Data Management and Data Analytics Capstone Topic Approval Form
Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Student Name: [redacted]

Student ID: [redacted]

Capstone Project Name: Machine Learning of Breast Cancer Diagnostic

Project Topic: Discover Main Features of Malignant Tumors of Breast Cancer through Machine Learning

Research Question: Utilize machine learning techniques to conduct diagnostic assessments using pre-existing breast cancer datasets and discern the predominant shared features of malignant tumors.

Hypothesis:
Are some the features in 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.

Context:
Breast cancer is one of the serious and potentially life-threatening disease. Advancements in technology have spurred comprehensive research and studies aimed at gaining deeper insights into this disease and finding innovative ways to combat it. In this project, I will leverage recently acquired machine learning techniques to conduct an in-depth analysis of 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. 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.

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. However, it is important to note that 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 our understanding of breast cancer and develop diagnostic and predictive models without the need for specific permissions or ownership considerations.

> "
> Features in 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) ID number
> 2) Diagnosis (M = malignant, B = benign)
>
> Ten real-valued features are computed for each cell nucleus:
> a) radius (mean of distances from center to points on the perimeter)
> b) texture (standard deviation of gray-scale values)
> c) perimeter
> d) area
> e) smoothness (local variation in radius lengths)
> f) compactness (perimeter^2 / area - 1.0)
> g) concavity (severity of concave portions of the contour)
> h) concave points (number of concave portions of the contour)
> i) symmetry
> j) fractal dimension ("coastline approximation" - 1)
> "

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

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

Data Analytics Tools and Techniques:
With the dataset being preprocessed and devoid of missing data, first I can 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, and F1 score. Subsequently, I can employ grid search techniques to optimize the model's hyperparameters. Furthermore, to enhance the understanding of the dataset and its diagnostic potential, I can identify the top features crucial for breast cancer diagnosis by examining feature importance attributes.

Justification of Tools/Techniques:
The tools and techniques I will be using will enable me to develop an effective and efficient breast cancer diagnostic with a well-trained machine-learning model. 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.

Application Type, if applicable (select one):

- ☐ Mobile
- ☒ Web
- ☐ Stand-alone

Programming/Development Language(s), if applicable: Python

Operating System(s)/Platform(s), if applicable: N/A

Database Management System, if applicable: N/A

Project Outcomes: Choose the optimal machine learning model for assessing breast cancer and uncover the top key features for distinguishing malignant tumors.

Projected Project End Date: [redacted]2023

Sources: Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

---

Human Subjects or Proprietary Information

Does your project involve the potential use of human subjects? (Y/N): N

Does your project involve the potential use of proprietary company information? (Y/N): N

STUDENT SIGNATURE

__________[redacted] ____________________

By signing and submitting this form, you acknowledge that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

TO BE FILLED BY A COURSE INSTRUCTOR

The capstone topic is approved by a course instructor.

COURSE INSTRUCTOR SIGNATURE:

James R Ashe

Jim Ashe, Ph.D. Mathematics

COURSE INSTRUCTOR APPROVAL DATE:

[redacted]/23

Project Compliance with IRB (Y/N): Y