Exploratory Data Analysis: Wisconsin Diagnostic Breast Cancer (WDBC)

1.1 Introduction

This report analyzes the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to identify key features distinguishing malignant from benign tumors. The data features were computed from digitized images of fine needle aspirates (FNA) of breast masses, describing the characteristics of the cell nuclei present in the image (Wolberg et al. 2003) .

1.2 Data Acquisition

The raw data was retrieved directly from the UCI Machine Learning Repository to ensure reproducibility. The dataset consists of 569 instances with 30 real-valued input features and one binary target variable (`Diagnosis`).

2. Data Cleaning, Schema Mapping, and Data Validation

The raw dataset lacks semantic column headers. To facilitate analysis, we implemented a schema mapping strategy based on the `wdbc.names` metadata. The 30 features represent ten distinct cell nucleus characteristics (e.g., Radius, Texture) computed in three statistical forms.

We applied the following suffix mapping transformation: * **Mean Value:** Suffix `1` -> `_mean` * **Standard Error:** Suffix `2` -> `_se` * **Worst (Max) Value:** Suffix `3` -> `_max`

This step ensures all features are semantically interpretable for the subsequent EDA.

To ensure the dataset is clean, consistent, and ready for modeling, we validated the data by implementing the following checks:

- Correct Data File Format The cleaned dataset was exported as a standard CSV (breast_cancer_cleaned.csv) with UTF-8 encoding. It successfully loaded in pandas without errors, confirming proper file format and readability.

- Correct Column Names All 30 feature columns follow the expected naming convention (_mean, _se, _max). The Diagnosis column contains only "Benign" and "Malignant" values, and the total number of columns is 31, as expected.
- No Empty Observations All rows contain complete observations. There are no fully empty rows, and no partial missing values were detected.
- Missingness Within Expected Threshold A threshold of 5% missingness per column was applied. No columns exceeded this limit, ensuring all features are sufficiently complete for reliable modeling.

By combining schema mapping with these validation checks, the dataset is fully consistent, correctly formatted, and reproducible, providing a robust foundation for downstream modeling and analysis.

Outlier Analysis & Scaling Strategy

Due to the massive disparity in magnitude between features (e.g., `Area` > 2000 vs. `Smoothness` < 0.2), a **Symmetric Log (Symlog) scale** was applied to the visualizations. This effectively mitigates the compressing effect of the `Area` feature, allowing the distribution and spread of smaller-scale variables to be clearly observed without losing the information from larger values.

Post-scaling inspection reveals numerous outliers (points beyond whiskers), particularly in Malignant samples (Orange). * **Significance:** These are **not data errors**. In the context of breast cancer, extreme values in features like `Area`, `Concavity`, and `Perimeter` are characteristic of malignant tumor growth. * **Conclusion:** These points represent high-priority **biological signals** essential for classification.

Preprocessing Recommendation * **Action: Do not drop** these outliers, as removing them would discard critical diagnostic information. * **Strategy:** To handle the skewness and scale differences during modeling: 1. Apply **Log Transformation** (`np.log1p`) to right-skewed features (Area, Perimeter) to normalize distributions. 2. Apply **Standard Scaling** (`StandardScaler`) to all features to ensure the model treats all dimensions with equal weight.

9. Target/response variable follows expected distribution

- We validate that the target variable `Diagnosis` is not severely imbalanced. If one class is much rarer than the other, this can hurt model performance and may require special handling (for example, resampling or adjusting evaluation metrics).

10. No anomalous correlations between target and features

- We check how strongly each feature is associated with `Diagnosis`. Extremely high predictive power for a single feature can indicate data leakage or unexpected dependencies that should be investigated.

Figure 1: No outlier or anomalous values

Figure 2: Target/response variable follows expected distribution

Figure 3: No anomalous correlations between target and features

11. No anomalous correlations between features

- We examine pairwise correlations between features. If many feature pairs are highly correlated, this suggests redundancy or multicollinearity, which may require feature selection or dimensionality reduction.

The `FeatureFeatureCorrelation` check fails the condition we set. In our data there are many feature pairs with very high correlation, for example `radius_mean`, `perimeter_mean` and `area_mean`, as well as their corresponding `_max` and `_se` versions. This pattern is expected for this dataset because these variables all describe related geometric properties of the tumor, so strong correlations are not a data quality error but a sign of redundancy and multicollinearity.

3. Data Profiling: Structure and Statistics

Purpose: * **`df.info()`:** Used to verify data integrity by checking for null values and ensuring all feature columns are of `float64` type. * **`df.describe()`:** Used to examine the central tendency and spread of numeric features. This highlights differences in **magnitude** (scales) across variables.

Observation: The dataset is complete (no missing values). However, `describe()` reveals massive scale disparities (e.g., `area_mean` ranges up to 2500, while `smoothness_mean` is < 0.1), confirming the necessity for **Feature Scaling** (Standardization) before modeling.

4. Correlation Analysis: Pearson vs. Spearman

Method: * **Pearson Correlation:** Measures linear relationships. * **Spearman Correlation:** Measures monotonic rank relationships (non-linear). Comparing both helps identify if relationships are strictly linear or just trending in the same direction.

Purpose: To detect **Multicollinearity**—redundant features that increase model complexity without adding information.

Results: Both metrics show near-perfect correlation (> 0.95) between `Radius`, `Perimeter`, and `Area`. This confirms these features are geometrically redundant. We should retain only one (e.g., Radius) and drop the others to improve model stability.

5. Pairwise Separability Analysis

Purpose: To visualize 2D decision boundaries. We look for feature combinations where the **Benign (Blue)** and **Malignant (Orange)** clusters are clearly distinct with minimal overlap.

Results: * **High Separability:** Features related to size (`radius_mean`) and shape complexity (`concavity_mean`) separate the classes well. * **Non-linear**

Figure 4: No anomalous correlations between features

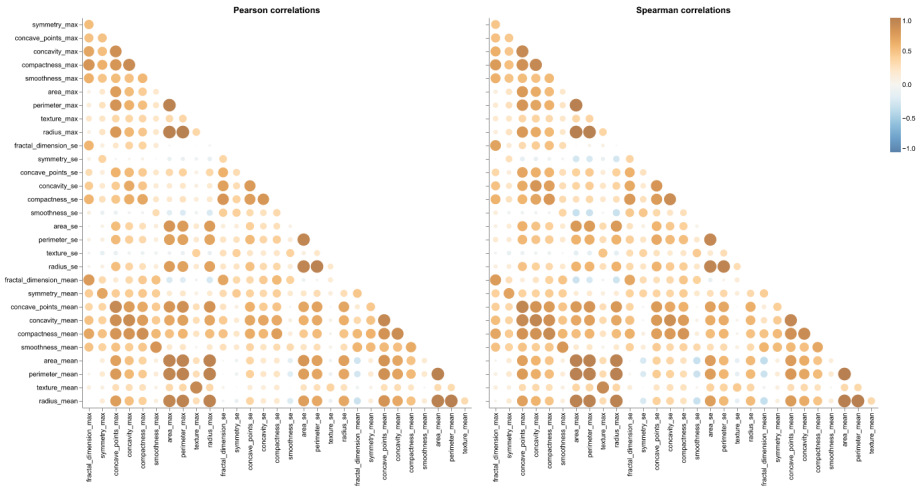

Figure 5: Multicollinearity

patterns: The curved relationship between `area` and `radius` is clearly visible, reinforcing the geometric redundancy found in the correlation analysis.

6. Distribution Analysis

Purpose: To inspect the univariate "shape" of the data. We look for **Skewness** (asymmetry) and **Outliers** that could bias linear models.

Results: * **Skewness:** Features like `area_se` and `concavity_mean` are heavily **right-skewed** (long tail to the right). This indicates that **Log Transformation** is required to normalize these distributions. * **Overlap:** "Texture" and "Smoothness" show high overlap between classes, suggesting they are less informative on their own compared to "Size" features.

EDA Findings

- **Class Separation:**
  - **High Separability:** Features related to **size** (`radius`, `perimeter`, `area`) and **concavity** (`concave_points`, `concavity`) show clear distinction between Benign and Malignant classes (Malignant samples generally have higher values).
  - **Low Separability:** Texture, Smoothness, and Fractal Dimension show significant overlap, indicating they are weaker individual predictors.
- **Distributions:**
  - **Skewness:** "Area" and "Concavity" features (both `_mean` and `_se`) are heavily **right-skewed**.
  - **Outliers:** Visible in the upper tails of `area_max` and `perimeter_se`.

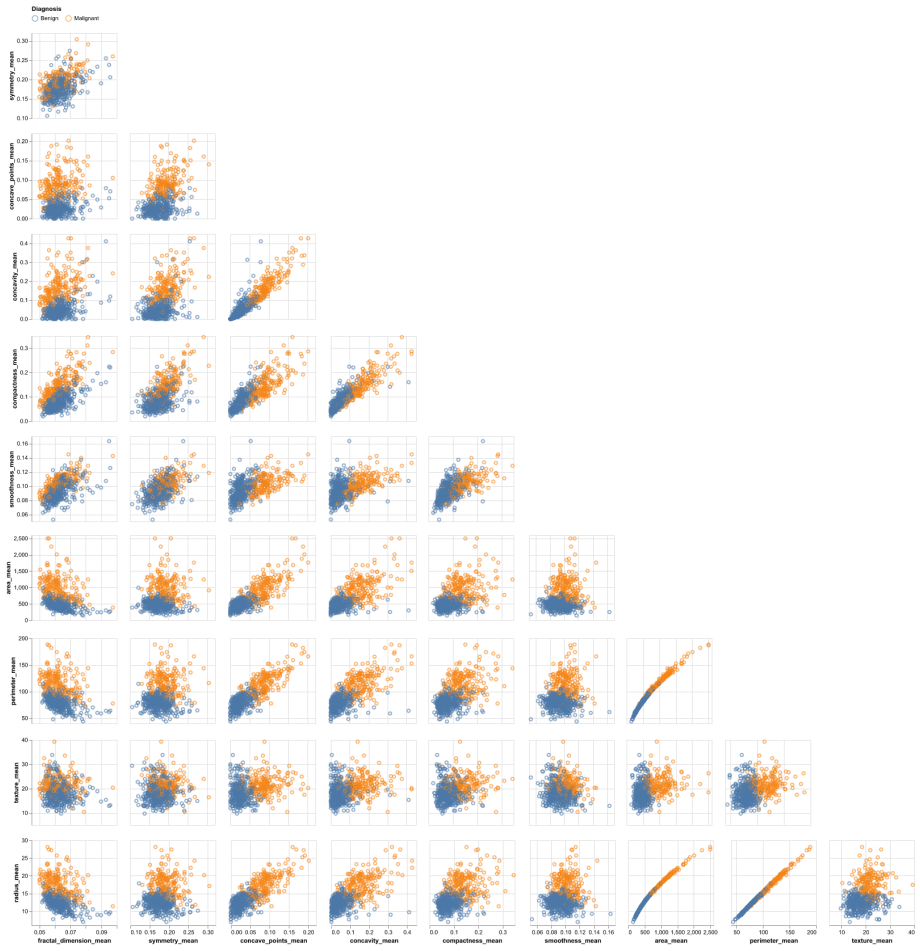

Figure 6: Pairwise Separability

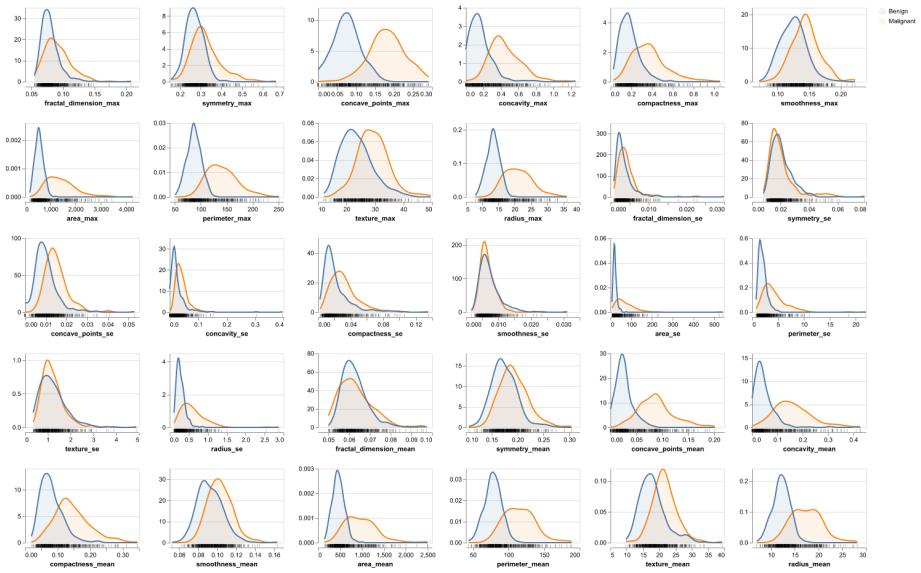

Figure 7: Distribution Analysis

- **Correlations (Multicollinearity):**
  - **Severe Multicollinearity:** `radius`, `perimeter`, and `area` are perfectly correlated ($R \approx 1$). This is expected geometrically but redundant for models.
  - `concavity`, `concave_points`, and `compactness` also exhibit very high positive correlation.

Preprocessing Recommendations

Based on the above, the following pipeline is suggesued:

1. **Feature Selection / Drop:**
   - Remove redundant features to reduce multicollinearity. **Keep `radius`** (or `perimeter`), but **drop `area`** and `perimeter` as they duplicate information.
2. **Transformation:**
   - Apply **Log Transformation** to skewed features (e.g., `area`, `concavity`) to normalize distributions.
3. **Scaling:**
   - Features vary vastly in scale (e.g., `area` > 1000 vs. `smoothness` < 0.2). Use **`StandardScaler`** to standardize all features to unit variance.
4. **Imputation:**
   - None needed (Data is clean).

Onto Creating a Classification Model

Discussion:

Our model demonstrated strong performance, achieving high accuracy on the test set derived from the UCI Machine Learning Repository (Frank 2010) and correctly classifying nearly all cases. This result was generally expected given the strong feature patterns observed during Exploratory Data Analysis (EDA), which suggested a clear separation between benign and malignant tumours.

However, the main concern is the occurrence of a single false negative, where a malignant tumour was incorrectly predicted as benign. Although rare, such an error carries significant clinical risk. Accurate classification is critical for determining appropriate treatment pathways (Board 2025), and missed diagnoses can severely impact patient outcomes given the prevalence and nature of the disease (Society 2007). This highlights that the model, while statistically strong, requires further refinement before it is reliable enough for real-world medical deployment.

These results suggest that future work should explore methods aimed at reducing false negatives, such as adjusting class weights, employing cost-sensitive training, or validating on external datasets to assess robustness.

References

Board, PDQ Adult Treatment Editorial. 2025. "Breast Cancer Treatment (PDQ®)." *PDQ Cancer Information Summaries [Internet]*.

Frank, Andrew. 2010. "UCI Machine Learning Repository." *Http://Archive. Ics. Uci. Edu/Ml*.

Society, American Cancer. 2007. *Breast Cancer Facts & Figures*. American Cancer Society.

Wolberg, Alisa S, Dougald M Monroe, Harold R Roberts, and Maureane Hoffman. 2003. "Elevated Prothrombin Results in Clots with an Altered Fiber Structure: A Possible Mechanism of the Increased Thrombotic Risk." *Blood, The Journal of the American Society of Hematology* 101 (8): 3008–13.

Figure 8: Heatmap for SVM GridSearchCV Mean Test Score

Figure 9: Heatmap for Confusion Matrix