Northeastern University

Project Proposal

Breast Cancer Diagnosis: Using Machine Learning Algorithms and Kaggle Dataset to improve the accurately diagnose 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 diagnosis. 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.

Project Objective

- Analyze the Kaggle breast cancer dataset to extract meaningful features
- Develop a machine learning model that can accurately diagnose breast cancer
- Evaluate the performance of the model using various performance metrics, such as accuracy, precision, recall, and F1 score.
- 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.
2. Data Preprocessing: We will preprocess the dataset to handle missing values, outliers, and other anomalies. We will also perform feature selection and extraction to identify the most relevant features for diagnosis.
3. Model Selection: We will evaluate the performance of various machine learning algorithms, such as logistic regression, decision trees, K-means and support vector machines, to identify the best model that can accurately diagnose breast cancer.

4. Model Tuning: We will perform hyperparameter tuning on the selected machine learning algorithm to improve its performance.
5. Model Evaluation: We will evaluate the performance of the developed model using various performance metrics, such as accuracy, precision, recall, and F1 score, and compare it with existing state-of-the-art methods.

Expected Outcome

We expect to develop a machine learning model that can accurately diagnose breast cancer with high precision and recall. We also anticipate that the proposed model will outperform existing state-of-the-art methods in terms of accuracy and F1 score. The project's findings will contribute to improving breast cancer diagnosis and ultimately lead to better treatment outcomes for breast cancer patients.

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.