Northeastern University

Project Proposal

Using Machine Learning Algorithms and Kaggle Dataset to Detect Breast Cancer

Professor: Dr. Karl Ni

Author: Jie Zhang

Course: CS 6220 Data Mining

Introduction

Breast cancer is a significant public health concern, affecting millions of women worldwide. Early detection of breast cancer is critical for successful treatment and improving the chances of survival. With the increase in the availability of breast cancer data, machine learning techniques are becoming increasingly relevant in improving the accuracy of detection. In this project, we aim to use machine learning algorithms to analyze the breast cancer dataset from Kaggle and develop a model that can accurately detect breast cancer in the early step.

Project Objective

- Analyze the Kaggle breast cancer dataset to extract meaningful features
- Develop a machine learning model that can accurately detect breast cancer
- Evaluate the performance of the model using various performance metrics, such as accuracy.
- Compare the performance of the proposed model with existing state-of-the-art methods

Methodology

1. Data Collection: We will use the publicly available breast cancer dataset from Kaggle, which dataset size is 5GB and .cvs type.

   During this part, it is hard to choose a proper dataset from Kaggle, the main reason is that the quality of the dataset directly affects the accuracy and reliability of the models built using it. After searching and taking a look at all the dataset related to "Breast Cancer" I found there are two majority dataset types in Kaggle: images, regular dataset. For this project , the better way is to use both of them and explore machine learning methods.

   However, after read some professional papers, I choose a part of data set from Kaggler which focus on the three main techniques in the project:

   1)Improve model performance and evaluate different Model (use k-fold validation)

2)Use HeatMap to improve the data : use this method to let the data more clearly show the correlation vs coefficients.

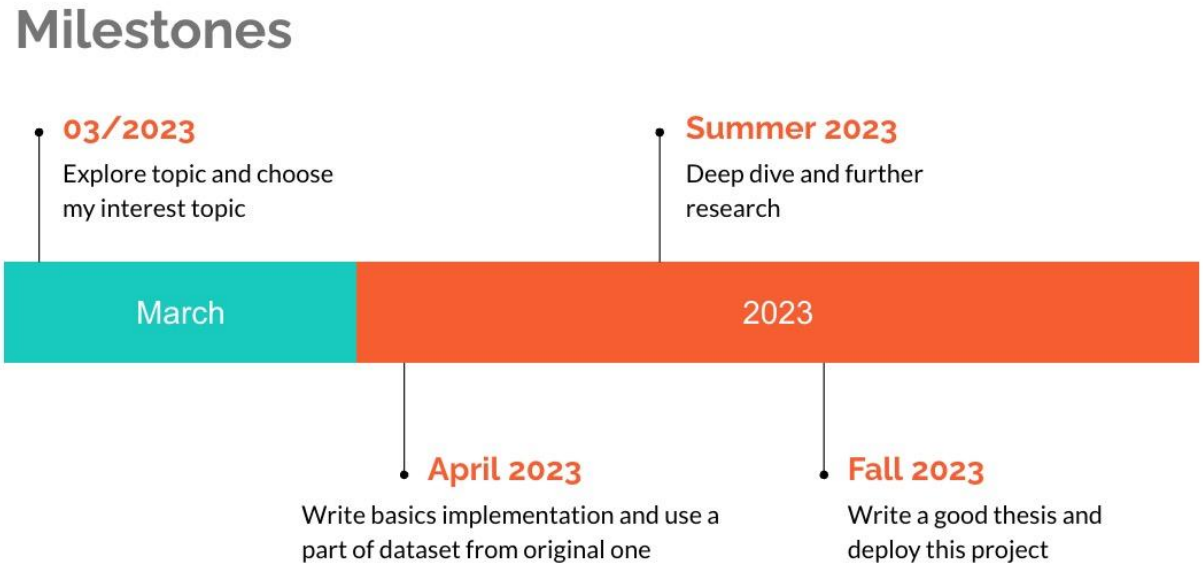
3) Do some data look around and find the more : missing value and null value, map the column with 0 or 1.

Data Preprocessing: Preprocess the dataset to handle missing values or null values. Choose some features and select them to identify a proper model for detection.

2. Model Selection: Evaluate the performance of various machine learning algorithms,expencailly k-fold validation and heatmap to find the associate relationship for the selected features.

3. Model Tuning: Perform hyperparameter tuning on the selected machine learning algorithm to improve its performance.
4. Model Evaluation: Evaluate the performance of the developed model using various performance metrics, such as accuracy, precision, recall, and compare it with existing state-of-the-art methods.

Project Milestone:



Milestones

03/2023
Explore topic and choose my interest topic

Summer 2023
Deep dive and further research

April 2023
Write basics implementation and use a part of dataset from original one

Fall 2023
Write a good thesis and deploy this project

Outcome:

Part1:

New Knowledges and Techniques:

1)Heatmap:

Display the relationship between different values in a matrix or table of data, Color encoding ,Correlation vs coefficients between two variables, Displaying the correlation between different features in a multidimensional dataset.

2)K-Fold Cross Validation:

Detect whether the model is overfitting

Improve the accuracy of evaluation.

3) How to handle the data type issue:

Missing value

After map to binary value 1 or 0, I need to consider the type.

Null value check.

Data explore and play.

Part2:

Gain more interest to research part:

Use the NEU library to read more intensive topics and research papers.

Use google scholar to find some related articles.

Use online resources to understand some concepts and algorithms.

Refresh the lecture and lab code, implementation, environment setting techniques.

Reference:

1.https://onesearch.library.northeastern.edu/discovery/fulldisplay?docid=cdi_crossref_primary_10_3991_ijoe_v18i05_29197&context=PC&vid=01NEU_INST:NU&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,breast%20cancer%20machine%20learning&offset=10

2.https://www.kaggle.com/code/malik12345/breast-cancer-detection-using-ml/notebook

3.https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset

4.https://www.kaggle.com/code/usakshaya/breast-cancer-prediction

5.https://machinelearningmastery.com/k-fold-cross-validation/

6.https://en.wikipedia.org/wiki/Histogram

7.https://towardsdatascience.com/different-ways-to-connect-google-drive-to-a-google-colab-notebook-pt-1-de03433d2f7a

8.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/

9.https://www.freecodecamp.org/news/how-to-handle-missing-data-in-a-dataset/

10..https://www.v7labs.com/blog/overfitting

11.https://www.youtube.com/watch?v=qxpKCBV60U4

12.https://scikit-learn.org/stable/modules/cross_validation.html

13.https://stackoverflow.com/questions/73443407/svm-problem-name-model-svc-is-not-defined

14.https://www.youtube.com/watch?v=tLhunN5Jhqs

15.https://www.youtube.com/watch?v=Hlk5psu5yFw

Conclusion

Breast cancer is a significant public health concern, and early detection is crucial for successful treatment. In this project, we aim to use machine learning algorithms to analyze the breast cancer dataset from Kaggle and develop a model that can accurately diagnose breast cancer. We believe that the project's findings will contribute to improving breast cancer diagnosis and ultimately lead to better treatment outcomes for breast cancer patients.