Machine Learning in Breast Cancer Diagnostic Dataset

[redacted]

Western Governors University

A. Project Overview

1. 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.

2. Context and Background

Breast cancer is one of the serious and potentially life-threatening diseases. Advancements in technology have spurred comprehensive research and studies to gain deeper insights into this disease and find innovative ways to combat it. In this project, I will leverage recently acquired machine learning techniques to analyze historical breast cancer data. I will utilize features extracted from digitized images of fine needle aspirates (FNAs) of breast masses to identify prevalent characteristics associated with malignant tumors. Subsequently, by assessing the precision values associated with features in the dataset, I aim to determine if specific tumor features significantly influence the classification of tumors as benign or malignant. This analysis will contribute to my comprehension of the field of machine learning while allowing me to identify areas where I can improve my skills by contrasting the disparities between the present and past data analysis methodologies.

3. Three Published Works of Relate Research

a. How Published Work Informs the Development of Project

In this project, several published works have served as general guidelines, including:

1. "**Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison**" (Published by Pfob, A., Lu, SC. & Sidey-Gibbons in 2022): This recent publication will provide valuable insights into techniques for data preprocessing, hyperparameter tuning, and model comparison in the context of machine learning in medicine. It provides me with an overall process of how to conduct statistical tests to compare different models, allowing me to understand whether differences in model performance are in fact statistically significant.

2. "**Machine learning in medicine: a practical introduction**" (Published by Sidey-Gibbons, J., Sidey-Gibbons, C. in 2019): The authors provide a practical and accessible introduction to the utilization of machine learning in the field of medicine. The paper includes detailed examples demonstrating the implementation of Logistic Regression, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs). These examples effectively illustrate the theory and practical application of machine

learning, making the content accessible and beneficial for clinicians and medical researchers.

3. "**Machine Learning Approaches for Cancer Detection**" (Published by Ayush Sharma, Sudhanshu Kulshrestha, Sibi B Daniel in 2018): This paper concentrates on the application of machine learning techniques to the detection of cancer. It provides valuable insights into how various metrics derived from machine learning models can influence decision-making within the medical domain. Moreover, it demonstrates a deep understanding of how a single model can exhibit varying behaviors when confronted with different datasets.

The knowledge and methodologies presented in these papers helped me establish a solid foundation for the project and provided valuable guidance on how to approach my hypothesis.

4. Summary Data Analytics Solution

The objective of this project is to employ a machine learning model to identify the most influential features in the dataset about distinguishing between benign and malignant tumors. Furthermore, I will conduct a t-test by examining the precision values between the primary features and secondary features, thereby determining whether there exists a statistically significant difference between the means of these two groups.

Utilizing the publicly available dataset known as WDBC, the "Breast Cancer Wisconsin (Diagnostic) dataset". I anticipate obtaining two key outputs from this analysis:
The first output will consist of the five most influential features identified by the trained and turned machine learning model. These features are essential for distinguishing between benign and malignant tumors.

The second output will be a p-value obtained through a t-test. This p-value will determine the statistical significance of the results obtained in the first output. It will indicate whether the difference in precision scores between the selected top features and the remaining features is statistically meaningful. To verify my initial hypothesis.

5. Benefit And Support of Decision-Making Process.

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.

B. Data Analytics Project Plan

1. Goals, Objectives, and Deliverables

Goal: Identify whether specific features hold greater significance than others in the determination of tumor malignancy.

- Objective A: Identify and select the top 5 features from the dataset using a fine-tuned machine learning model. This objective focuses on pinpointing the most influential features by simply looking at the exiting data.

  Deliverables:

Items	Time
Download data from UCI website	Sep 19,2023
Cleaned and processed data	Sep 20,2023
Selected Models after training and prediction	Sep 21,2023
Selected fine-turn model	Sep 25,2023
Top 5 features selected by model	Sep 27,2023

- Objective B: Utilize a statistical test (t-test) to obtain p-values. These p-values serve as a crucial metric to assess the statistical significance of the findings from Objective A. It helps us determine whether the observed differences in feature importance are statistically meaningful and can be used to answer our underlying hypothesis.

  Deliverables:

Items	Time
2 groups of data split by selected top 5 features	Sep 28,2023
Lists contain precision scores from groups	Sep 29,2023
p_value produced by t-test	Sep 29,2023
Summary of result	Oct 2,2023

2. 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.

3. Project Planning Methodology

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

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

- **Design**
  1. Data cleaning and processing.
  2. Data splitting, features scaling to ensure data equality and performance.
  3. Models training and verification.
  4. Model turning.
  5. Features select based on weight.
  6. Statistics test with divide group by features for significant level.

- **Implementation**
  1. Implement code for design step 1 to 5.

- **Evaluation:**
  1. Testing model with testing data to ensure accuracy.
  2. Testing metric scores to ensure output significant level.

4. 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

Trained and tested models	Sep 21,2023	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
Groups of data for statistic test	Sep 28,2023	Sep 28,2023	1 day
Precision scores for statistic test	Sep 29,2023	Sep 29,2023	1 day
p_value from t-test	Sep 29,2023	Sep 29,2023	1 day
Result summary from analysis	Oct 2, 2023	Oct 2, 2023	1 day

5. List of Resources and Associated Costs

Item	Cost
Breast Cancer Diagnostic Dataset	Free
Jupyter Notebook	Free
Python Libraries	Free

6. Criteria of Success

Recognizing the widespread popularity of the WDBC dataset in the field of machine learning, the original contributor has generously shared their preliminary findings regarding the performance of various machine learning models. These findings will serve as a valuable reference point for me as I embark on my own exploration of suitable models and expected outcomes.

In the initial phase of my project, I aim to meticulously evaluate multiple machine learning models using the WDBC dataset. My objective is to proficiently discern the most suitable model based on a comprehensive comparison of their respective performances. The visual representation of these models' performance metrics should closely resemble the results available on the UCI dataset repository, maintaining consistency with the original data presentation.

In the project's final analysis, the significance of various features will be established through the p-values generated by a t-test. By examining the relationship between these p-values and the alpha threshold, we will ascertain whether the features identified during the model fine-tuning process holds greater importance compared to others in the dataset.

C. Design of Data Analytics Solution

1. Hypothesis

Are some of the features in the dataset more important than others in predicting the malignancy of tumors?

Null Hypothesis (H0): All features in the breast cancer dataset are equally important in predicting breast cancer.

Alternative Hypothesis (H1): Certain features in the breast cancer dataset are more important than others in predicting breast cancer.

2. Analytical Method

a. Justify of Analytical Method

With the dataset being preprocessed and avoided of missing data, first I will perform some feature scaling operations to ensure data are concise and within the same standard. Then I can systematically explore the data using various Supervised Learning Models such as decision trees, random forest, and logistic regression. I aim to determine the model that best suits the data based on criteria such as execution time, accuracy, precision and F1 score. Subsequently, I will employ grid search techniques to optimize the model's hyperparameters. Furthermore, to enhance the understanding of the dataset and its diagnostic potential, I will identify the top features crucial for breast cancer diagnosis by examining feature importance attributes.

Once the top 5 features have been identified, I will proceed with conducting a t-test. This test aims to yield a p-value as part of my final analysis, assessing whether a statistically significant difference exists in the precision or accuracy scores between the dataset comprising the top 5 features and the dataset containing the remaining features.

3. Tools and Environments

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

4. Methods and Metrics of Statistical Significance

a. Justify Methods and Metrics

<u>Model:</u>

Model: LogisticRegression, RandomForestClassifier, DecisionTreeClassifier or more

Metric: fbeta_score, accuracy_score, precision_score

Benchmark: Which ever model obtain the highest precision and accuracy scores with less execute time. Output should be similar with result on UCI (Baseline Model Performance).

While the database is relatively small in terms of the number of instances, it contains a substantial number of features, with each instance having 30 of them. Feature selection techniques, as well as machine learning methods like tree-based models, can efficiently and rapidly label and group them, and pinpoint the most influential features within this dataset. By comparing multiple models, we can intuitively identify the module that is most suitable for handling this type of data. The implementation of grid search can help us look for the optimal combination of parameters for the given machine learning algorithm, to improve the overall performance of the model.

Statistical test:

Null hypothesis: All features in the breast cancer dataset are equally important in predicting breast cancer.

Plan: Calculate the precision or accuracy score after selecting top 5 features by model from dataset, comparing lists of scores between top 5 features and other features to see if the difference is significant.

Metric: p-value from t-test

Alpha: 0.05

A t-test is a statistical hypothesis test used to determine if there is a significant difference between the means of two groups or samples. It is particularly useful in this situation when I want to compare the means of two groups (top feature and others) and determine if the differences we observe are likely to be due to chance or if they are statistically significant.

5. Practical Significance of the Data Analytics Solution

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.

6. Tools and Graphical Representations

I will employ a variety of visualization techniques in this project. First, I will create a correlation matrix heatmap, offering an initial insight into the relationships between features and the target variable. Additionally, a pie chart will visually represent the distribution of malignant and benign tumors in the dataset. Lastly, I will utilize side-by-side bar charts to highlight differences in model metrics, simplifying the identification of superior-performing models. Seaborn and Matplotlib will be the primary tools employed to craft and present the visualizations in this project.

D. Description of Datasets

1. Sources 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. The dataset includes 569 instances and 32 attributes. It was originally compiled and made publicly accessible by researchers for the purpose of advancing breast cancer research and machine learning applications in healthcare. The dataset is typically credited to the University of Wisconsin, where researchers like Dr. William H. Wolberg and his colleagues W. Nick Street and Olvi L. Mangasarian from the Computer Sciences Department were involved in its creation and distribution.

2. Appropriate of Dataset

Features in WDBC dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

“Features are:

1) diagnosis (M = malignant, B = benign)
2) radius (mean of distances from center to points on the perimeter)
3) texture (standard deviation of gray-scale values)
4) perimeter
5) area
6) smoothness (local variation in radius lengths)
7) compactness (perimeter^2 / area - 1.0)

8) concavity (severity of concave portions of the contour)
9) concave points (number of concave portions of the contour)
10) symmetry
11) fractal dimension ("coastline approximation" - 1)"

(Wolberg et al., 1995)

Note that above features information was provided by the data's original contributor. This dataset is appropriate for our project because it collectively provides a comprehensive set of characteristics that can help healthcare professionals and machine learning algorithms differentiate between benign and malignant breast tumors. By analyzing these features, it becomes possible to identify patterns associated with cancer and aid in the early and accurate diagnosis of breast cancer, which is crucial for effective treatment and patient outcomes.

3. Data Collection Methods

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

4. Observations on Data Quality and Completeness

The dataset has already been cleaned and processed by the original contributor, and it is free from missing values. The operations I will perform are dropping the ID column and splitting them into subgroups when necessary.

5. Data Governance, Privacy, Security, Ethical, Legal, And Regulatory Compliance

a. Precautions

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. It was originally compiled and made publicly accessible by researchers for the purpose of advancing breast cancer research and machine learning applications in healthcare. It is a publicly available dataset with no restrictions on its use for research and educational purposes. 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.

E. Sources and Reference

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

Pfob, A., Lu, S.-C., & Sidey-Gibbons, C. (2022). Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01758-8

A, J., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4

Sharma, A., Kulshrestha, S., & B Daniel, S. (2018). Machine Learning Approaches for Cancer Detection. *International Journal of Engineering and Manufacturing*, *8*(2), 45–55. https://doi.org/10.5815/ijem.2018.02.05