Cancer Prediction Using Machine Learning

Introduction

Cancer prediction and diagnosis are critical areas of medical research, with significant implications for patient outcomes and treatment strategies. Machine learning, a subset of artificial intelligence, has shown immense potential in enhancing cancer prediction by analyzing complex datasets and identifying patterns that may not be immediately apparent through traditional statistical methods. This project focuses on using a regression model to predict cancer diagnoses based on a comprehensive dataset of breast cancer cases.

Data Set Description

The dataset used in this project consists of 569 instances and 33 attributes, detailing various characteristics of breast cancer cell nuclei present in the digitized images of fine needle aspirate (FNA) of breast masses. The data has the following structure:

Data Shape

- **Rows:** 569
- **Columns:** 33

Data Columns

The dataset includes the following columns:

1. **id:** Unique identifier for each observation
2. **diagnosis:** Diagnosis of breast cancer (B = benign, M = malignant)
3. **radius_mean:** Mean of distances from center to points on the perimeter
4. **texture_mean:** Standard deviation of gray-scale values
5. **perimeter_mean:** Mean size of the core tumor area
6. **area_mean:** Mean area of the tumor

7. **smoothness_mean:** Mean of local variation in radius lengths
8. **compactness_mean:** Mean of compactness (perimeter^2 / area - 1.0)
9. **concavity_mean:** Mean of concavity (severity of concave portions of the contour)
10. **concave points_mean:** Mean of number of concave portions of the contour
11. **symmetry_mean:** Mean symmetry of the tumor
12. **fractal_dimension_mean:** Mean of fractal dimension ("coastline approximation" - 1)
13. **radius_se:** Standard error for the radius
14. **texture_se:** Standard error for the texture
15. **perimeter_se:** Standard error for the perimeter
16. **area_se:** Standard error for the area
17. **smoothness_se:** Standard error for the smoothness
18. **compactness_se:** Standard error for the compactness
19. **concavity_se:** Standard error for the concavity
20. **concave points_se:** Standard error for the number of concave portions
21. **symmetry_se:** Standard error for the symmetry
22. **fractal_dimension_se:** Standard error for the fractal dimension
23. **radius_worst:** "Worst" or largest mean value for the radius
24. **texture_worst:** "Worst" or largest mean value for the texture
25. **perimeter_worst:** "Worst" or largest mean value for the perimeter
26. **area_worst:** "Worst" or largest mean value for the area

27. **smoothness_worst:** "Worst" or largest mean value for the smoothness

28. **compactness_worst:** "Worst" or largest mean value for the compactness

29. **concavity_worst:** "Worst" or largest mean value for the concavity

30. **concave points_worst:** "Worst" or largest mean value for the number of concave portions

31. **symmetry_worst:** "Worst" or largest mean value for the symmetry

32. **fractal_dimension_worst:** "Worst" or largest mean value for the fractal dimension

33. **Unnamed: 32:** Unnamed column which may be discarded or further investigated

Model Used: Regression Model

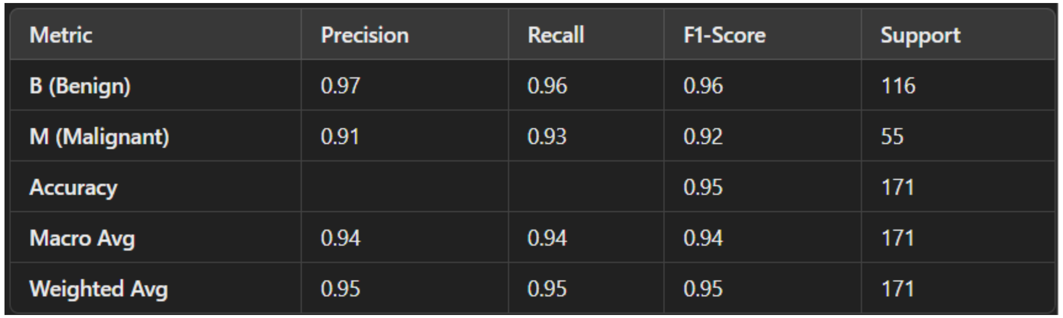
In this project, a regression model was utilized to predict the likelihood of a breast cancer diagnosis being benign (B) or malignant (M). Regression models are powerful tools in predicting continuous outcomes and can be adapted to classification problems with appropriate adjustments and decision boundaries.

Conclusion

The performance of the regression model was evaluated using several key metrics, with the following results:

- **Accuracy:** 0.94736

The detailed classification report is as follows:



Metric	Precision	Recall	F1-Score	Support
B (Benign)	0.97	0.96	0.96	116
M (Malignant)	0.91	0.93	0.92	55
Accuracy			0.95	171
Macro Avg	0.94	0.94	0.94	171
Weighted Avg	0.95	0.95	0.95	171

These results indicate a high level of accuracy and robustness of the regression model in predicting breast cancer diagnoses, with precision, recall, and F1-scores all demonstrating the model's effectiveness across both benign and malignant cases. This underscores the potential of machine learning models in aiding medical professionals to make more accurate and timely diagnoses, ultimately contributing to better patient outcomes.