# CISM 623 – Practical Lab 2 Documentation (Breast Cancer Dataset)

## 1. Dataset Overview

• Dataset: Breast Cancer (Scikit-learn)  
• Shape: 569 rows × 32 columns  
• Features: 30 predictive attributes (all numerical)  
• Target Variable: 0 = malignant, 1 = benign  
• Added extra column 'diagnosis' for clarity  
  
**Conclusion:** The dataset is cleanly structured and ready for analysis.

## 2. Missing Values

• Results: No missing values were found in any column.  
• Conclusion: No preprocessing needed for missing data.  
  
This saves preprocessing hardwork by guaranteeing that no imputation was necessary.

## 3. Feature Types

• All 30 predictive features are numerical (float64).  
• 'target' is integer, and 'diagnosis' is categorical (object).  
  
**Conclusion:** Since all features are numerical, models that require numerical input (Logistic Regression, SVM, kNN, Random Forest) can be applied directly.

## 4. Class Distribution

• Benign: 357 (62.7%)  
• Malignant: 212 (37.3%)  
•Observation: The dataset is imbalanced, with more benign cases.  
  
**Conclusion:** I understood that while accuracy can still be a fair metric, precision/recall provide deeper insight for the malignant (minority) class.Accuracy alone may not be reliable.It is important to prioritize metrics like F1-score, precision, and recall.

## 5. Descriptive Statistics

• Features show different ranges and scales.

• Example:  
 - Mean radius: ~6–29  
 - Mean area: ~143–2501  
 - Mean smoothness: 0.05–0.16  
  
**Conclusion:** I understood that this scale difference would affect distance-based models (kNN) and gradient-based models (Logistic Regression, SVM).

Solution: apply StandardScaler.

• Standardization is necessary because of the wide variation, particularly for distance-based models like kNN.

## 6. Visualizations (EDA)

• Histograms/Boxplots: Showed skewness in some features (e.g., perimeter, area).  
• Correlation Heatmap:  
 - Strong correlations: mean radius, perimeter, and area.  
 - Multicollinearity Found → Tree-based models (Random Forest) can handle this better.  
  
**Conclusion:** I understood that this redundancy could negatively affect linear models. Tree-based models like Random Forest handle correlated features better.Feature correlation helps predict which models may perform better.

## 7. Model Training & Results

• Models Tested: Logistic Regression, kNN, Random Forest.  
• Accuracy Scores (approximately):  
 - Logistic Regression → ~93%  
 - kNN → ~91%  
 - Random Forest → ~96%  
  
**Conclusion:** Based on EDA, Random Forest is expected to perform strongly because:

• It handles non-linear decision boundaries (malignant/benign separation is not purely linear).

• It is robust against correlated features (feature bagging reduces reliance on redundant features).

• Linear models like Logistic Regression may perform decently but could be affected by feature redundancy.

• Thus, Random Forest scored better than the others because it captures complex non-linear decision boundaries and manages multicollinearity well.

## 8. Impact of Class Imbalance (HYPOTHETICAL)

If the dataset had been highly imbalanced, issues would include:

• Accuracy being misleading (predicting all 'benign' could give high accuracy).

• Precision/Recall trade-off with malignant class likely having lower recall (more false negatives).

• F1-score becoming more important than accuracy for fair evaluation.

• Solutions: Stratified sampling in splits, class weighting, or oversampling techniques.

## 9. Future Improvements

Future improvements can be made by Hyperparameter Tuning:

• Optimize Random Forest, Logistic Regression, and kNN using GridSearchCV.

• Improve accuracy and recall, especially for malignant cases.

**Conclusion:** Proper tuning may push Logistic Regression and kNN closer to Random Forest performance.

## Summary:

- Dataset: 569 samples × 32 columns, no missing values.  
- Target: benign (357) vs malignant (212) → moderate imbalance.  
- Features: different scales and strong correlations → scaling & handling needed.  
- Visualizations confirmed feature distributions and redundancy.  
- Improvements Needed: Applying ML models, evaluate with accuracy, precision, recall, F1-score, and tune hyperparameters.  
- Overall, I understood how to prepare, interpret, and analyze the Breast Cancer dataset before modeling.

**Final Conclusion:** Through EDA and model testing, Random Forest was found to be the most suitable model for the Breast Cancer dataset due to its robustness against correlated features and ability to handle complex decision boundaries. However, awareness of class imbalance and future hyperparameter tuning could further improve results.