Machine Learning in Breast Cancer Diagnostic Dataset

[redacted]

Western Governors University

A. Project Overview

Research Question

Utilize machine learning techniques to conduct diagnostic assessments using pre-existing breast cancer datasets and test whether certain features are more important than others in determining tumor malignancy.

Scope Of Project

In this study, all training and testing data used is sourced from the WDBC (Wisconsin Diagnostic Breast Cancer) dataset. Specifically, the analysis focuses solely on the features available within the dataset, which include parameters such as "radius" (the mean of distances from the center to points on the perimeter), "texture" (the standard deviation of gray-scale values), and "perimeter" and more. Notably, demographic information about the patients, such as age, race, gender, and other personal characteristics, is not considered or analyzed in this study. The emphasis is solely on the selected quantitative features related to breast tumors in the dataset.

The primary focus of this study is to assess whether some features play a more significant role in distinguishing between benign and malignant tumors. To obtain a definitive assessment of feature importance or validate the model's selected features, further analysis is necessary after hypothesis testing. It's important to note that this phase of analysis will not be included as part of this project.

Overview Of Solution, With Tools and Methodologies

The objective of this project is to employ a machine learning model to identify the most influential features in the breast cancer dataset about distinguishing between benign and malignant tumors. I anticipate obtaining two key outputs from this analysis:

1) The relationship between features determined by the correlation metric.
2) The top five most influential features identified by the trained and tuned machine learning model.

The project executed utilizing Jupyter Notebook, leveraging various Python libraries, including NumPy and Pandas.

I employed ADDIE methodology in this project. The specific arrangements are outlined as follows:

a) Analysis:

Define the project scope and objective, identify what data will be using. The platform, programming language, machine learning models or statistics test will be selected at this initial phase, after a detail analysis of the topic.

b) Design

1. Data cleaning and processing.
2. Correlation metric for brief introduction of relationship
3. Data splitting, features scaling to ensure data equality and performance.
4. Models training and verification.
5. Model turning.
6. Features select based on weight.

c) Implementation

1. Implement code for design step 1 to 6.

d) **Evaluation:**

1. Testing model with testing data to ensure accuracy.
2. Comparing model output with correlation metric output

B. Summarize Execution of Project

Project Plan

The goal of the project is to identify whether specific features hold greater significance than others in the determination of tumor malignancy. However, when using additional statistical tests to verify the findings, I chose to change the process, I used a correlation metric instead of a t-test in the process.

After changing of the original plan, the actual processes are:

Step	Time
Download data and background study,	Sep 19,2023
Cleaned and processed data	Sep 20,2023
Correlation Metric	Sep 21,2023
Models training and testing	Sep 22,2023
Fine-turn model	Sep 25,2023
Top 5 features selected by model	Sep 27,2023
Compare results from correlation metric and selected model	Sep 29,2023
Summary	Oct 2,2023

Project Planning Methodology

I employed ADDIE methodology in this project. The specific arrangements are outlined as follows:

a) Analysis:
Define the project scope and objective, identify what data will be using. The platform, programming language, machine learning models or statistics test will be selected at this initial phase, after a detail analysis of the topic.

b) Design
1. Data cleaning and processing.
2. Correlation metric for brief introduction of relationship
3. Data splitting, features scaling to ensure data equality and performance.
4. Models training and verification.
5. Model turning.
6. Features select based on weight.

c) Implementation
1. Implement code for design step 1 to 6.

d) **Evaluation:**
1. Testing model with testing data to ensure accuracy.
2. Comparing model output with correlation metric output

Timeline with Milestones for Project

Milestone	Start Date	End Date	Duration
Data collection and background study	Sep 19,2023	Sep 19,2023	1 day
Cleaned and processed data	Sep 20,2023	Sep 20,2023	1 day
Correlation metric	Sep 21,2023	Sep 21,2023	1 day
Trained and tested models	Sep 22,2023	Sep 22,2023	2 days
Fine-turn model	Sep 25,2023	Sep 26,2023	2 days
Top 5 features selected by fine-turn model	Sep 27,2023	Sep 27,2023	1 day
Compare results	Sep 29, 2023	Sep 29, 2023	2 days
Summary	Oct 2, 2023	Oct 2, 2023	1 day

C. Data Selection and Collection Process Including:

Plan and Actual Process

The data I used in this project are public on UC Irvine Machine Learning Repository. The dataset is clean, complete and can be downloaded directly from UCI website. The actual processes as expected.

Obstacles Encounter

The dataset has already been cleaned and processed by the original contributor, and it is free from missing values. Relative information can be found online and free for reference.

Issues Handled

I did not encounter any issues since it is publicly available and well processed data.

1. Advantages And Limitation of Data

The Breast Cancer Wisconsin (Diagnostic) dataset, often referred to as the WBCD dataset, is publicly available and not owned by any specific entity or individual. Researchers and practitioners are allowed to use this dataset for machine learning and data analysis to advance the understanding of breast cancer and develop diagnostic and predictive models without the need for specific permissions or ownership considerations. However, the dataset is small and could be outdate since it is a dataset from 25 years ago.

D. Data Extraction and Data Preparation Processes

The dataset has already been cleaned and processed by the original contributor, and it is free from missing values. I first read data into data frame format on Jupyter notebook with Python. However, the data does not have columns name attach with it, I manually created a list contain columns name, combined the columns and data when reading data. After that, the main operations I performed was dropping the ID column and splitting them into subgroups when necessary. I also mapped “diagnosis” column values to 0 and 1, with 1 indicating a malignant tumor.

E. Report Data Analysis Process

1. Methods Used in Analyze

In this analysis process, I first used a heatmap with Pearson's Correlation Coefficient to display the correlation metrics, assessing the linear relationship between pairs of numerical features. After introducing the dataset briefly, I employed multiple machine learning models, including Gaussian NB, Decision Tree Classifier, and Random Forest

Classifier etc., to compare their performance. Lastly, after selected a model with high accuracy/precision score, I implement feature_importances_ attribute to display the features with more weight in the dataset. The result from this step will be used to compare with the result from correlation metric.

2. Discuss Advantages and Limitation of Tools and Techniques

In our breast cancer dataset, correlation metrics provide a numerical measure of the strength and direction of relationships between variables. It helps quantify the extent to which two variables are associated. However, correlation metrics may not capture non-linear associations, which can be significant in some datasets.

The Decision Tree is a widely used and intuitive machine-learning algorithm for both classification and regression tasks. One well-known example of a decision tree classifier is in Breast Cancer Diagnosis. However, decision trees are prone to overfitting when they become overly complex.

Random Forests have found applications in tasks like fraud detection and image classification, including object recognition and disease detection in medical images. While employing a Random Forest can yield accurate results, the inherent complexity of the model can sometimes obscure its steps, leading to reduced interpretability. Moreover, overfitting remains a concern if the individual trees become excessively deep or the overall model complexity becomes unwieldy.

By analyzing the frequency of feature presence in the dataset, Gaussian Naive Bayes can calculate the likelihood of an individual having a malignant tumor. Gaussian Naive Bayes is especially beneficial for datasets with numerous features, and it can exhibit strong performance even when the assumption of independence is not entirely precise. This algorithm is rapid and efficient, making it a favorable choice for quickly exploring classification problems. However, it's important to note that Naive Bayes assumes feature conditional independence given the class, which might not always hold true in real-world scenarios.

By comparing the results displayed by the correlation metric and the tuned model, we can gain a preliminary understanding of whether certain features play a more significant role in the dataset for breast cancer determination. However, it's essential to acknowledge that additional analysis should be conducted in the future. The similarity in the results of the two tests could be attributed to the relatively small size of the data.

3. Explanation of Methods in Part E1, Verified Data Satisfied, Assumptions or Requirements

With the cleaned and preprocessed dataset, I initiated the analysis by creating a correlation matrix with a heatmap. This step aimed to provide an initial insight into the frequency and relationships between each feature and the target variable.

Next, I conducted feature scaling operations to ensure data uniformity and consistency. This process involved splitting the data into training and test sets.

Following the data preparation, I systematically explored the dataset using various Supervised Learning Models, including decision trees, random forests, and GaussianNB etc. The objective was to identify the model that best suited the data based on criteria such as execution time, accuracy, precision, and F1 score. I used training and validation data in this part.

Subsequently, I employed grid search techniques to fine-tune the model's hyperparameters for improved performance. I used test data to ensure the model final performance.

Lastly, with the optimized model, I aimed to enhance the understanding of the dataset and its diagnostic potential by identifying the top features crucial for breast cancer diagnosis through an examination of feature importance attributes.

F. Evaluate Success if Project

1. Statistical Significance of Analysis with Accurate Calculations

Model: Logistic Regression, RandomForestClassifier, DecisionTreeClassifier or GaussianNB, SVC (linear)

Process:

1) Splitting data to training, validation, and testing group.
2) fitted and predicted data with model.
3) Calculate the time and each metric score.
4) Plot metric scores into bar chart to compare their performance.
5) Select model with overall highest performance score.

Metric: Time, beta_score, accuracy_score, precision_score

Benchmark: Whichever model achieves the highest precision and accuracy scores with less execution time should be selected. The output should resemble the results on UCI (Baseline Model Performance) where the Random Forest classifier achieved nearly 100% accuracy and precision scores. Our results should closely resemble it.

Conclusion: The RandomForestClassifier exhibits the highest accuracy and precision scores. With the results obtained from its feature_importances_ attribute and the

display of the correlation metric, we reject the null hypothesis. There is clear evidence that certain features in the breast cancer dataset are more important than others in predicting breast cancer.

2. Practical Significance of Solution with Examples

Identifying the pivotal features among the numerous variables can significantly streamline the medical diagnosis process and reduce potential losses. For instance, after an initial diagnosis of a malignant tumor, a rapid feature reduction test can be employed to enhance the accuracy of the initial diagnosis. Simultaneously, gaining insights into the primary characteristics of cancer malignancy diagnosis can have a favorable impact on prognosis and subsequent detection procedures.

3. Overall Success and Effectiveness

In summary, I will conclude that the analysis was a success. By examining the results displayed from correlation matrices and important features selected by the random forest model, we can confirm that certain features, such as tumor size, concave points, and perimeter may have played a prominent role in determining whether a tumor is benign or malignant in the breast cancer database.

G. Summarize Of Key Takeaway from Analysis

1. Conclusions from Analysis

In conclusion, by examining the results displayed from correlation matrices and machine learning model, we can confirm that certain features, such as tumor size, concave points, and perimeter may have played a prominent role in determining whether a tumor is benign or malignant in the breast cancer database.

2. Explain Why Chosen Tools and Graphical Representation for Storytelling

I employed a variety of visualization techniques in this project. First, I created a correlation matrix heatmap, which provided an internal insight into the relationships between features and the target variable. Additionally, a pie chart visually represented the distribution of malignant and benign tumors in the dataset. This allowed us to rapidly comprehend the proportion of tumor diagnoses. Lastly, I utilized side-by-side bar charts to highlight differences in model metrics, simplifying the identification of superior-performing models. Furthermore, within the side-by- side bar chart, I displayed the weights of the top 5 features selected by the tuned model, serving as the summary visualization of the answer to our hypothesis.

3. Two Courses of Action Based on Findings

As outlined in the project scope, to achieve a conclusive assessment of feature importance and validate the model's selected features, additional analysis is required after the initial hypothesis question is answer.

The first course of action would involve exploring similar datasets to enhance the accuracy of feature identification, specifically to determine the most important features for diagnosing malignant tumors.

The second course of action would entail collecting more recent data on the same topic to assess whether the predominant characteristics of breast cancer have changed over time.

H. Panopto Link for Recording of Project and Findings

Panopto Recording For [redacted]

I. Appendices

Project Code

The analysis has been completed on the Jupyter notebook, and an HTML format file named "***WDBC_ML_Project.html***" has been attached within the same folder.

Other Files

All files that contain in same folder are:

a. Task3.pdf
b. WDBC_ML_Project.html
c. WDBC_ML_Project.ipynb
d. Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository.pdf
e. Wdb_data folder with original data

Sources of Data

Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B