A comparison of Logistic regression and Random forest on Breast Cancer Analysis

Description and motivation of the problem

The aim of this research/analysis is to address the binary classification task of accurately distinguishing between benign (B) and malignant (M) tumors using the Breast Cancer Wisconsin (Diagnostic) dataset. Supervised machine learning algorithms, which include Logistic Regression (LR) and Random Forest (RF), will be utilized in this task. By leveraging these two distinct approaches, their comparative effectiveness will be explored in a medical diagnosis context, where it is extremely important to do a correct classification as it can significantly impact a patient and how they will go about treating it. The models' performance will be essentially assessed against the 5 key metrics such as accuracy, precision, specificity, sensitivity and F1-score, providing a complete evaluation of their predictive capabilities. Moreover, the results will be compared with other results from previous research to contextualize the models' performance within the broader field of machine learning applications in healthcare. This research does not only aim to contribute to the advancement of diagnostic procedures but also it aims to reinforce the potential of machine learning as a tool for enhancing decision-making in healthcare.

Initial analysis of the dataset including basic statistics / EDA

The dataset visualized in the first image originates from biomedical measurements related to breast cancer tumors. Features included contain radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, all of which are measured as means, standard errors, and worst (largest three values) for each image, resulting in thirty features (columns)

- **Pair plot:**

In figure (3), Each feature is analyzed for its distribution and relationship to the diagnosis (Malignant or Benign), shown as histograms and scatter plots with linear regression lines.

- **Pie chart:**

Figure (2) depicts the distribution of type of cancer in a pie chart, categorizing tumors as either Benign or Malignant. Malignant tumors took up 37.25% of the total data which means 212 instances were malignant and 62.75% was benign (357 instances).

- **Correlation Matrix Heatmap:**

Figure (1) represents a correlation matrix heatmap that displays the relationships between all the numeric features in a dataset.

Darker colors represent stronger correlations, and this visualization helps in identifying which features are most strongly related to each other, guiding feature selection for modeling.

- **PCA Variance Analysis:**

Figure (4)shows a scree plot from Principal Component Analysis (PCA), displaying the cumulative explained variance by the number of components used.

This analysis helps in dimensionality reduction by indicating how many components should be retained to preserve a certain percentage of the data's total variance.

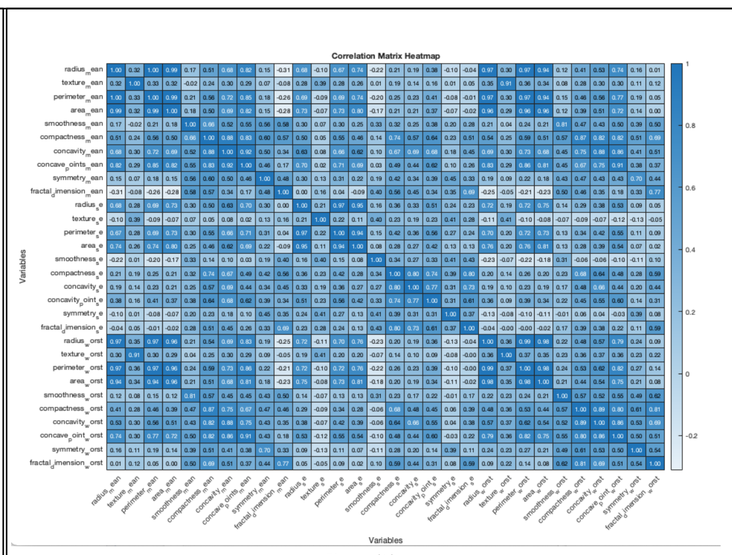

Figure (1)

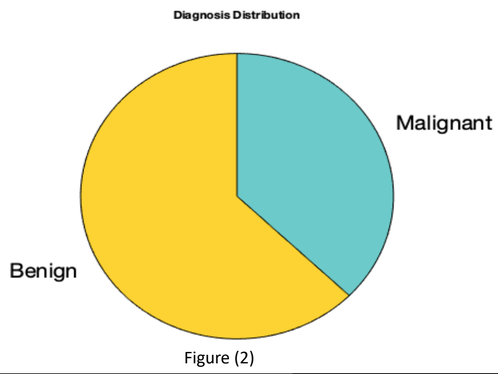

Figure (2)

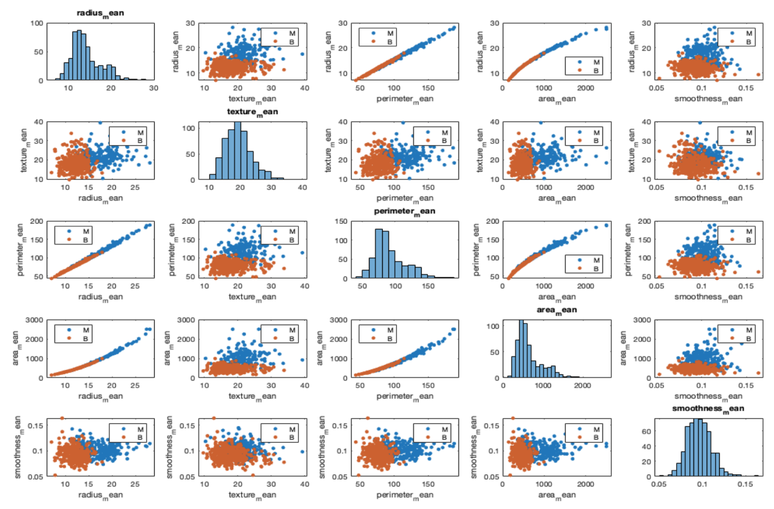

Figure (3)

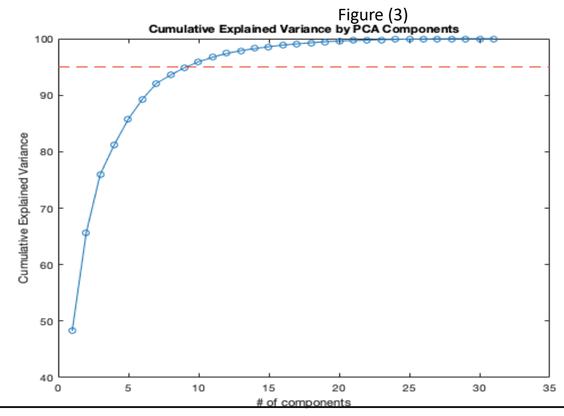

Figure (4)

Random Forest

Random Forest is an supervised learning algorithm that operates by constructing a multitude of decision trees at training time after that it outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. (in the case of this research, it performs classification) (Fleiss, 2023)

Pros:

- Handles high dimensionality and a large number of training examples well.(E R, 2021)
- Able to model complex interactions between features.
- Provides a measure of feature importance.(Donges, 2023)
- Usually has a high accuracy and good performance on many problems without overfitting.

Cons:

- Can be computationally intensive, leading to longer training times.
- Model outcomes can be difficult to interpret compared to simpler models.
- Less effective on very sparse data, like text data.
- Requires a good balance between the number of trees and depth to avoid overfitting and underfitting.

Logistic Regression

Logistic Regression is a statistical model where a logistic function is used to model a binary dependent variable, on the other hand other complex extensions exist to make it usable for multiclass classification and regression.

Pros:

- Quick to train and easy to apply regularization to using techniques like L1 or L2 penalties. (Rout, 2020)
- Outputs have a probabilistic interpretation, which can be a threshold to classify observations.
- Well-suited for binary classification problems. (Rout, 2020)
- Model is inherently simple, which makes it interpretable.(Rout, 2020)
- **Cons:**
- Assumes a linear relationship between the log odds of the outcome and the input variables, which may not always hold.
- More complex models can outperform it in datasets with more complex relationships.
- Tends to overfit; requires feature selection or regularization to mitigate.
- Not suitable for non-linear problems unless you transform features or extend the model.

Hypothesis statement

- Based on literature reviews, Logistic Regression is known for being highly interpretable and does training quickly for datasets that have linearly separable classes (James et al., 2013).
- On the other hand, Random Forest usually performs better in handling complex, non-linear data patterns, but has higher computational costs (Breiman, 2001).
- In the case of having imbalance in the dataset, the probability of Random Forest misclassifying the minority class is lower because of its ensemble approach, potentially offering a more robust solution against overfitting compared to LR.

Description of the choice of training and evaluation methodology

- Set a random seed to make the code reproduceable.
- The original dataset got split into training and testing sets in a 80:20 split, and the test set remained unseen to models until they were trained on the training set.
- Encoded labels to the two classification where 0 represented (Benign) and 1 represented Malignant.
- Implement a scaler
- Apply 5-fold cross validation on training set for Logistic Regression, and applying out of bag error estimation in Random Forest.
- Using OOB error and cross validation to find the mot optimal hyperparameters that would yield the highest scores in the performance metrics.
- After yielding the most optimal hyperparameters, Train the models and check the training score.
- Test the two models on the test set using the hyperparameters, check the performance metrics, and compare results.

Choices of parameters and experimental results

Logistic Regression

- Using multiple regularization strengths defined by the hyperparameter 'C', the model is fitted.
- By using 5-fold cross validation, the optimal C value is obtained. In which, the model complexity is balanced as well as the prediction accuracy
- To prevent overfitting, regularization is then applied where Lambda is inverse of C.

Parameters

- The most optimal C value was 100 and it was determined from a set [0.01,0.1,10,100]
- Hence, the most optimal Lambda value which is the inverse of C was 0.01.

Random Forest

- The Random Forest model is optimized over a grid of hyperparameters, including the number of trees and whether the data gets bootstrapped or no.
- variations in the number of estimators (n_estimators) and the method for splitting nodes (criterion) are included in the search..

The choice of parameters:

- The most optimal size for the number of trees was 20 and it was searched within a predefined set [20, 50, 100, 150, 200].
- Splitting criterion is chosen between gdi and deviance, and whether to bootstrap samples or not.
- There was the choice between 'gdi' and 'deviance' where the optimal answer was 'gdi'.
- The bootstrapping was shown as '1' which means it would be best to add bootstrapping, which is already added by default in the TreeBagger.

Note: the confusion matrix for random forest isn't shown because it got 100% accuracy which means that everything was correctly classified.

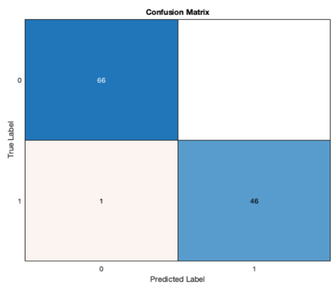

Figure (1)

```
The best parameters for the classifier
    n_estimators: 20
       criterion: 'gdi'
       bootstrap: 1

The best out-of-bag score is: 0.719

Accuracy=1.000
Sensitivity=1.000
Specificity=1.000
Precision=1.000
F-measure=1.000

Accuracy = 1
Sensitivity = 1
Specificity = 1
Precision = 1
F_measure = 1
```

Figure (2)- RF

```
The best parameters for the classifier are:
                C: 100
           Lambda: 0.0100

The best training score is: 0.991

Accuracy=0.991
Sensitivity=0.979
Specificity=1.000
Precision=1.000
F-measure=0.989

Accuracy = 0.9912
Sensitivity = 0.9787
Specificity = 1
Precision = 1
F_measure = 0.9892
```

Figure (3)- LR

Analysis and Critical Evaluation of Model Results

Logistic Regression Analysis

The Logistic Regression model exhibits outstanding performance on the test dataset, with an accuracy of 0.991, sensitivity of 0.979, specificity of 1.000, precision of 1.000, and an F-measure of 0.989. These metrics indicate that the model not only predicts the majority class well but also effectively identifies the minority class without significant false positives, as evidenced by the perfect precision and specificity scores.

Optimal Parameters and Overfitting Consideration: The optimal hyperparameters, with C set to 100 and Lambda to 0.0100, suggest that the model benefits from a lower degree of regularization, allowing for more complexity in the decision boundary. This is appropriate given the high accuracy score, which signifies that the model has learned to generalize rather than merely memorize the training data. However, the slight difference between training and test accuracy implies that we should remain vigilant about potential overfitting, especially in future applications with different or larger datasets.

Evaluation of Fit and Predictive Power: The confusion matrix for Logistic Regression shows a near-perfect classification with just one misclassification. This demonstrates the model's robustness and its potential for deployment in scenarios where high sensitivity is crucial. However, we must be cautious about interpreting these results as they may be influenced by the test set's size and composition. Additional validation using techniques such as k-fold cross-validation on a larger dataset could provide further confirmation of the model's effectiveness.

Random Forest Analysis

The best out-of-bag score in random forest was 0.719, and the optimal parameters that were used was 20 trees with gdi criterion and enabling bootstrap. The oob score is usually a great estimate of test accuracy and in the case of our model, it indicates that the model is more likely to perform well to unseen data, but less than the logistic regression model.

Hyperparameter Tuning and Model Complexity: Interestingly, the optimal number of trees is relatively low, which may indicate that adding more trees does not significantly improve performance, possibly due to the nature of the dataset being well-represented by fewer decision boundaries. The use of gdi as the criterion for node splits could suggest that the default measure of impurity reduction was sufficient for the dataset at hand. The choice to implement bootstrapping is consistent with standard Random Forest practice, promoting model diversity and robustness against overfitting.

Performance Metrics and Interpretation: In this model, the optimal number of trees is low which is seemingly weird but it indicates that adding more trees will not change anything possibly due to the dataset's nature where its well-represented by fewer decision boundaries. Using 'gdi' suggests that the default measure of impurity reduction was sufficient for our dataset. The implementation of bootstrapping in random forest is a consistent with its practice, because it promotes model diversity and increases robustness against overfitting.

General Observations: The confusion matrix confirms the high accuracy of the Random Forest model. However, the distribution of errors and the slightly lower out-of-bag score relative to the Logistic Regression model suggest that while the Random Forest is powerful, it may require careful tuning and possibly more data to achieve its full potential, particularly in managing the balance between variance and bias

Future Work

- Use larger datasets and use k-fold cross validation more to ensure robustness.
- Implement strategies to address imbalanced data
- Combining models could leverage their strength.
- Using more advanced techniques,like Bayesian optimization to find the most optimal hyperparameters.

Lessons Learned

- This project highlighted the importanceof chossing the right model complexity
- Pay attention to data imbalances, and look at the whole performance metrics not just accuracy

References

- archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository*. [online] Available at: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.
- Blei, D. (2012). *Model-Based Classification Probability models*. [online] Available at: https://www.cs.princeton.edu/courses/archive/spring12/cos424/pdf/model-classification.pdf [Accessed 20 Dec. 2023].
- Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.
- Donges, N. (2023). *What Is Random Forest? A Complete Guide | Built In*. [online] builtin.com. Available at: https://builtin.com/data-science/random-forest-algorithm#:~:text=One%20big%20advantage%20of%20random [Accessed 20 Dec. 2023].
- E R, S. (2021). *Random Forest | Introduction to Random Forest Algorithm*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=One%20of%20the%20most%20important.
- Fleiss, A. (2023). *What are the advantages and disadvantages of random forest?* [online] Rebellion Research. Available at: https://www.rebellionresearch.com/what-are-the-advantages-and-disadvantages-of-random-forest#:~:text=February%2019%2C%202023%20What%20are [Accessed 20 Dec. 2023].
- Gareth Michael James, Witten, D., Hastie, T.J. and Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. New York: Springer.
- Khairunnahar, L., Hasib, M.A., Rezanur, R.H.B., Islam, M.R. and Hosain, M.K. (2019). Classification of malignant and benign tissue with logistic regression. *Informatics in Medicine Unlocked*, 16, p.100189. doi:https://doi.org/10.1016/j.imu.2019.100189.
- OpenGenus IQ: Computing Expertise & Legacy. (2020). *Advantages and Disadvantages of Logistic Regression*. [online] Available at: https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/#:~:text= [Accessed 20 Dec. 2023].
- Rout, A.R. (2020). *Advantages and Disadvantages of Logistic Regression*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/.