**Data Analytics & Visualization (CS/IT 312)**

**Mini Project Submission & Feedback Form**

**Student Information**

* **Student Name:** Rajat Kumar Thakur
* **Student Roll No.:** 202211070
* **Mini Project Title:** Breast Cancer Classification Using Machine Learning
* **Are you working with anyone else on the same project (data)?** No

**Mini Project Details**

1. **Type of Data Considered:** Text Data
2. **Number of Observations / Subjects:**
   * Total number of samples: 569
   * Categories: B (Benign): 357 samples, M (Malignant): 212 samples
3. **Project Type:** Classification
4. **Data Source:** [breast cancer wisconsin diagnostic](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)
5. **Data Representation Before Feature Extraction:**
   * How did you represent the raw data? Using tables, and graphs
   * How many graphs did you use? 5
     1. **Box plot:** Box plots are used to visualize the distribution, central tendency, and variability of numerical features in the dataset. They help compare feature differences between categories (e.g., Benign vs. Malignant) and easily highlight outliers. This makes them useful for spotting patterns relevant to diagnosis.
     2. **Joint plot:** Joint plots are used to visualize the relationship between two numerical variables along with their distributions.
     3. **Violin plot:** Violin plots show the distribution and density of a numerical feature across different categories. They combine the benefits of box plots and KDE (kernel density estimation) to reveal variations between groups like Benign and Malignant.
     4. **Swarm plot:** A swarm plot can be used to represent a breast cancer dataset for binary classification because it visually shows the distribution of data points for each class. It helps in understanding the spread and density of both malignant and benign cases.
     5. **Heatmap:** A heatmap is useful for representing a breast cancer dataset for binary classification because it visually displays correlations between features, highlighting relationships and patterns that can influence classification. It allows easy identification of feature interactions.
6. **Feature Extraction / Creation Details:**
   * Total number of features extracted: 4
   * Names of features and their formulas (if possible), and brief explanation of their use.
     1. **volume\_mean** = 4/3π (radius\_mean)^3 : Represents the estimated volume of the tumor using the mean radius, helping assess tumor size.
     2. **volume\_worst** = 4/3π (radius\_worst)^3 : Represents the estimated volume using the worst radius, indicating the largest tumor size.
     3. **surface\_area\_mean** = 4π (radius\_mean)^2 : Calculates the surface area of the tumor using the mean radius, providing insights into tumor shape.
     4. **surface\_area\_to\_volume\_ratio** = surface\_area\_mean / volume\_mean :Compares surface area to volume, helping identify tumor growth patterns.
7. **Data Representation After Feature Extraction:**
   * Number of graphs used to represent extracted features: 4
   * Reason for choosing each graph:

**Graphs Used**: Four histogram plots with KDE curves were used to visualize feature distributions:

**Initial Feature**: radius\_mean

**New** **Features**: volume\_mean , surface\_area\_mean, surface\_area\_to\_volume\_ratio

**Reason for Choosing**: Histograms with KDE curves clearly show the distribution and shape of continuous features. They help compare the original feature with newly engineered ones and assess if the new features add meaningful variation. Using distinct colors and subplots improves clarity and visual comparison.

1. **Feature Selection Techniques Used:**
   * Methods used (e.g., Filter, Wrapper, Embedded, or any discussed in class/PPTs): **Filter Method**

**20 features selected**

Highly correlated groups such as radius\_mean, perimeter\_mean, and area\_mean were reduced by keeping only the most informative feature from each group (e.g., area\_mean). This reduces redundancy and multicollinearity. The selection was guided by the heatmap correlation matrix to retain representative features and improve model performance and interpretability.

1. **Feature Transformation Techniques Used:**

**Standardization** was applied to scale features to have zero mean and unit variance. This ensures all features contribute equally to PCA, preventing those with larger scales from dominating the results.

1. **Feature Reduction Techniques Used:**

Principal Component Analysis (PCA) was used to reduce the dimensionality of the dataset while preserving as much variance as possible. The original dataset contained 30 features, many of which were highly correlated. This can lead to redundancy and overfitting in machine learning models.

By applying PCA, I transformed the data into a smaller set of uncorrelated components that still captured most of the important patterns. This not only improves computational efficiency but also helps in visualizing the data and enhancing model performance by removing noise and multicollinearity.

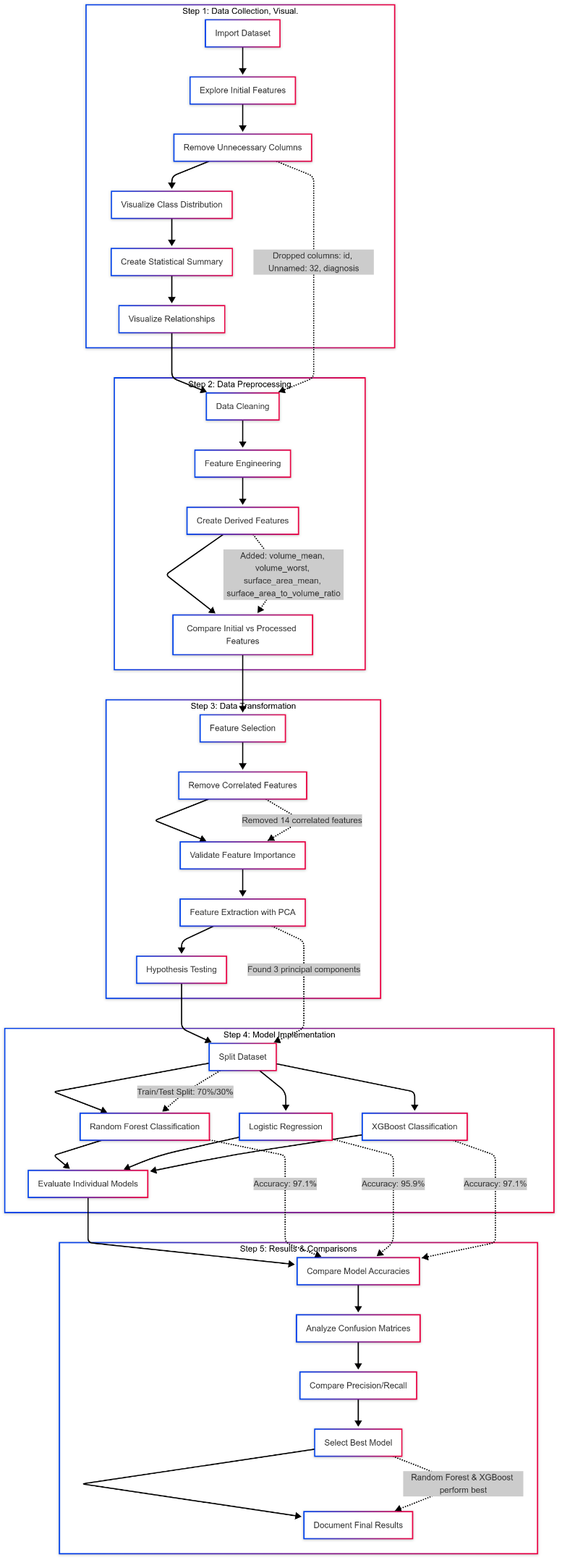
1. **Hypothesis Testing Methods Used:**

A **paired t-test** was performed between radius\_mean and area\_mean to check if there is a statistically significant difference between their means. This test helps identify redundancy between features — in this case, high correlation and a non-significant p-value would support removing one of them to reduce multicollinearity.

1. **Models Employed:**
   * Random Forest Classification
   * Logistic Regression
   * Xgboost Classification
2. **Best Model Selection Criteria (Beyond Accuracy):**
   * Precision,
   * Recall, and
   * F1-Score
3. **Document & Code Upload:**

**Github repo:** <https://github.com/rajat-kumar-thakur/Breast-Cancer-Classification-Using-Machine-Learning>

1. **Workflow Diagram:**



**Optional Course Feedback**

1. **What did you learn from the Data Analytics & Visualization course?**  
   *Through lectures, labs, and the mini project, I gained a solid foundation in key data analytics and visualization techniques. I learned to use Python and specifically R for data manipulation and data visualization tools. I also explored data preprocessing techniques, exploratory data analysis (EDA), and how to draw insights from data. The mini project helped me apply these concepts practically, particularly in using real-world datasets and presenting results clearly through dashboards and plots.*
2. **Was this course helpful in learning new concepts or improving your problem-solving skills?**  
   *Yes, the course was very helpful. It introduced me to structured approaches to analyzing data, cleaning datasets, and visualizing results for interpretation. Working through practical labs and the mini project improved my problem-solving skills significantly, especially when handling messy data, choosing the right visualization for different data types, and deriving actionable insights.*
3. **Suggestions for Improvement (Excluding Internet Issues):**  
   *Include a few sessions focused specifically on using visualization tools like Power BI or Tableau for broader tool exposure.*
4. **Difficult Topics That Need More Explanation:**  
   *Time-series analysis and visualizations were a bit challenging and could use more dedicated sessions.*
5. **Personal Feedback on the Subject or Faculty:**  
   *The instructor was clear throughout the course. I especially appreciate how the course balanced theory and application.* *The practical examples used during lectures made complex topics easier to grasp.*