**[Classification] Breast Cancer Wisconsin (Diagnostic) Data set**

|  |  |
| --- | --- |
| **Dr. Katherine Shoemaker**  Project Director, DOED Research 2021  Ph.D Mentor Research  Department of Science and Technology  University of Houston – Downtown  Houston, TX  [shoemakerk@uhd.edu](mailto:shoemakerk@uhd.edu) | **Dr. Mary Jo Parker**  Project PI, DOED Research 2021  Executive Director, Scholar Academy  Department of Natural Sciences  College of Science and Technology  University of Houston – Downtown  Houston, TX  [Parkerm@uhd.edu](mailto:Parkerm@uhd.edu) |
| **Duc Thanh La**  Project Author, DOED Research 2021  Student Assistant, Scholar Academy  College of Science and Technology  University of Houston – Downtown  Houston, TX  [Lad1@gator.uhd.edu](mailto:Lad1@gator.uhd.edu) | **Son Luong**  Project Author, DOED Research 2021  Student Assistant, Scholar Academy  College of Science and Technology  University of Houston – Downtown  Houston, TX  [luongs3@gator.uhd.edu](mailto:luongs3@gator.uhd.edu) |

**Abstract**

Breast cancer is the most common cancer, accounting for approximately 33% of all cancers diagnosed among women in Wisconsin. It has been 40 years since the first case was collected by the Wisconsin Cancer Report system (WCRS). Early detection plays an essential role in breast cancer treatment by boosting the chance for timely intervention. Moreover, the increase in the incidence of this cancer shows no sign of stopping, which is why this topic has received much attention from researchers, including expert systems and machine learning techniques. Therefore, this study aims to predict and detect breast cancer early with non-invasive and technology-involved methods by using one of the most popular machine learning algorithms, which is stepwise logistic regression. Using stepwise regression will help the system remove redundant values and retain valuable values in the data, thereby using the logistic regression system to calculate relevant predictions such as accuracy, AIC, the probability. The study works on the Wisconsin Breast Cancer dataset taken from UCI Machine learning Repository. The outcome of this research shows that the stepwise logistic regression model achieved a classification accuracy of nearly 98%. Furthermore, this research will continue to be developed in terms of accuracy to detect breast cancer for early intervention.

Project category: Data Science - Undergraduate research & DOED MSEIP

Keywords: Breast Cancer, Classification, Machine Learning, Data Science, Logistic Regression, Stepwise Regression, Data Analysis, Data Visualization

References

American Cancer Society. (2000). Breast cancer facts and figures 2000. *Cancer Practice*, *8*(2), 55–57. https://doi.org/10.1046/j.1523-5394.2000.82001.x

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.