Exploration of feature significance in Breast cancer diagnosis

Palak Jagani_(0826226)

2024-03-16

Background

Breast cancer stands as a significant global health challenge, affecting millions of women each year. Recognized by the World Health Organization (WHO) as the most prevalent cancer among women, it impacts approximately 2.1 million women annually and is the leading cause of cancer-related deaths in the female. These statistics underscore the critical importance of early detection and precise diagnosis in enhancing patient survival rates and overall quality of life. The analysis presented in this report leverages the Breast Cancer Wisconsin dataset to identify patterns that could significantly aid in the early detection and diagnosis of breast cancer, addressing this urgent health concern. Research into breast tumor characteristics, including radius, area, compactness, smoothness, and symmetry, has laid the groundwork for investigating differences between malignant and benign tumors. This body of work forms the basis for our hypothesis that discernible variations in the mean values of these features exist between the two tumor types.

Mehak et al. explored the diagnostic implications of quantitative image features extracted from breast biopsy images. Their research highlighted the critical role of area, smoothness, and compactness in differentiating between cancerous and non-cancerous tissues, stating that quantification of tumor image could substantially improve diagnostics [1]. Rasool et al. developed predictive models for breast cancer diagnosis using exploratory techniques and applied these models to the WDBC dataset. Techniques such as feature distribution, correlation, elimination, and hyperparameter optimization were explored to classify tumors into malignant and benign classes [2].

Seddik et al. proposed a method based on tumor variables for a binary logistic model to diagnose breast cancer WDBC data . The proposed model classifies the WDBC data into malignant and benign and accomplished the 98% average classification accuracy by finding that area, texture, concavity, and symmetry are significant WDBC features [3]. Mert et al. developed a breast cancer prediction model using a k-nearest neighbor classifier and independent component analysis for feature reduction, achieving 91% accuracy. This method efficiently distinguished malignant from benign cases with a simplified dataset [4].

Wolberg et al. discovered that computer-analyzed features like nuclear size and shape in breast cancer cells are strong survival indicators, potentially outperforming traditional staging methods. This finding suggests a less invasive way to assess patient prognosis and plan treatments [5].

These studies highlight the WBCD's role in distinguishing between benign and malignant breast tumors by analyzing tumor characteristics. They advance our understanding of breast cancer diagnostics and confirm the value of computational methods in improving diagnostic precision.

Data Description

Our main goal is to see if there are clear differences in certain features of tumors that can tell us if a tumor is malignant (cancerous) or benign (not cancerous). In our analysis of the Breast Cancer Wisconsin dataset, we specifically focus on a subset of feature variables: *radius_mean, perimeter_mean, area_mean, compactness_mean, smoothness_mean*, and *symmetry_mean*, along with targeted binary feature *diagnosis which classifies case to be either Benign or Malignant.*

This deliberate selection is driven by the aim to concentrate on variables that are most relevant to our study's objectives, ensuring a focused analysis. The dataset, comprising over 30 variables, includes numerous measures that, while informative, may not directly contribute to our current analysis. Narrowing down these key features, not only enhances the clarity and efficiency of our analysis but also avoid the potential dilution of results that could arise from incorporating less targeted set of variables.

Below table provides description of feature variables along with targeted variable *Diagnois :*

Table 1: Breast Cancer Wisconsin Dataset Features

Features	Description	DataType
id	Identifier	Integer
diagnosis	Malignant or Benign tumor	Categorical
radius_mean	Mean of distances from center to points on the perimeter	Float
texture_mean	Mean of gray scaled values	Float
perimeter_mean	Mean size of the core tumor	Float
area_mean	Mean area of the tumor	Float
smoothness_mean	Mean of local variation in radius lengths	Float
compactness_mean	Mean of perimeterˆ2 / area - 1.0	Float

Below table provides an efficient pathway to explore the dataset's characteristics. It offers a statistical overview that directly supports. This exploratory analysis is crucial for validating the proposed hypothesis and sets the stage for further statistical testing.

Table 2: Statistical Summary of Selected Variables

	mean	median	sd	min	max
radius_mean	14.1385018	13.375000	3.5169865	6.98100	28.1100
texture_mean	19.2804049	18.835000	4.2991664	9.71000	39.2800
perimeter_mean	92.0465845	86.290000	24.2498190	43.79000	188.5000
area_mean	655.7234155	551.400000	351.6606424	143.50000	2501.0000
smoothness_mean	0.0964373	0.095895	0.0139560	0.06251	0.1634
compactness_mean	0.1044479	0.093125	0.0527977	0.01938	0.3454

Method

To investigate the significance of a group of continuous "mean" variables in determining the diagnosis of breast cancer (benign B or malignant M), logistic regression was chosen as the primary analytical approach. Logistic regression is particularly well-suited for this analysis due to its ability to handle binary outcome variables (such as tumor diagnosis) and continuous predictors (such as the "mean" variables representing various tumor characteristics). This method is well-suited for analyzing dichotomous outcomes in relation to continuous features (the "mean" variables). The logistic regression test in this scenario can be described by the below formula :

$$\boldsymbol{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1} \cdot \text{radius\_mean} \boldsymbol{+ \beta_2} \cdot \text{smoothness\_mean} \boldsymbol{+} \ldots$$

Unlike linear regression, logistic regression models the probability of an event (e.g., tumor malignancy) based on predictor variables, making it suitable for investigating the collective impact of continuous variables on diagnosis. It provides interpretable coefficients and significance tests for each predictor, allowing for the identification of significant contributors to the predictive model. Overall, logistic regression offers a robust and interpretable framework for analyzing the relationship between continuous predictors and binary outcomes in breast cancer diagnosis. To conduct the targeted analysis I have formed following hypothesis :

- Null Hypothesis (H0): The group of continuous "mean" variables collectively doesn't have any statistical significance in determining whether a tumor is benign (B) or malignant (M).

$$\boldsymbol{H_0 :} \beta_{\text{radius}} = \beta_{\text{texture}} = \beta_{\text{perimeter}} = \beta_{\text{area}} = \beta_{\text{smoothness}} = \beta_{\text{compactness}} = \mathbf{0}$$

- Alternative Hypothesis (H1): The group of continuous "mean" variables is collectively significant in determining whether a tumor is benign (B) or malignant (M).

$$\boldsymbol{H_1 :} \text{ At least one } \boldsymbol{\beta_i \neq 0\ (i =} \text{ radius, texture, perimeter, area, smoothness, compactness)}$$

Essentially, the test aims to identify at least one feature variable that significantly relates to the target variable, enabling us to reject the null hypothesis. As described in the code below, bsefore conducting the analysis, I preprocessed the data to ensure its suitability for logistic regression. Specifically, I converted the diagnosis variable into a binary format, where B was mapped to 0 and M was mapped to 1, to facilitate the modeling process.

```r
data$diagnosis <- ifelse(data$diagnosis == "B", 0, 1)

X <- data[c("radius_mean", "texture_mean", "perimeter_mean","area_mean",
            "smoothness_mean","compactness_mean")]
X <- cbind(Intercept = 1, X)
y <- data$diagnosis

# conducting binomial logistic regression test
m <- glm(y ~ ., data = X, family = binomial(link = "logit"))
library(broom)
m1 <- tidy(m)
m1$significant <- m1$p.value < 0.05
m2 <- m1 %>%
  filter(term != "(Intercept)")  # Exclude the intercept
m3 <- m2 %>%
  filter(term != "Intercept")  # Exclude the intercept
```

Results

The analysis of these variables provided mixed results in terms of P-Value and coefficiency. Some variables, like *radius_mean,texture_mean,area_mean and smoothness_mean* are statistically significant predictors of malignancy, showing strong evidence against the null hypothesis (H0) that these variables are essential in determining the nature of the tumor. In contrary, variables like *perimeter_mean* and *compactness_mean* did not show statistical significance at the usual 0.05 threshold, indicating no evidence to reject the null hypothesis for these predictors specifically.

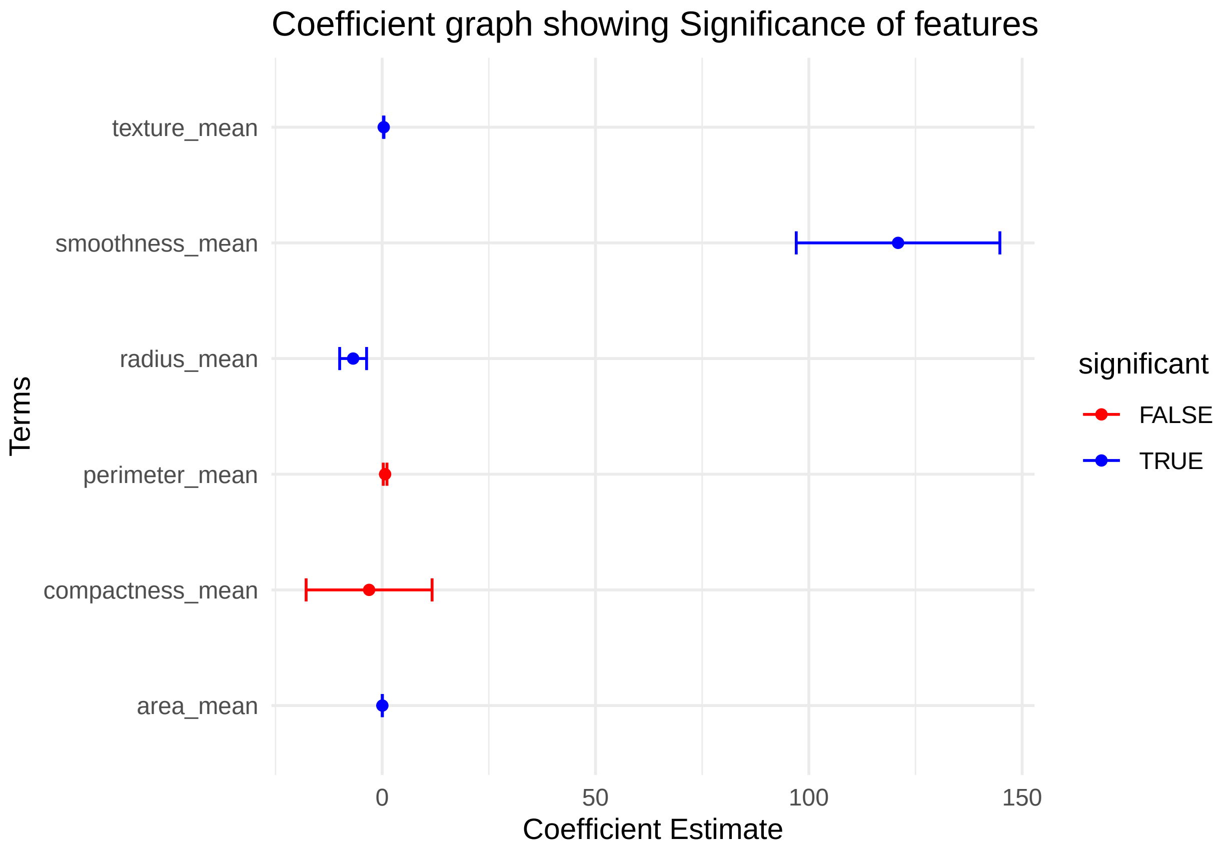
However an interesting observation regarding the negative coefficient estimates for radius_mean (estimate= -0.68057686) indicates that increasing radius_mean by 1 unit, the log odds of the tumor being malignant decrease by 0.68057686.

Overall, considering the significance of multiple variables in our model, there is evidence to reject the null hypothesis in favor of the alternative hypothesis that the group of continuous "mean" variables has statistical significance in determining the nature of the tumor. However, the contribution of each variable varies, underscoring the importance of examining each predictor's role in the model. These results are formatted in the below summary table.

Table 3: Regression Test Summary

term	estimate	std.error	statistic	p.value	significant
radius_mean	-6.8057686	3.1566175	-2.156032	0.0310812	TRUE
texture_mean	0.3642435	0.0600322	6.067473	0.0000000	TRUE
perimeter_mean	0.6903039	0.4401178	1.568453	0.1167755	FALSE
area_mean	0.0416519	0.0137787	3.022916	0.0025035	TRUE
smoothness_mean	120.9118087	23.8613785	5.067260	0.0000004	TRUE
compactness_mean	-3.0605672	14.7607487	-0.207345	0.8357404	FALSE

The logistic regression coefficient graph shown below depicts the estimated parameters and their significance, with red indicating non-significant and blue signifying significant P values rejecting null hypothesis.



Conclusion

Our analysis of the Breast Cancer Wisconsin (Diagnostic) dataset has provided important insights into the diagnosis of breast cancer, focusing specifically on distinguishing between benign and malignant tumors based on selected features. The analysis focused on specific tumor features from the Wisconsin Breast cancer dataset such as radius_mean, area_mean, compactness_mean, smoothness_mean, and symmetry_mean, and establish their significance with tumor outcomes 'Benign' and 'Malignant'.

The analysis primarily focused on Binary logistic regression analysis due to continuous nature of independent predictor variable and dependent predictor variable. It suggested that certain features within the dataset, notably texture_mean, area_mean, and smoothness_mean, and radius_mean significantly contribute to differentiating between malignant and benign breast tumors. These findings are in line with prior research mentioned in the backgroun section, highlighting the relevance of these features in diagnostic processes. While some variables demonstrated a clear statistical significance in predicting tumor malignancy, others, such as perimeter_mean and compactness_mean, did not show a significant impact at the conventional threshold. This variance underscores the complexity of breast cancer and suggests that a multifactorial approach is necessary for accurate diagnosis.

Apart from the significance analysis, this study also showed the positive/negative correlation between features and predictor. An intriguing point was the discovery that an increase in radius_mean actually lowers the chance of a tumor being malignant. The reverse trend was found in all other significant variables such as texture_mean, area_mean and smoothness_mean where probability of tumor being malignant increases with the increase in these features.

However this doesn't capture the entire complexity of breast cancer. Not all variables showed statistical significance, which reminds us of the disease's multifaceted nature. Cancer's development and progression are influenced by a a whole lot of other features beyond the scope of our study, including genetics, lifestyle, and environmental aspects. This approach also has some methodological constraints. Although the regression test is powerful, it has its limitations such as the potential for oversimplification of complex biological interactions.

This analysis not only supports findings from previous studies cited but also extends their narrative by providing a fresh perspective through the logistic regression approach. The study helps us see interplay of various factors in breast cancer diagnosis, thereby contributing to the advancement of detection and treatment strategies for the disease.

References

[1] Mehak, S., Ashraf, M. U., Zafar, R., Alghamdi, A. M., Alfakeeh, A. S., Alassery, F., Hamam, H., & Shafiq, M. Z. (2022). Automated grading of breast cancer histopathology images using multilayered autoencoder. Computers, Materials & Continua, 71(2), 3407–3423. https://doi.org/10.32604/cmc.2022.022705

[2] Rasool, A., Bunterngchit, C., Luo, T., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved Machine Learning-Based Predictive Models for breast cancer diagnosis. International Journal of Environmental Research and Public Health, 19(6), 3211. https://doi.org/10.3390/ijerph19063211

[3] A. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK, 2015, pp. 150-154, doi: 10.1109/IntelliSys.2015.7361138.

[4] Mert, A., Kılıç, N., Bilgili, E., & Akan, A. (2015). Breast Cancer Detection with Reduced Feature Set. Computational and Mathematical Methods in Medicine, 2015, 1–11. https://doi.org/10.1155/2015/265138

[5] Wolberg, W. H., W, N. S., & Mangasarian, O. L. (1999). Importance of nuclear morphology in breast cancer prognosis. American Association for Cancer Research. https://aacrjournals.org/clincancerres/article/5/11/3542/286159/Importance-of-Nuclear-Morphology-in-Breast-Cancer

[6] Breast Cncer Wisconsin Dataset https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data