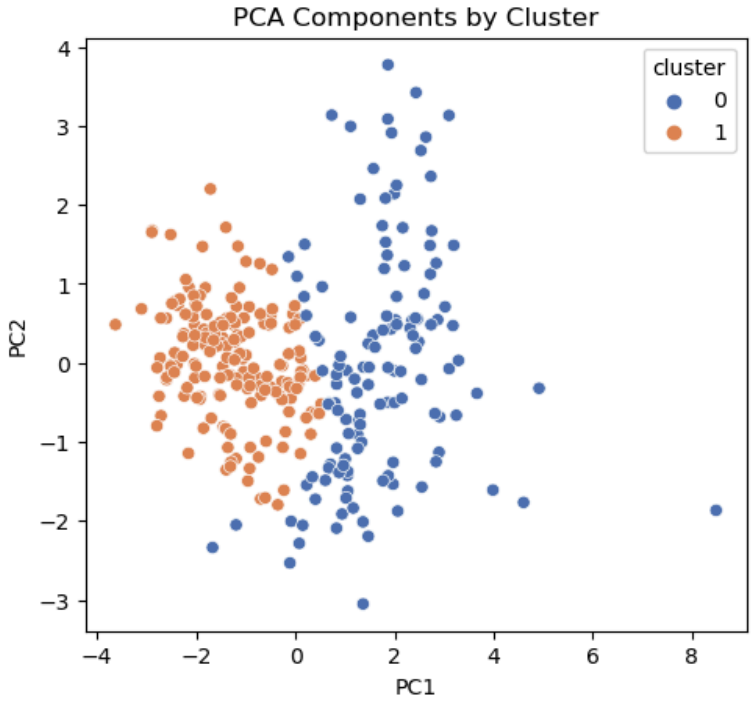

Figure 5: GMM on PCA-reduced data

- **Support Vector Machines**

Data was split into train and test sets before SVM model training. After training, predictions were made on the PCA-reduced test set. The SVM test set visualization also demonstrated some extent of separation between abnormal/normal as shown in Figure 6.

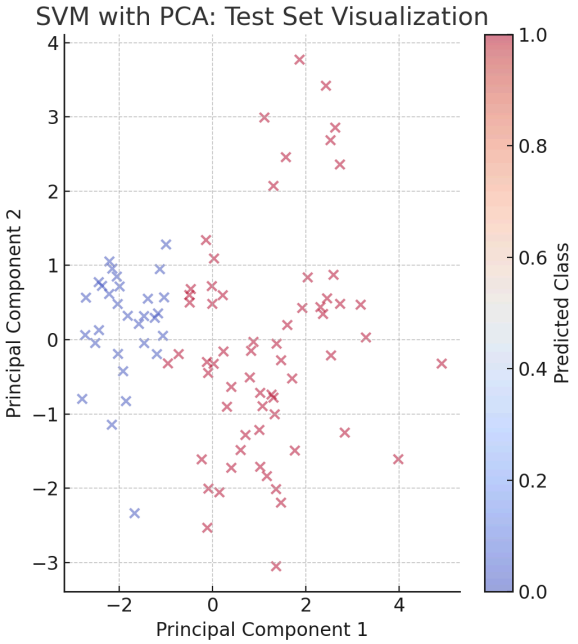

Figure 6: SVM on PCA-reduced data

To analyse the performances of both dimensionally reduced models, the accuracy scores and confusion matrices were obtained (Table 7). The accuracy scores which demonstrated how much the test set predictions matched the true labels indicated that SVM captured the underlying data distribution better than GMM.

Model	Accuracy scores	Confusion Matrix
GMM	67.7%	[[210 0] [100 0]]
SVM	72.0%	[[14 10] [16 53]]

Figure 7: Evaluation metrics for models

Based on the evaluation metrics, on top of having higher accuracy the SVM algorithm producing 53 true negatives and 14 true positives also seemed more sensible than the GMM which apparently demonstrated 0 values of false and true negatives. This gives more confidence in the supervised classification method than unsupervised learning, which is in line with current literature of similar contexts - as previously mentioned.

Despite the inexact match, mapping GMM clusters to the actual labels indicates a certain extent of alignment but also highlights the challenges in using unsupervised learning into classifications in medical data. Contrary to that, SVM performance showed robustness in differentiating between classes through direct evaluation on how supervised learning performance is true to the actual classifications - with its higher accuracy and minimal imbalances.

Recommendations for future application should explore other feature engineering in data preprocessing for a more cleaned dataset. Furthermore, utilising other kernels for SVM and cross-validation in models could also be employed for a more refined model. Other modelling techniques in both supervised and unsupervised learning could also be attempted to compare the most accurate approach.

In spite of some datapoint overlapping in its feature spaces as seen in both models' scatterplots, this would be sensible due to the complex and non-linear medical data. Regardless, both models successfully demonstrated the ability to classify patients based on the normality of biomechanical features even with reduced dimensionality. Using GMM for unsupervised clustering as well as SVM for classification captured the underlying patterns of orthopaedic patients which provided insights into the robustness of the model performance in the decision-making of diagnoses from orthopaedic patients' vertebral columns.

(1090 words)

References

Berry, M.W., Mohamed, A. and Yap, B.W. eds., 2019. Supervised and Unsupervised Learning for Data Science. *Springer Nature*.

Cortes, C. and Vapnik, V., 1995. Support-Vector Networks. *Machine Learning*, 20, pp.273-297.

Gupta, C., Gondhi, N.K. and Lehana, P.K., 2019. Analysis and Identification of Dermatological Diseases using Gaussian mixture Modeling. *IEEE Access*, 7, pp.99407-99427.

Khanmohammadi, S. and Chou, C.A., 2016. A Gaussian mixture model based discretization algorithm for associative classification of medical data. *Expert Systems with Applications*, 58, pp.119-129.

Kumari, V.A. and Chitra, R., 2013. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2), pp.1797-1801.

Reynolds, D.A., 2009. Gaussian Mixture Models. *Encyclopedia of Biometrics*, 741(659-663).

Sidey-Gibbons, J.A. and Sidey-Gibbons, C.J., 2019. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19, pp.1-18.