Project Proposal

"Breast Cancer Diagnosis Prediction"

Group 03 – Team members' detail

- Yen Nga Le
- Tehsin Shaikh
- Srilakshmi Gummadidala

Introduction

We aim to develop a predictive model for breast cancer diagnosis using machine learning techniques applied to a dataset containing clinical cases reported by Dr. Wolberg. Our goal is to create a model that can accurately classify breast tumors as benign or malignant based on various attributes such as clump thickness, cell size uniformity, and mitoses. By leveraging the power of machine learning, we seek to enhance the accuracy and efficiency of breast cancer diagnosis, ultimately contributing to improved patient outcomes and treatment strategies.

Related Work

Previous researches in the field of breast cancer diagnosis have explored various machine learning approaches, including decision trees, support vector machines, and neural networks. These studies have demonstrated promising results in accurately predicting breast cancer outcomes based on clinical and histopathological data. Our project builds upon this existing body of work by focusing on developing a predictive model using readily available clinical data, with the aim of making accurate diagnoses accessible to a wider range of healthcare settings.

Relevance with the Course

This project aligns closely with the concepts covered in our course on machine learning. Specifically, it applies techniques learned throughout the course, including data preprocessing, feature selection, model training, and evaluation. By working on this real-world problem of breast cancer diagnosis, we have the opportunity to apply the theoretical knowledge gained in the course to practical applications in healthcare.

Potential challenges

The primary challenge encountered in this project stems from the absence of demographic, geographic, and lifestyle information within the dataset. The limited scope restricts the depth of analysis regarding the potential influence of these factors on breast cancer characteristics.

Demographic Information: Characteristics like age, gender, race, and socioeconomic status offer insight into population composition, aiding in analyzing health outcomes by demographic groups.

Geographic Information: Spatial details such as location, neighborhood, or city provide context to health data, revealing patterns and disparities in disease prevalence and outcomes.

Lifestyle Information: Behaviors like diet, exercise, smoking, and sleep patterns impact health outcomes, aiding in identifying risk or protective factors for specific conditions.

Temporal Shifts: Findings from a dataset collected in 1992 may not be generalizable to diverse populations or specific demographic groups today. The applicability of the dataset's findings may be limited, particularly if demographic or cultural shifts have occurred over time.

Motivation

Our project holds significant societal importance due to the widespread impact of breast cancer on individuals and communities worldwide. Breast cancer is one of the most prevalent forms of cancer, affecting millions of people each year. Early detection is crucial for successful treatment outcomes, and accurate diagnosis plays a vital role in initiating timely interventions. Our project has the potential to make a tangible difference in the lives of patients by improving the accuracy and efficiency of diagnosis. Timely identification of malignant tumors can lead to earlier treatment initiation, which may ultimately improve survival rates and quality of life for individuals affected by breast cancer.

From a personal learning perspective, this project provides an opportunity to apply and reinforce machine learning concepts learned throughout our coursework. It allows us to gain practical experience in data analysis, model development, and evaluation within the context of a real-world healthcare problem. The project's dual significance in addressing a critical health issue and advancing our learning journey in machine learning makes it both compelling and rewarding.

Evaluation

Successful Outcome

We will consider whether the project is successful or not if we can:

- Develop a predictive model for breast cancer diagnosis with high accuracy in distinguishing between benign and malignant tumors.
- Provide actionable insights into the factors contributing to breast cancer diagnoses, such as tumor characteristics.
- Communicate the model's findings effectively through data visualizations and reports, facilitating informed decision-making in clinical settings.

Measurement of Success

Success will be measured by:

- Accuracy and reliability of the predictive model in classifying breast tumors as benign or malignant.
- The depth and quality of insights generated regarding the factors influencing breast cancer diagnoses.
- The practicality and effectiveness of recommendations in guiding clinical practice and improving patient outcomes.
- Assuring ethical practices in Data Analytics throughout every stage of this project due to the high sensitivity of healthcare datasets.

Resources

Breast Cancer Dataset: Our primary data source will be the breast cancer dataset provided by Dr. Wolberg, comprising clinical cases and tumor characteristics. This dataset serves as the foundation for our predictive modeling and analysis:

> https://data.world/uci/breast-cancer-wisconsin-original/workspace/project-summary?agentid=uci&datasetid=breast-cancer-wisconsin-original

Data Analysis Tools: We will leverage Python programming language along with libraries such as Pandas, NumPy, and Scikit-learn for data preprocessing, model development, and evaluation. Additionally, visualization tools like Tableau, Power Bi, Matplotlib and Seaborn will be utilized for data visualization tasks. Excel may also be employed for basic data manipulation and analysis.

Relevant Course Materials: Our project will draw upon the knowledge and concepts covered in our machine learning and data analysis courses. Course lectures, textbooks, and supplementary materials on machine learning algorithms, data preprocessing techniques, and model evaluation will provide the theoretical framework and guidance for our analysis.

Contribution

Initially, every member focused on identifying relevant healthcare projects aligned with our primary objectives. Subsequently, we collaboratively deliberated and selected the most pertinent dataset for our analysis. While individual responsibilities may vary, everyone will work together to achieve the project goals. Specifically:

Member 1: Responsible for data preprocessing tasks, including handling missing values, feature engineering, and dataset preparation. Additionally, will assist in model development and evaluation stages, contributing insights and expertise to enhance the overall analysis.

Member 2: Will focus on model development and optimization, experimenting with various machine learning algorithms and techniques to create an accurate predictive model for breast cancer diagnosis. Additionally, will collaborate with other team members to interpret model results and refine the analysis.

Member 3: Tasked with data visualization and communication efforts, creating informative and visually engaging representations of the project findings. Will work closely with other team members to translate complex analytical results into clear and actionable insights for stakeholders.

References

1) O. L. Mangasarian and W. H. Wolberg (1990). Cancer diagnosis via linear programming. *SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18*.

2) William H. Wolberg and O.L. Mangasarian (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196*.

3) O. L. Mangasarian, R. Setiono, and W.H. Wolberg (1990). Pattern recognition via linear programming: Theory and application to medical diagnosis in: "Large-scale numerical optimization". *Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30*.

4) K. P. Bennett & O. L. Mangasarian (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers)*.