S3 MCA (2024-26 Batch)

M24CA1L306 Data Science Lab

Lab Exercise #04

Date of Submission: 12 Sept 2025

- Download the dataset Breast_Cancer.csv
- Using Python, visualize the various characteristics of the dataset.
- Use Google Colab for implementation.
- Define your own conditions to learn the characteristics
- After each visualization write your findings
- EDA Questions

1. What is the distribution of malignant vs benign cases?
    - Use count plots.
2. Which features differ the most between malignant and benign tumors?
    - Compare boxplots of key features like radius_mean, texture_mean, area_mean.
3. Is there a strong correlation between radius, perimeter, and area?
    - Heatmap of correlation matrix.
4. Which features are most useful in distinguishing malignant from benign?
    - Use violin plots for selected features (e.g., concavity_mean, smoothness_mean).
5. How do "worst" features compare to "mean" features in classification power?
    - Compare distributions of mean vs worst for malignant/benign.
6. Are there multicollinearity issues among features?
    - Inspect correlation heatmap for redundancy.
7. Which top 5 features show the largest separation between malignant and benign tumors?
    - Use feature importance pairplots.
8. What is the average tumor size (radius/area) for malignant vs benign?
    - Use grouped bar plots or descriptive stats.
9. Are there any outliers in the dataset? How might they affect classification?
    - Boxplots or scatter plots for area and radius.
11. Which features show the highest variance, and do they contribute to classification?
    - Plot variance of each feature and compare across classes.
12. How does texture differ between malignant and benign tumors?
    - Use histograms or KDE plots for texture_mean and texture_worst.
13. Do tumors with higher concavity tend to be malignant?
    - Compare distribution of concavity_mean for both classes.
14. What is the relationship between smoothness and compactness?
    - Scatter plot colored by diagnosis.
15. Which class tends to have higher values of fractal dimension?
    - Use violin plots for fractal_dimension_mean.