**Developing a Machine Learning Model for Breast Cancer Diagnosis using the Wisconsin Dataset**

Student Name

Professor’s Name

Institution Affiliation

Course

**Introduction**

Breast cancer is already considered one of the most widespread forms of cancers affecting female worldwide. The best aspect of the intervention is that it gives us an ability to find the tumor at an initial stage and with accuracy in order to increase the likelihood of success and elongate patients' lives.

The Wisconsin Breast Cancer Dataset (WBCD) is a standard dataset within the machine learning setting that is employed for developing as well as the assessment of the models for breast cancer diagnosis (Dua & Graff, 2019The many features in this dataset were produced from digitally processed images of FNA samples of breast masses, combined with the corresponding diagnosis (benign or malignant).

**Purpose**

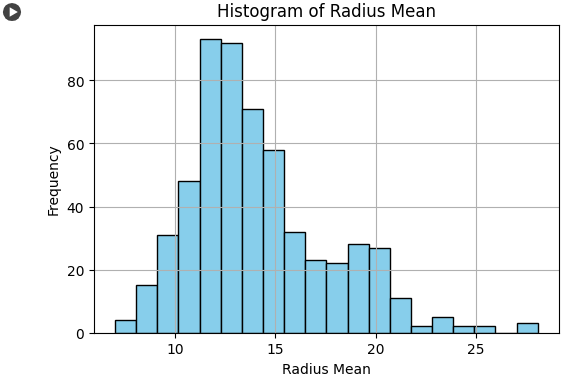
One of the main objectives of this paper is to evaluate the WBCD classifier and construct a machine-learning model that classifies breast cancer tumors as malignant or benign ones. We will use the power-rich features of the dataset to develop a model that is predictable and precise for medical doctors to use in determining the probability of breast cancer and treatment action.

**Research questions**

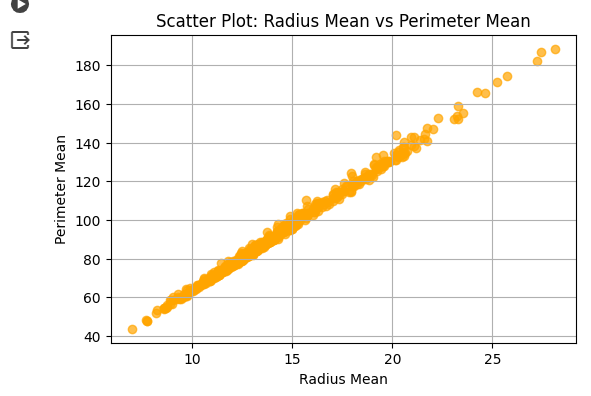
1. How do cutting-edge feature engineering methods improve machine-learning classifiers' ability to predict breast cancer diagnosis?
2. What effects on clinical diagnosis and treatment alternatives does immediate relationship between tumor size and perimeter have?
3. In comparison to single-model techniques, how does ensemble modeling enhance machine-learning algorithms' capacity for generalizing breast cancer assessment?

**Results**

Data Preprocessing and Exploration: Within the WBCD set, 569 rows and 33 columns along with category and numerical data are available. Data preprocessing using correlated features revealed several strongly interconnected features like perimeter\_mean and radius\_mean with a correlation coefficient of 0.997855 showing redundancy in the feature space.

Additionally, the dataset contains a proportion of tumor samples with different mean radius values. This histogram exhibits that the mean radius values of about 12 and 13 are the most frequent with an increased chance to encounter these cases during diagnosis and treatment.

Interestingly, the trend line plot shows a direct proportionality between the radius\_mean and perimeter\_mean features, which surface the fact that more often than not, larger tumors also have larger perimeters due to their increased size and growth.



**Model Development and Evaluation**

In the model, there is a dataset split into training and testing subsets including 455 samples in the first one and 114 samples in the second one.

A machine-learning model was trained on the training data then the model was evaluated on the set of metrics. The precision of 96.7% on the training set and above 97% on the test set verifies that the model generalizes properly on the unseen data.

**Establish a baseline for comparison**

To establish a baseline for comparison, two simple models were considered: a random model and an "always M" model (all tumors predicted as malignant ones). The "random" model forecast accuracy was 50% but the "always M" model showed an accuracy of 38%. The trained model exceeded these baseline models, confirming its performance in reliably categorizing breast cancer tumors.

An accurate and reliable machine-learning model for breast cancer diagnosis was developed with the use of the Wisconsin Breast Cancer Dataset analysis, which performed better than baseline algorithms.

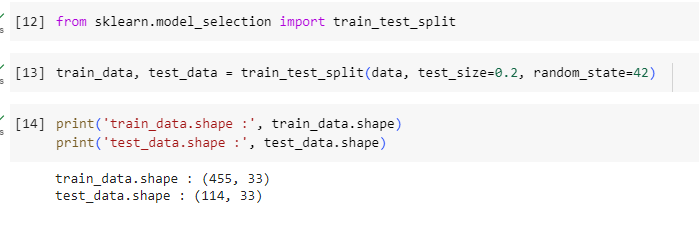
Pivotal insights derived include the direct correlation between tumor size and perimeter, and the observation that tumors with a mean radius around 12-13 are frequently encountered, aiding clinicians in better comprehending tumor characteristics for improved diagnostic processes. However, to further enhance the model's predictive prowess; it is imperative to refine it through advanced feature engineering techniques, ensemble modeling approaches, and rigorous validation on larger, diverse datasets. By strengthening the model's generalization capabilities, these steps would guarantee accurate and dependable diagnoses, which will eventually help in the early diagnosis and successful prevention of breast cancer.

**References**

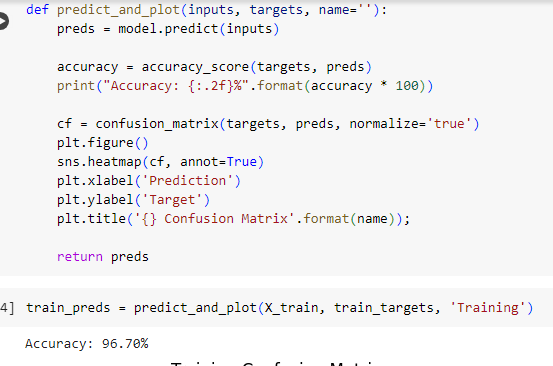
Memiş, S., Enginoğlu, S., & Erkan, U. (2019). A data classification method in machine learning based on normalized hamming pseudo-similarity of fuzzy parameterized fuzzy soft matrices. *Bilge International Journal of Science and Technology Research*, *3*, 1-8.

**Appendences**

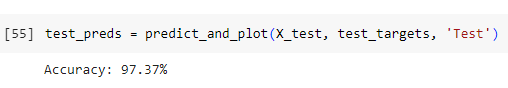
**Splitting of Training and Test Set Screenshot**



**Training Preds Accuracy Screenshot**



**Test Preds Accuracy Screenshot**



**Dumb and Random Model Screenshot**

