Breast cancer prediction application report

Palak Jagani

2024-04-08

1. Background and past work

Breast cancer remains one of the most significant health challenges facing women worldwide, with millions of new cases diagnosed annually. The disease's impact is profound, as it not only affects the health of individuals but also imposes significant emotional and financial burdens on patients and their families. Early detection is critical in improving treatment outcomes and survival rates, as it allows for the timely initiation of therapy before the cancer advances to more lethal stages.Traditionally, breast cancer detection has relied on physical examinations, mammography, and biopsies, which, while effective, also come with limitations such as false positives and negatives, as well as being invasive. Advances in computational sciences, particularly in machine learning and artificial intelligence, have paved the way for the development of predictive models that can potentially revolutionize how breast cancer is diagnosed. These models can process vast datasets to identify patterns and predict outcomes with high accuracy, offering a valuable tool for clinicians.

In a previous assignment, the focus was placed on the statistical analysis of diagnostic variables associated with breast cancer, to identify their significance in determining the benign or malignant nature of tumors. Utilizing regression test and evaluating p-values, we were able to identify which features held statistically significant correlations with cancer outcomes. This methodical approach allowed for a precise understanding of how individual diagnostic measurements—such as tumor size, shape, and texture—contributed to tumor classification.

Reza et. al Combining multiple risk factors in modeling for breast cancer prediction could help the early diagnosis of the disease with necessary care plans. Collection, storage, and management of different data and intelligent systems based on multiple factors for predicting breast cancer are effective in disease management [1]. Rasool and his team created predictive models for breast cancer using exploratory data analysis techniques on the WDBC dataset, employing methods like feature distribution, correlation, elimination, and hyperparameter tuning to categorize tumors as malignant or benign [2]. Seddik and associates developed a binary logistic model using tumor variables that successfully classified the WDBC data into malignant and benign categories with an average accuracy of 98%, identifying area, texture, concavity, and symmetry as key features [3]. Mert and his group formulated a breast cancer prediction model utilizing the k-nearest neighbor classifier and independent component analysis for feature reduction, which achieved a 91% accuracy rate and effectively differentiated between malignant and benign cases using a simplified dataset [4].

Building on this foundation work, the current Breast Cancer Prediction application extends these insights into the practical domain of predictive modeling through advanced machine learning techniques such as SVM, Random Forest, KNN and Logistic regression. On one hand, the initial project provided a baseline understanding of the relationships between variables and cancer outcomes, this application leverages that data to train sophisticated algorithms capable of making real-time predictions. By applying these models the application not only predicts the nature of tumors but also enhances the predictive accuracy beyond the capabilities of traditional statistical tests.

2. Introduction

The Breast Cancer Prediction application uses R and Shiny to help detect and classify breast cancer early. This tool turns complex statistical methods and machine learning algorithms into a user-friendly interface that users can easily use to evaluate the characteristics of the breast cancer tumor. In its advanced mode, It can be designed to help professionals quickly understand how different factors related to breast cancer tumor are connected, which is crucial for making fast and accurate diagnoses.

The application includes tools for displaying data visually. These tools show how various diagnostic factors are related, making it easier for medical professionals to grasp complicated information quickly. Additionally, the application can analyze patient data instantly, providing fast feedback that is vital for making decisions in medical settings.

Although the main aim of this application is to give the non-technical user the ability to predict the tumor diagnosis, for the technical users the application provides a good overview of the different machine learning application and their performance comparision.

The web application can be accessed using this link : https://palak2803.shinyapps.io/BreastCancerApplication_Final/

The source code can be accessed using the following github link : https://github.com/palak2803/BreastCancerPrediction/tree/main

The interface of the Breast Cancer Prediction application is straightforward and easy to use, it's important to note that this is still a application designed for academic purposes. It works well within the scope of its dataset, but might not yet handle the variety and complexity of real-world clinical data. This report will explore the main features of the application, including its data processing, data visualization the types of models it uses for predictions, and its user interface. The aim is to show how these elements combine to make this application a useful tool for diagnosing and treating breast cancer.

3. Implementation

3.1 Data Handling

The application makes use of detailed patient data retrieved from the Wisconsin Breast Cancer Database (WBCD), which includes diagnostic features such as the mean radius, texture, perimeter, area, and various other cellular characteristics of breast masses. To prepare this data for analysis, initial preprocessing steps are undertaken to convert categorical outcomes from 'Malignant' (M) or 'Benign' (B) into a binary format (1 for Malignant, 0 for Benign). This conversion makes it easy to do classification process and ensures that the data fed into the machine learning algorithms is clean, standardized, and optimally formatted for analysis. No other preprocessing is required since all the feature variables are continuous and mostly follows the normal distribution. The following R code snippet demonstrates this preprocessing step:

```r
# Load the dataset
data_temp <- read.csv("wdbc.csv", stringsAsFactors = TRUE)

# Convert diagnosis to a binary factor
data_temp$diagnosis <- as.factor(ifelse(data_temp$diagnosis == "M", 1, 0))
```

3.2 Model Development

The core functionality of the application is supported by four distinct predictive models, each selected for its proven capabilities in handling binary classification tasks efficiently. Each model is trained using 80% of the dataset, reserving the remaining 20% for testing. This training-validation split ensures that the models are not only trained on comprehensive data but are also capable of performing well on unseen data, simulating real-world application. Model validation is rigorously conducted, assessing metrics such as accuracy, precision, recall, and F1 scores. ROC curves and confusion matrices are generated to provide a complete performance overview:

1. Logistic Regression Logistic Regression is utilized to estimate the probability of a tumor being malignant based on the input variables. It provides a fundamental probability model that serves as a baseline for comparison with more complex algorithms. The logistic regression model is implemented as follows:

```
# Train Logistic Regression Model
logisticModel <- glm(diagnosis ~ ., data = cbind(trainX, diagnosis = trainY), family = "binomial")
```

This model is particularly useful for its interpretability and the straightforward probability scores it generates.

2. Support Vector Machine (SVM) SVM is deployed to find the optimal hyperplane that separates malignant cases from benign ones within the high-dimensional space of breast cancer features. This method is robust against overfitting and is known for its accuracy in classification tasks involving complex datasets:

```
# Train SVM Model
svmModel <- svm(diagnosis ~ ., data = cbind(trainX, diagnosis = trainY), type = 'C-classification', ker
```

3. Random Forest Random Forest builds multiple decision trees and merges them to produce a more accurate and stable prediction. This model also provides significant insights into which features are most influential in diagnosing breast cancer, thereby enhancing the interpretability of the results:

```
# Train Random Forest Model
randomForestModel <- randomForest(diagnosis ~ ., data = cbind(trainX, diagnosis = trainY), importance=T
```

The ensemble approach of Random Forest minimizes overfitting and is excellent for handling large datasets with high feature dimensionality.

4. K-Nearest Neighbors (KNN) KNN classifies each case based on the most common outcome among its nearest neighbors, providing effective predictions especially when the decision boundary is ambiguous:

KNN is included for its simplicity and effectiveness, particularly useful in this application for validating the robustness of the other models.

```
# Train KNN Model
knnModel <- knn3(diagnosis ~ ., data = cbind(trainX, diagnosis = trainY), k = 5)
```

3.3 Model postprocessing :

After training machine learning models, several crucial post-processing steps were implemented to ensure their readiness for practical use in the app.

The first step involves calculating the accuracy of each model by comparing predicted outcomes against actual results from the test dataset. This provides a clear measure of each model's ability to classify breast cancer accurately

Following the accuracy evaluation, confusion matrices were generated for each model, offering a detailed view of their performance, including true positives, false positives, true negatives, and false negatives. This helps us in developing sensitivity and specificity metrics of the given models. It also helps us in identifying strengths and weaknesses in specific diagnostic scenarios, guiding potential refinements to improve precision.

Finally, the validated models are saved to disk using the `saveRDS` function. This ensures they can be easily reloaded for use within the `app.R` script, supporting the application's functionality. This capability allows the app to deliver immediate and precise diagnostic predictions, enhancing clinical decision-making processes.

```
# Calculate accuracies and generate confusion matrices
logisticPreds <- ifelse(predict(logisticModel, newdata = testX, type = "response") > 0.5, 1, 0)
logisticAccuracy <- mean(logisticPreds == testY)
...
logisticCM <- confusionMatrix(as.factor(logisticPreds), as.factor(testData$diagnosis))
...
```

3.4 Application execution :

Upon launching the web application, the `app.R` script begins by loading pre-trained machine learning models such as Logistic Regression, SVM, Random Forest, and KNN from saved RDS files from the same directory. It then runs the front-end and back-end components UI and server simultaneously. The UI components generate the layout of the appication that communicates with server components for their interactive capabilities. The user interface, displayed next, offers interactive elements including sidebar menus for navigating between data visualization, model predictions, and other functionalities. These functionalities are part of the components defined in the same file app.R.

The server component of `app.R` captures the input data, which is then formatted and fed into the machine learning models for inference. Each model outputs predictions on whether the tumor is benign or malignant, which are then displayed along with accuracy metrics and confusion matrices in the “Results” tab. Dynamic visualizations in the “Visualization” tab provide graphs and charts that reflect the latest inputs and model outputs. Links to comprehensive documentation, source code, and a downloadable report are also being handled by the same server component.

4. How to Use the Application

4.1 Access

To begin using the Breast Cancer Prediction application, navigate to its web address application link using any standard web browser. The application is deployed to shinyapps.io server using rsconnect() function and hosted online, allowing for easy and immendiate access without any local server installation or execution. The application source code can be found at the github link posted on the top right corner of the application interface.

4.2 Navigate

Once the application has loaded, sidebar navigation menu is visible on the left side of the screen. This menu is our primary tool for navigating between different sections of the application, each dedicated to specific functionalities:

- **About**: This tab provides a comprehensive overview of the application’s purpose, the data it uses, and the technology behind it.
- **Data**: This section gives users the ability to explore the complete dataset as it is used within predictive models and visualizations. The section includes interactive tables that display patient data for breast cancer tumor characteristics and associated results of malignancy or benign.The section has 3 subsections :
    1. Data : Provides complete picture of the raw dataset used in this application.
    2. Structure : It provides structural overview of the features used such as their description and datatypes
    3. Summary : It provides statistical summary of all the features such as their mean, median, max, min and standard deviation.
- **Visualization**: Users can explore various data visualizations that offer insights into the dataset’s distribution and feature correlations. This section helps in visualizing complex relationships within the data.
    1. Feature Distributions Plot : These histograms show the distribution for each diagnostic feature, such as radius mean and texture mean. They are crucial for assessing the data characteristics like variability and outliers, which influence model predictions.
    2. Correlation Heatmap : This heatmap illustrates the correlations between all quantitative features, using color coding to highlight the strength and nature of these relationships. It helps identify potentially redundant features or those that strongly influence others, which is valuable for refining model inputs.
    3. Feature Importance : The feature importance graph ranks the diagnostic features based on their impact on the model’s predictions. This visualization helps pinpoint which features are most predictive of breast cancer outcomes, guiding users in focusing on the most influential data points.
- **Model Prediction**: This is where users can make practical use of the predictive models by entering new patient data. These sections are designed to be intuitive for users of all technical levels, from medical professionals to data analysts interested in healthcare applications. This section has 3 subsections :

1. Input Data : In the ‘Model Prediction’ tab, you will find input fields for each diagnostic measurement required by the predictive models. The application is loaded with the default values. Alternatively, the user can also enter the values for each measurement based on the patient’s diagnostic data. These inputs are crucial as they directly influence the prediction outcomes. Ensure the accuracy of the data entered to receive reliable predictions.
2. Results : After entering the necessary data, click the ‘Predict’ button located beneath the input fields. This action triggers the application’s server to process the data through the integrated machine learning models—Logistic Regression, SVM, Random Forest, and KNN. The models evaluate the data and calculate the likelihood of malignancy based on the inputted diagnostic measurements.
3. Model performance : This tab provides a comprehensive evaluation into each of the predictive models performance. It runs features ROC curves and confusion matrices calculation in the background to display the information on accuracy, sensitivity, and specificity of the models.

5. Conclusion

The Breast Cancer Prediction application serves as a prototype that demonstrates how predictive algorithms can be applied to enhance breast cancer diagnostics. On one hand, it gives non-technical users the ability to predict the outcome by adding the tumor characteristics, it also successfully integrates various analytical approaches through tools—such as feature distributions, correlation heatmaps, and performance metrics. Although the application is trained on the widely accepted, cited and well-researched dataset, the application is not without its limitations, it may not capture the full complexity of real-world scenarios and the potential for over fitting given the limited scope of testing. As a result we can see that, some of the models have same accuracy despite having different sensitivity and specificity. For future work, it would be beneficial to expand the dataset, add-in more diverse diagnostic variables, and apply advanced model validation techniques such as Deep learning models to improve the robustness and accuracy of the predictions. This project establishes the groundwork for further exploration and refinement, aiming to evolve into a more comprehensive tool that can reliably support clinical decision-making in a real-world medical context.

6. References

[1] Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022, June 1). *Prediction of breast cancer using machine learning approaches*. Journal of biomedical physics & engineering. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/

[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