***"Enhancing Breast Cancer Diagnosis: CytoScan - Advanced Nuclei-Based Analysis of FNA Images Empowered by Predictive Modeling"***

Final Team 2 Project Technical Report

University of San Diego

ADS 503 Data Science Cloud Computing

Team 5: Trevor Sauerbrey, Mackenzie Carter, Justin Farnan

**Dataset:** [**https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic**](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)

**Github Repository:** [**https://github.com/trevsauer/ADS\_503\_team\_2\_final\_project.git**](https://github.com/trevsauer/ADS_503_team_2_final_project.git)

**Abstract:**

This project aims to accurately detect and diagnose breast cancer by utilizing predictive modeling techniques applied to digitized images of fine needle aspirates (FNA) of breast masses. The FNA images are used to compute various features that describe the characteristics of cell nuclei present in the images. These features include radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1). The dataset also includes an ID number and a diagnosis (malignant or benign) for each sample. By leveraging these features and employing predictive modeling techniques, the project seeks to achieve accurate breast cancer detection and diagnosis.

**Problem Statement**

*(justification for the proposed approach)*

This study aims to improve breast cancer detection and diagnosis by utilizing predictive modeling techniques applied to a large-scale dataset of fine needle aspirate (FNA) images. By analyzing computed features from the FNA images, the models enhance diagnostic accuracy and differentiate between benign and malignant masses more effectively. Leveraging the non-invasive nature of FNA, the approach offers a safer and less burdensome alternative to invasive procedures like biopsies. Furthermore, the objective and standardized assessment provided by computational algorithms minimizes subjectivity and potential biases in manual assessment. The potential for early detection enables timely interventions, while the models' versatility allows for future applications beyond breast cancer. Through these advantages, the project contributes to improved patient care and outcomes in breast cancer detection and diagnosis.

**EDA (Exploratory Data Analysis)**

*(graphical and non-graphical representations of relationships between the response variable and predictor variables)*

**Data wrangling and pre-processing**

*(handling of missing values, outliers, correlated features, etc.)*

*Note: there are no missing values in the dataset*

Note: All feature values are recoded with four significant digits.

**Data splitting**

*(training, validation, and test sets)*

Note: Class distribution: 357 benign, 212 malignant

**Predictive Modeling Strategies**

*(describing main research questions and appropriate analytics methods)*

**Validation and testing**

*(model tuning and evaluation)*

**Results and final model selection**

*(performance measures, etc.)*

**Discussion and conclusions**

*(address the problem statement and suggestions/solutions could go beyond the scope of the course)*

**References**

**Appendix**

*(reproducible and clean R code)*