Is It Possible to Predict the Diagnosis, Benign or Malignant, of Breast Cancer Based on Cellular Features?

Kayleigh Haydock

1 Executive Summary

Breast cancer is one of the most common cancers worldwide, and early detection can significantly improve treatment outcomes [1]. This project investigates the ability to predict the diagnosis of breast cancer, specifically distinguishing between benign and malignant tumours, using cellular features from breast cancer biopsies. The dataset used is the Breast Cancer Wisconsin (Diagnostic) dataset [2], which contains 30 features describing the characteristics of cell nuclei present in biopsy samples. The central research question addressed is: *Can we predict the diagnosis (benign or malignant) of breast cancer, based on cellular features?*

2 Dataset Overview

The widely used dataset comprises 569 samples describing various measurements like radius, texture and smoothness of tumour cells. The target variable indicates whether a tumour is benign (B) or malignant (M).

2.1 Machine Learning Techniques

Three algorithms were applied for classification and clustering:

- **Logistic Regression:** A linear model used for binary classification, which estimates the probability of an instance belonging to a class B or M.
- **Support Vector Machines (SVM):** Evaluated with both linear and RBF kernels to handle high-dimensional data and capture non-linear patterns.
- **K-Means Clustering:** An unsupervised learning technique used to group similar data points. The optimal number of clusters *(k=2)* was selected based on validation through both the elbow method and silhouette score analysis.

3 Evaluation Metrics

The models were evaluated using the following: **Accuracy**: Percentage of correctly classified instances. **Classification Report**: Provides precision, recall, and F1-score for both classes. **GridSearchCV**: Selects the hyperparameters that yield the best results on the validation data. **Confusion Matrix**: Visualises the true positives, true negatives, false positives, and false negatives. **K-Fold Cross-Validation**: 5-fold cross-validation ensures the model's generalisability. **Elbow Method**: Used to determine the optimal number of clusters in K-Means. **Silhouette Score**: To further validate that two clusters was the most suitable number.

3.1 Results and Discussion

The models achieved high classification performance, demonstrating their effectiveness for predicting breast cancer diagnosis. A summary for some of the key metrics is provided in Table 1.

Table 1: Model Performance Summary

Model	Accuracy	Precision (Malignant)	Recall (Malignant)
Logistic Regression	93%	0.97	0.84
SVM (Linear Kernel)	93%	0.97	0.84
SVM (RBF Kernel)	93%	1.00	0.81

Logistic Regression: The model provided a clear and interpretable baseline, identifying key features such as concavity and radius as influential. It achieved an overall accuracy of 93%.

Support Vector Machines (SVM): The linear kernel performed similarly to logistic regression, with 93% accuracy. The RBF kernel also achieved the same accuracy but required tuning and had a slight trade-off in malignant case recall, making the linear kernel the better choice for this research question.

K-Means Clustering: Unsupervised clustering identified two natural groups, aligning with the benign and malignant labels. Validation through the elbow method and silhouette scores confirmed the choice of $K = 2$, demonstrating clustering's utility for pattern discovery.

3.2 Strengths and Limitations

Strengths:

- Logistic regression was chosen due to its simplicity and interpretability, making it suitable for medical applications [3].
- SVM is effective for high-dimensional datasets and non-linear decision boundaries [4].
- K-Means clustering was also chosen to help uncover potential hidden patterns in the data, as seen in Figure 1. and provides a holistic approach to the research question.

Limitations:

- The dataset may suffer from class imbalance, potentially affecting model accuracy.
- Feature selection might not fully capture the complexity of the data.
- SVM's model's complexity requires careful tuning to avoid over fitting.
- K-Means does not deal well with noise and can be sensitive to outliers in the data, so there could potentially be bias.

4 Rationale for Method Selection

These algorithms were chosen for their complementary strengths. Logistic regression is straightforward and interpretable with its binary nature. SVM is well-suited for high-dimensional and complex data, offering robust performance; however, its complexity might be considered overkill for simpler datasets like this one, where more interpretable methods may suffice. K-Means clustering adds an unsupervised perspective, helping to validate natural groupings within the data. Together, these methods offer a well-rounded approach to classification and exploratory analysis in a medical setting. Overall, the models performed well, but further exploration of feature engineering and addressing data imbalance could enhance performance.

5 Conclusion

This project demonstrates that machine learning, particularly logistic regression and SVM, can effectively classify breast cancer diagnoses based on cellular features. The models identified key features such as radius and concavity, which strongly influence tumour malignancy. While the models achieved high accuracy, limitations such as data imbalance and sensitivity to outliers should be addressed in future work. Overall, this project highlights the potential of machine learning techniques to improve the efficiency and accuracy of breast cancer diagnosis.

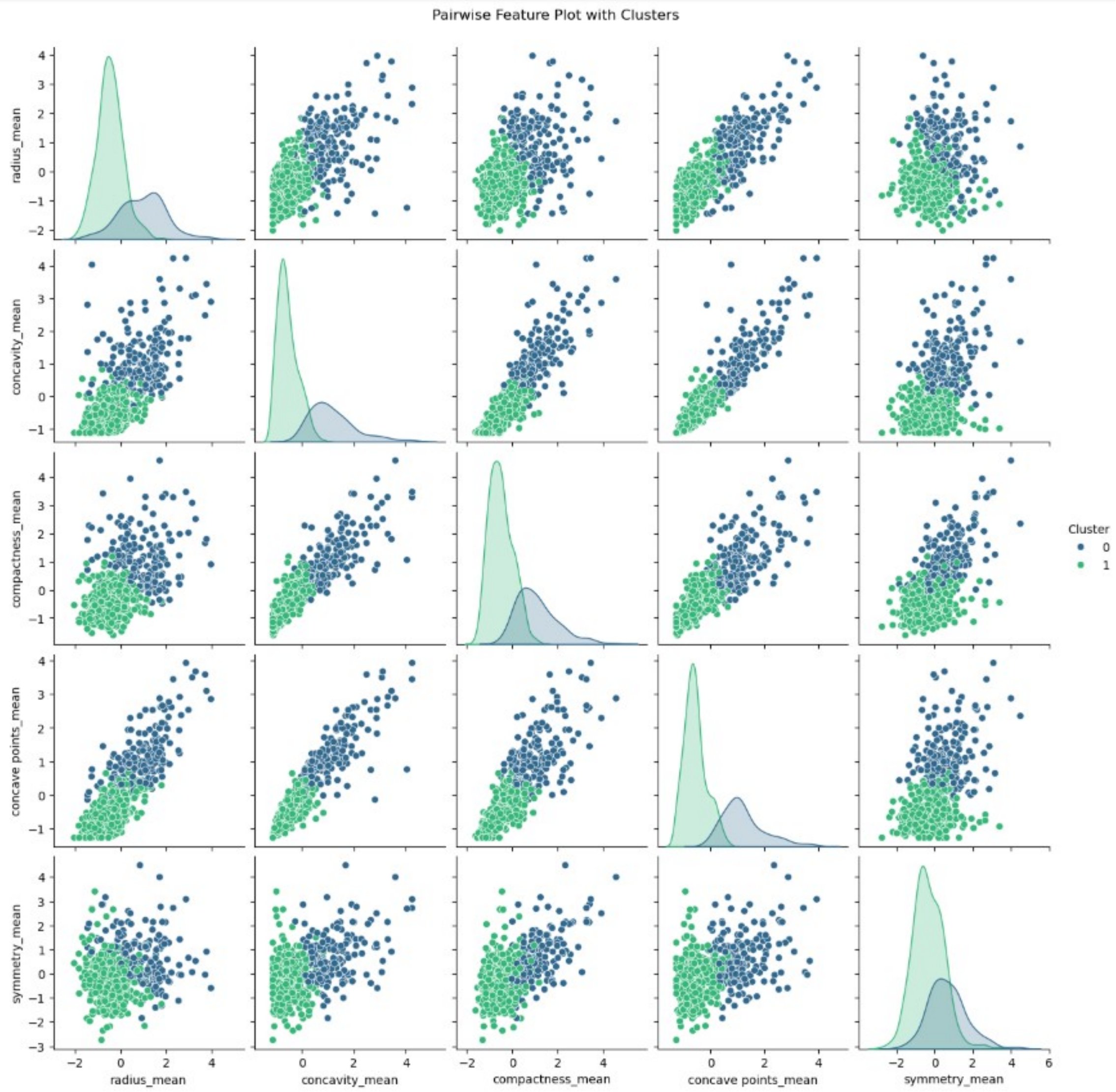

Figure 1: The figure visualises the relationships between all pairs of features in the dataset and uses colour to indicate their cluster.

References

[1] Annisa Maulidia, Lita Lidyawati, Lucia Jambola, and Lisa Kristiana. Analysis of logistic regression algorithm for predicting types of breast cancer based on machine learning. *AIP Conference Proceedings*, 2772(2):040005–1, 040005–8, 2023.

[2] UCI Machine Learning Repository. Breast cancer wisconsin (diagnostic) data set, 2024. Accessed: 2024-10-20.

[3] John D. Kelleher, Brian. Mac Namee, and Aoife D'Arcy. *Fundamentals of machine learning for predictive data analytics algorithms, worked examples, and case studies*. The MIT P., Cambridge, Mass, 2nd ed. edition, 2020. The Art of Machine Learning for Predictive Data Analytics.

[4] Ziba Khandezamin, Marjan Naderan, and Mohammad Javad Rashti. Detection and classification of breast cancer using logistic regression feature selection and gmdh classifier. *Journal of Biomedical Informatics*, 111:103591, 2020.