



Challenge 1 (Summary Report)

Name: **Anwai Archit**

Date of Submission: **19th March 2021**

Subject: **Machine and Deep Learning (Medical Imaging and Applications)**

Requisite Points of Concern:

1. Dataset – Breast Cancer Wisconsin (Diagnostic) – UCI Machine Learning Repository.
2. Using Scikit-Learn Library.
3. Training and Testing the Classifier Model and Studying the Features and Applications of:
 - a. Transformations
 - b. Dimensionality Reduction
 - c. Selection
4. Regular Accuracy of 3-NN Classifier – 93.01%
- Highest Accuracy Reached = **97.9%**

❖ *Summary of the Techniques Used:*

1. Standardization (*StandardScaler*) – scaling the data to obtain 97.2% accuracy.
2. Alternative Standardization: Scaling features to a range (*MinMaxScaler* and *MaxAbsScaler*) – scaling the data to obtain 97.2% accuracy.
3. Mapping to a Uniform Distribution (*QuantileTransformer*) – non-parametric transformation to obtain 94.41% accuracy.
4. Normalization (*Normalizer*) – normalizing the data to obtain 91.61% accuracy.
5. Linear Discriminant Analysis: LDA – 95.10% accuracy.
6. Removing Features with Low Variance (*VarianceThreshold*) – did not work.
7. Feature Selection using *SelectFromModel* (*LinearSVC* – 97.2% and *ExtraTreesClassifier* – 95.8%).

Model Presented: - Feature Transformation for Scaling the Data using *StandardScaler*, Linear Support Vector Classification using *SelectFromModel* and training a 3-NN Classifier and getting an accuracy of 97.9% on the test set.

Note: Sequential Feature Selection (sfs) theoretically looks appealing, but could not be imported due to version mismatch from `sklearn.feature_selection`.

*Silly Note: Changing the NN Classifier at instances gives us a decent performance (which is not allowed). Example: For instance, the *MinMaxScaler* with 5-NN Classifier gives a 98.6% accuracy.*