Predicting Breast Cancer Diagnosis using Machine Learning

A DATA-DRIVEN APPROACH TO CLASSIFYING BENIGN VS. MALIGNANT TUMOURS

KAYLEIGH HAYDOCK

Research Question

Can we predict the diagnosis (benign or malignant) of breast cancer based on cellular features such as radius, texture, perimeter, area, and others?

Why should it matter?

Early detection of breast cancer can save lives.

Accurate classification models can help doctors make quicker, more reliable decisions.

Dataset Overview

Title: Dataset: Breast Cancer Wisconsin (Diagnostic) Dataset

Source: Available on Kaggle or UCI Machine Learning Repository

Data Characteristics:

Number of Instances: 569

Number of Features: 30 (including radius, texture, perimeter, area, etc.)

Target Variable: Diagnosis (Benign=0, Malignant=1)

Machine Learning Approaches

Supervised Learning: Classification

Techniques:

- **Logistic Regression:** Linear model, interpretable, identifies feature importance via coefficients
- **Support Vector Machines (SVM):**
  - **Linear SVM:** Creates linear decision boundaries
  - **RBF Kernel SVM:** Handles non-linear data via higher-dimensional mapping
  - **GridSearchCV:** Fine tuning hyperparameters
- **K-Fold Cross-Validation:** Assesses model accuracy and generalisability

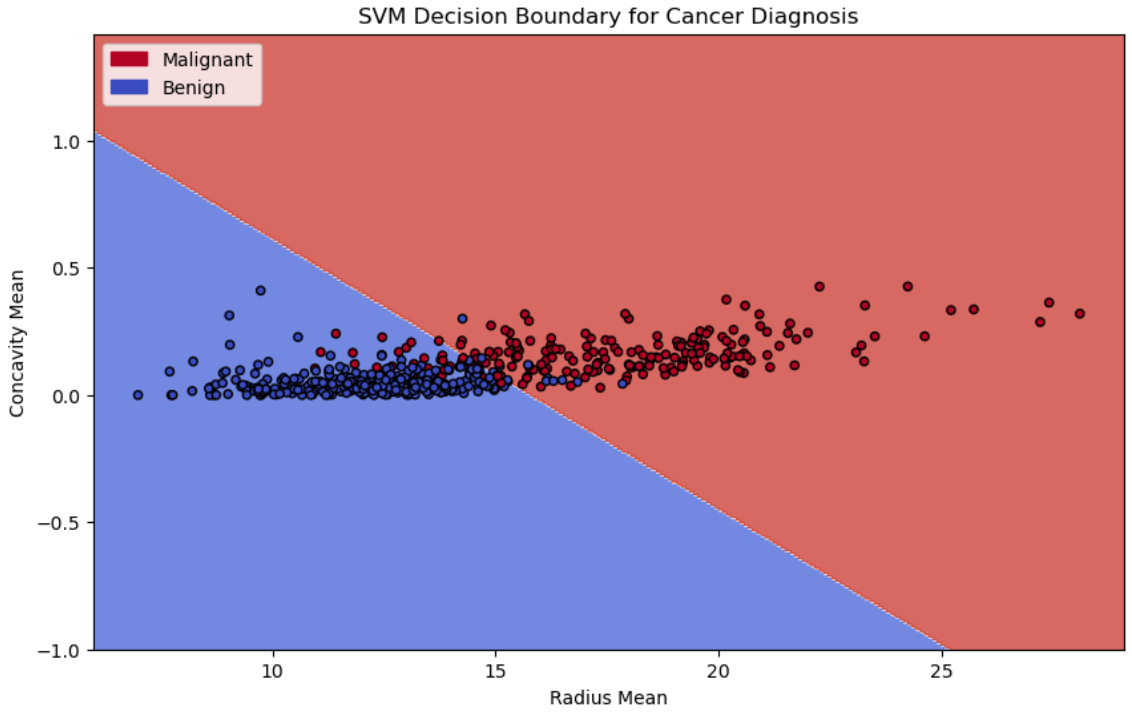
Evaluation:

- Metrics: Accuracy, Precision, Recall, F1 Score
- **Confusion Matrix:** Analyses true/false positives and negatives

Unsupervised Learning

Techniques:

- **K-Means Clustering:** Identifies hidden patterns in data
- **Elbow Method:** Determines optimal cluster count (*K*)
- **Silhouette Score:** Evaluates cluster quality and data point assignments

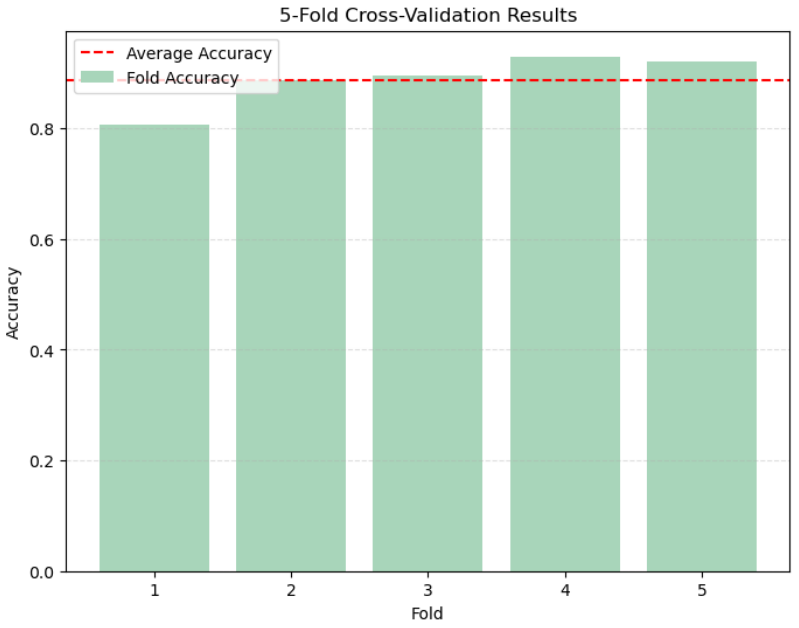


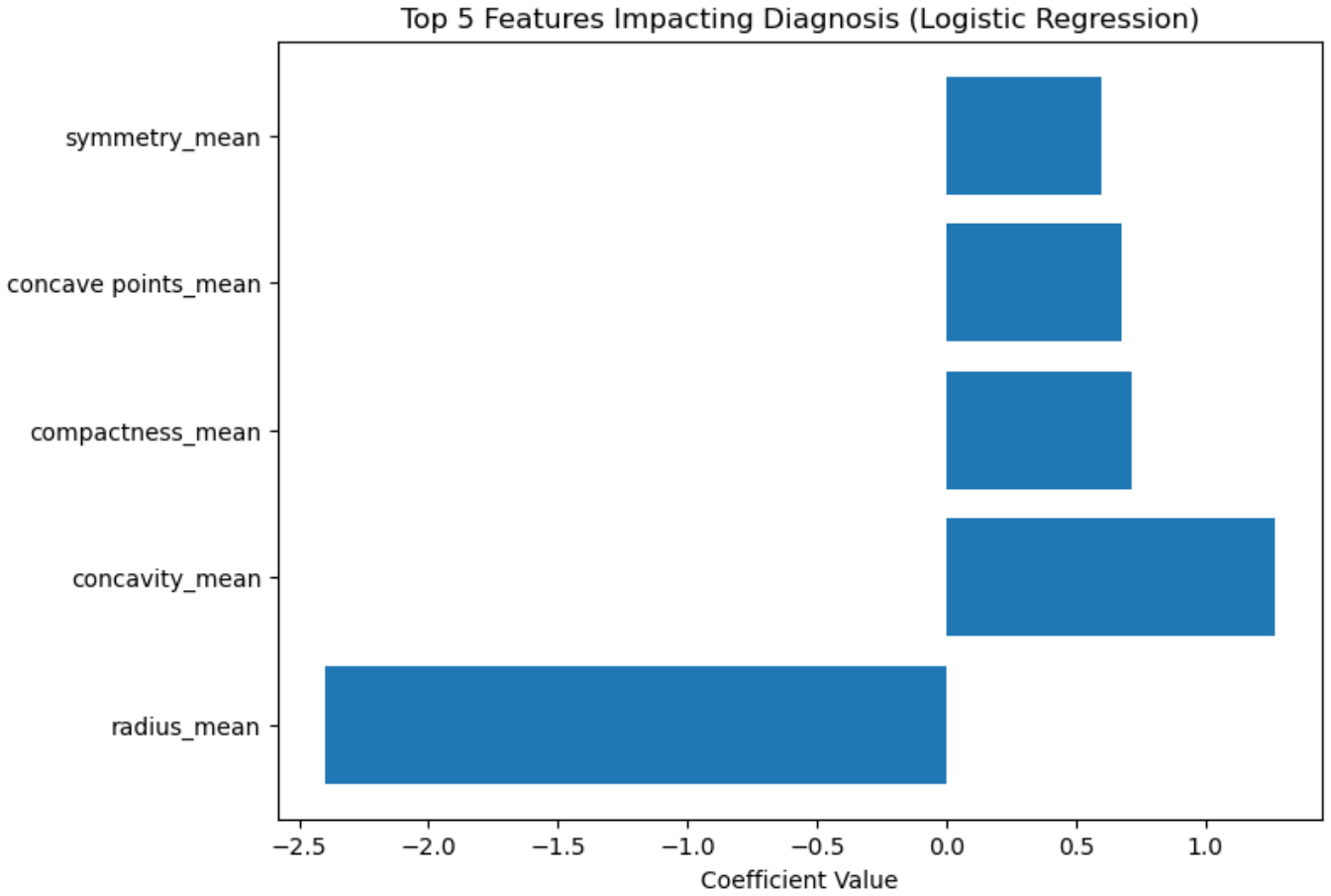
Data Analysis and Results

Logistic Regression:
- Top 5 most impactful features identified: Radius, Texture, Perimeter, Area, Smoothness
- Model accuracy on test data: 0.9298245614035088

SVM Performance:
- Linear and RBF SVM accuracy: 0.928245614035088
- 5-Fold Cross-Validation: [0.80701754, 0.88596491, 0.89473684, 0.92982456, 0.92035398]
- Average Cross-Validation Accuracy: 0.8875795683900014

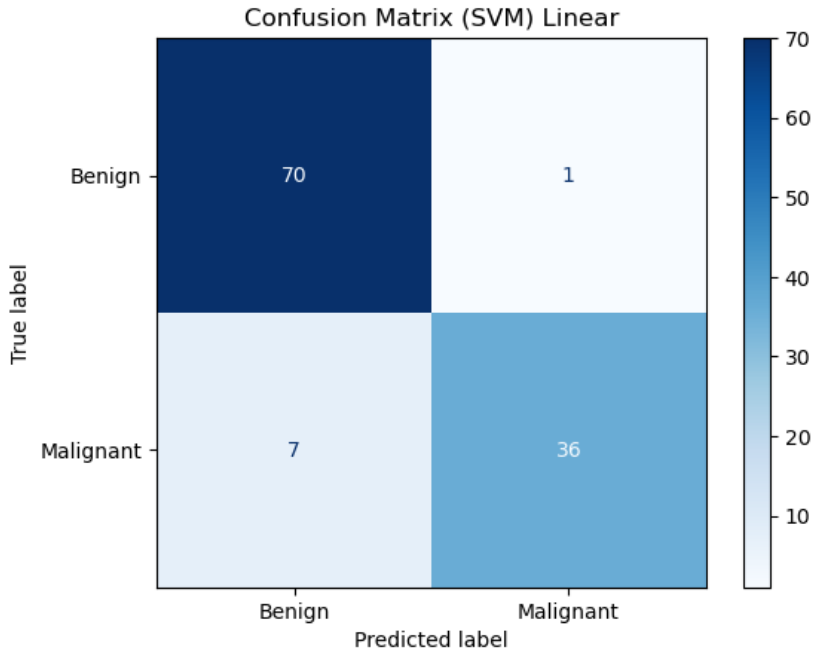
Clustering:
- K-Means: *K*=2
- Elbow Method
- Silhouette Score




Top 5 Features Impacting Diagnosis (Logistic Regression)
symmetry_mean
concave points_mean
compactness_mean
concavity_mean
radius_mean
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
Coefficient Value

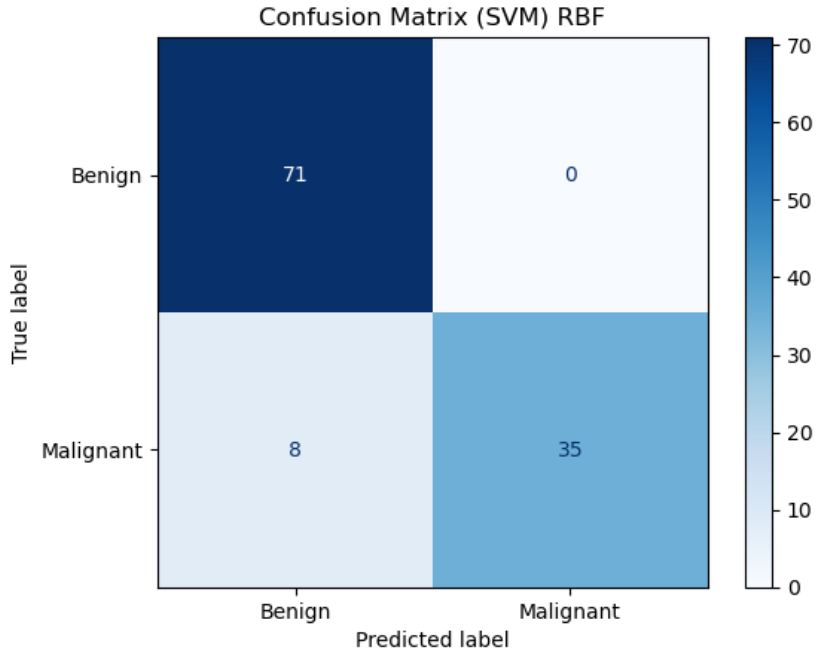
```python
# Classification report
print("Classification Report (SVM) Linear:")
print(classification_report(y_test, y_pred_svm))

# Confusion Matrix
conf_matrix_svm = confusion_matrix(y_test, y_pred_svm)
ConfusionMatrixDisplay(conf_matrix_svm, display_labels=["Benign", "Malignant"]).plot(cmap="Blues")
plt.title("Confusion Matrix (SVM) Linear")
plt.show()
```

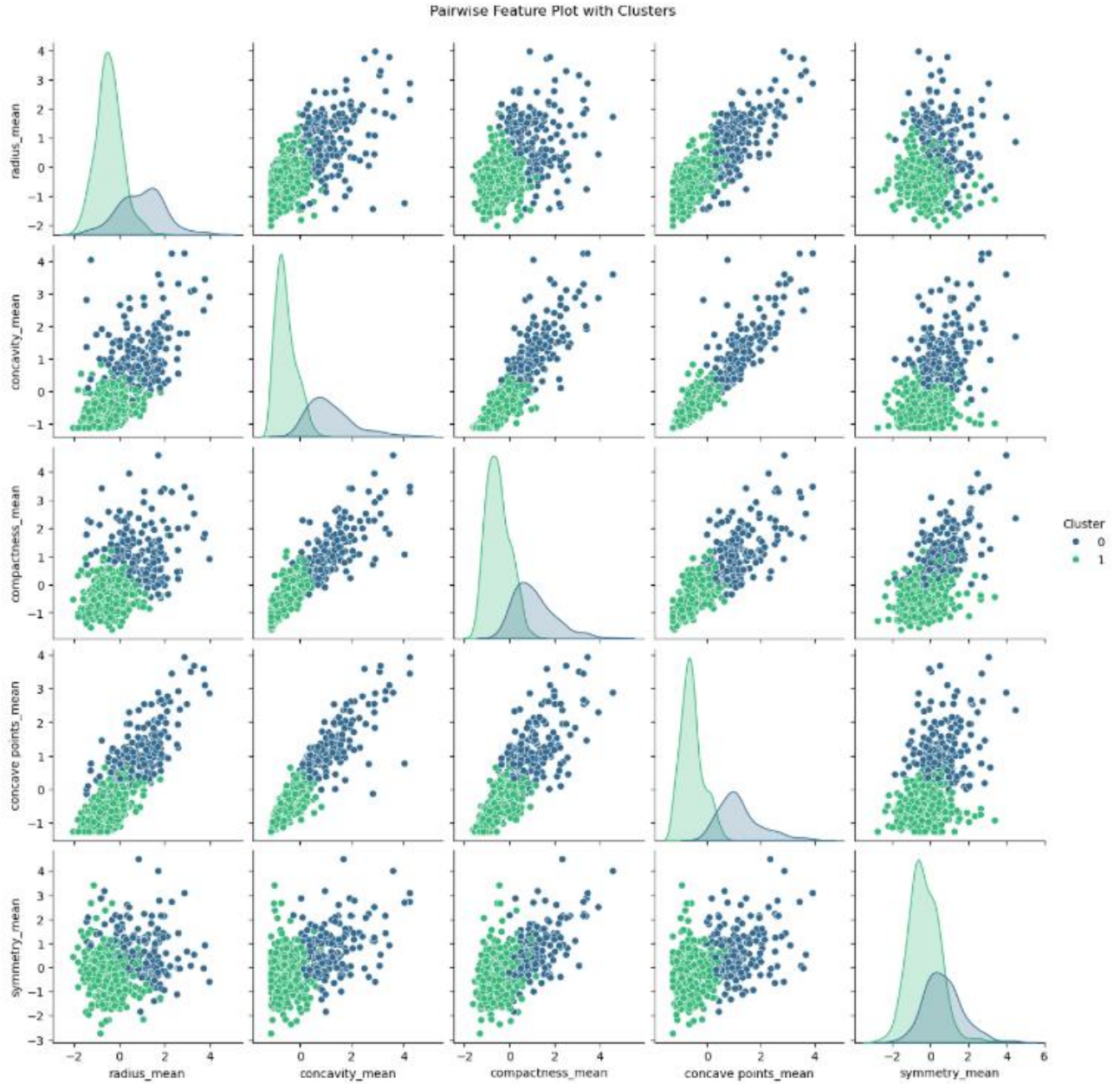
```
Accuracy of SVM Model with a linear kernel: 0.9298245614035088
Classification Report (SVM) Linear:
              precision    recall  f1-score   support

           0       0.91      0.99      0.95        71
           1       0.97      0.84      0.90        43

    accuracy                           0.93       114
   macro avg       0.94      0.91      0.92       114
weighted avg       0.93      0.93      0.93       114
```




```python
# Classification report
print("Classification Report (SVM) RBF:")
print(classification_report(y_test, y_pred_rbf))

# Confusion Matrix
conf_matrix_rbf = confusion_matrix(y_test, y_pred_rbf)
ConfusionMatrixDisplay(conf_matrix_rbf, display_labels=["Benign", "Malignant"]).plot(cmap="Blues")
plt.title("Confusion Matrix (SVM) RBF")
plt.show()
```

```
Accuracy of SVM Model with RBF kernel: 0.9298245614035088
Classification Report (SVM) RBF:
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        71
           1       1.00      0.81      0.90        43

    accuracy                           0.93       114
   macro avg       0.95      0.91      0.92       114
weighted avg       0.94      0.93      0.93       114
```




Pairwise Feature Plot with Clusters
radius_mean
concavity_mean
compactness_mean
concave points_mean
symmetry_mean
Cluster
0
1

What I've Learned

Machine Learning Insights:
- Feature selection significantly impacts model performance (Logistic Regression)
- SVMs can capture complex patterns in the data
- Clustering can highlight deeper insights to data that provide a well-rounded view
- Establishing measurable performance indicators is vital for assessing and improving models effectively

Real-World Implications:
- Early breast cancer diagnosis using machine learning could be a valuable tool in medical decision-making
- The importance of validating models with cross-validation and other techniques to ensure robustness

Key Takeaway:
- The choice of machine learning methods depends on the specific research question or task, along with the trade-offs inherent to each approach

Limitations

Data Limitations:
- Small dataset size (~569 instances), which might not generalize well to larger populations
- Potential bias in the dataset (e.g., non-representative sample)

Model Limitations:
- SVM performance might degrade if the data is noisy or too high-dimensional
- Logistic Regression assumes a linear relationship, which might not capture non-linear patterns

Generalization:
- Models could overfit if not properly tuned or validated on larger, more diverse datasets

Conclusion

Summary:

- The research question was addressed by applying both Logistic Regression, Support Vector Machines and Clustering to predict breast cancer diagnosis
- The results showed good model accuracy and demonstrated the importance of choosing the right features

Future Work:

- Consider more advanced models (e.g., Neural Networks)
- Use larger datasets to improve model generalization
- Experiment with more hyperparameter tuning (RandomizedSearchCV)
- Consider using density based clustering such as DBSCAN

Model development is not just a science—it's an art, requiring careful consideration of goals and context.

Appendix

Kaggle Dataset: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

GitHub: https://github.com/KaysHaydock/machine_learning_breast_cancer