SWINBURNE UNIVERSITY OF TECHNOLOGY

COS40007

Portfolio Assessment 1

Studio 1 - 3

Duc Thuan Tran|104330455

Table of Contents

Overview

This project examines the prediction of breast cancer diagnosis using the Wisconsin Diagnostic dataset, which contains 30 features characterizing cell nuclei from fine needle aspirate (FNA) images. By organizing features into mean, standard error, and worst measurement groups, the exploratory data analysis revealed strong correlations between size-related measurements and boundary irregularity features with diagnosis outcomes. After transforming the binary classification into a balanced 5-class risk assessment model, various feature engineering and selection approaches were implemented. Decision tree models trained on five different feature sets demonstrated that strategic feature selection significantly outperforms using all available features. The Representative Features set achieved

95.91% accuracy with only 7 carefully selected features, matching the performance of the 10-feature Worst Features set while using fewer variables. This work demonstrates the importance of effective feature selection in medical diagnostics, showing that model simplicity and performance can be simultaneously optimized through data-driven feature engineering.

1. Dataset Selection

1.1. Selected Dataset

Out of seven provided datasets in the portfolio, Breast Cancel Wisconsin is the one that will be used for the portfolio. Here is the link to the dataset:
https://archive.ics.uci.edu/dataset/17/breast%2Bcancer%2Bwisconsin%2Bdiagnostic

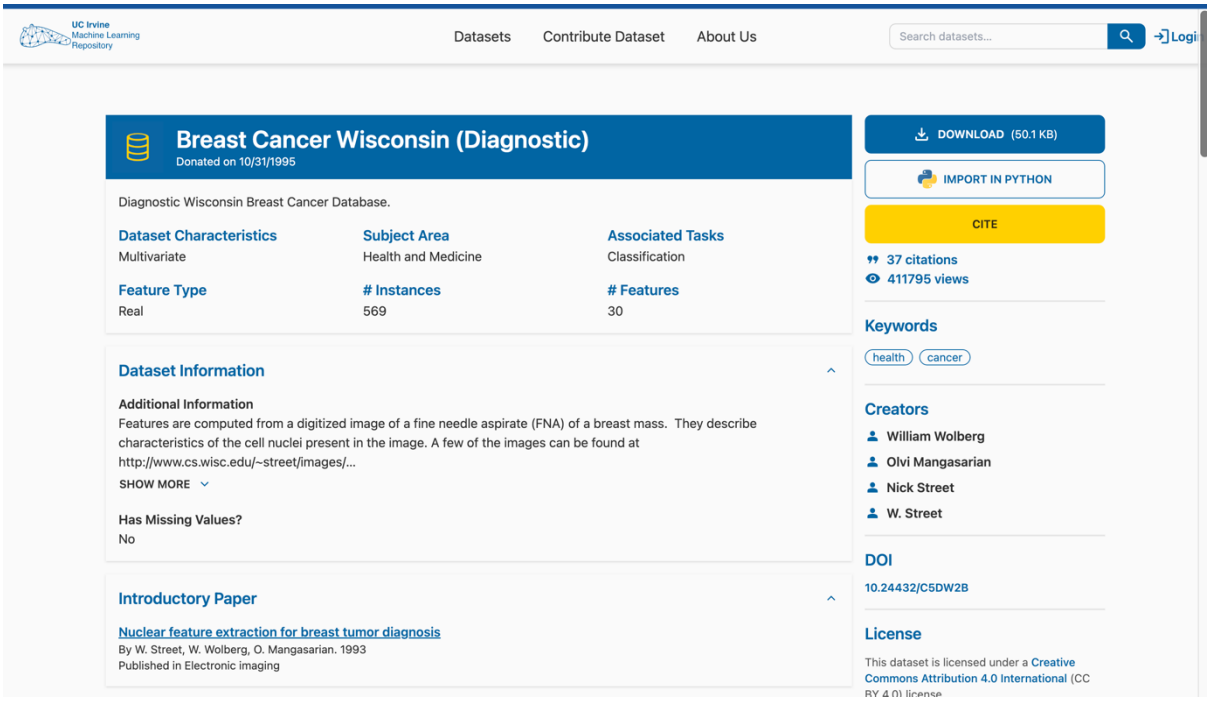

Figure 1. Breast Cancel Wisconsin dataset source.

1.2. Choosing Reason

I chose the Breast Cancer Diagnosis dataset because, as an Artificial Intelligence major, I am interested in how AI can benefit in biomedical engineering field. AI can improve Early Detection & Diagnosis, Treatment Planning, and Prediction Disease Outbreak, making it a powerful tool for healthcare. This dataset allows me to explore how AI can assist in diagnosing medical conditions, particularly breast cancer, and understand its potential to improve patient outcomes through early and accurate detection.

2.Summary of Exploratory Data Analysis

2.1. Variable Identification

Target variables: diagnosis

Predictors:

+ radius1, texture1, perimeter1, area1, smoothness1, compactness1, concavity1, concave_points1, symmetry1, fractal_dimension1.

+ radius2, texture2, perimeter2, area2, smoothness2, compactness2, concavity2, concave_points2, symmetry2, fractal_dimension2.

+ radius3, texture3, perimeter3, area3, smoothness3, compactness3, concavity3, concave_points3, symmetry3, fractal_dimension3.

2.2. Univariate Analysis

There are two types of variables in univariate analysis: continuous variables and categorical variables. Continuous variables are numerical values that can take any value within a range, while categorical variables represent distinct groups or categories.

2.2.1. Continuous Variable

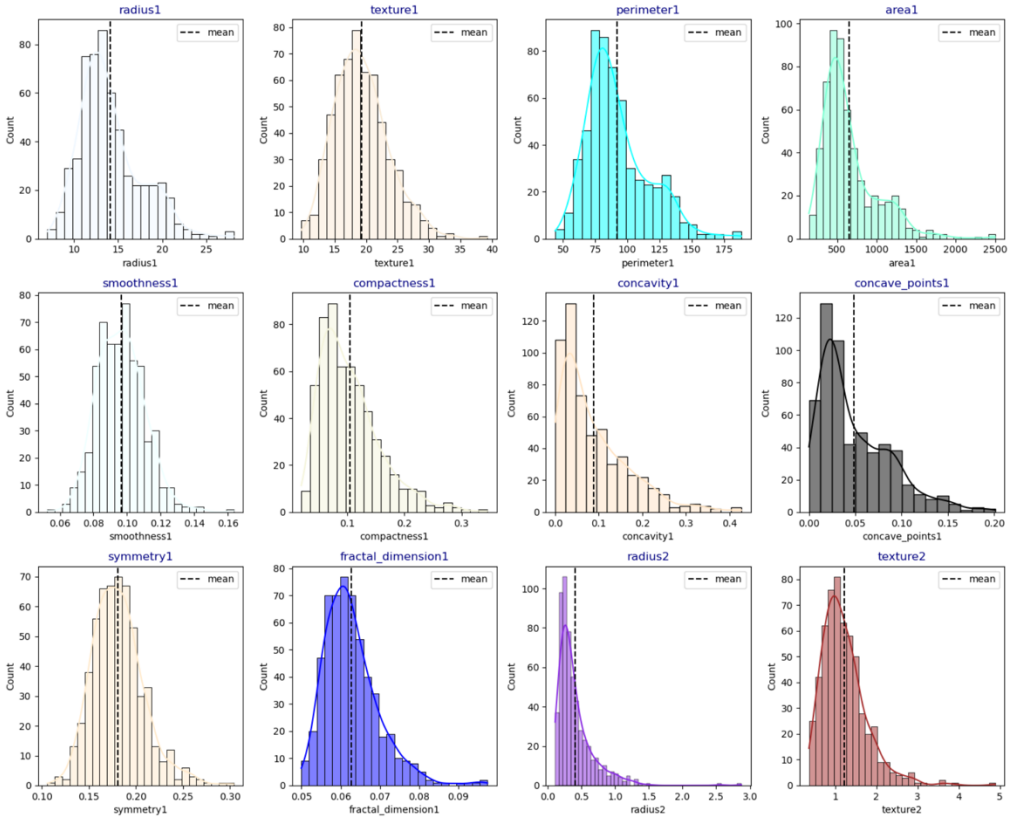

Figure 2. Histograms of Breast Cancer Wisconsin Feature Distributions (1 of 3).

radius1 - Moderately right-skewed distribution -- radius1 is skewed to higher values.

texture1 - Approximately normal distribution -- texture1 is symmetrically distributed around the mean.

perimeter1 - Right-skewed distribution -- perimeter1 is skewed to higher values.

area1 - Strongly right-skewed distribution -- area1 is highly skewed to higher values.

smoothness1 - Approximately normal distribution -- smoothness1 is symmetrically distributed.

compactness1 - Right-skewed distribution -- compactness1 is skewed to higher values.

concavity1 - Right-skewed distribution -- concavity1 is skewed to higher values.

concave_points1 - Right-skewed distribution -- concave_points1 is skewed to higher values.

symmetry1 - Right-skewed distribution -- symmetry1 is skewed to higher values.

fractal_dimension1 – Right-skewed distribution – fractal_dimension1 is skewed to higher values.

radius2 - Moderately right-skewed distribution – radius2 is skewed to higher values.

texture2 - Moderately right-skewed distribution – texture2 is skewed to higher values.

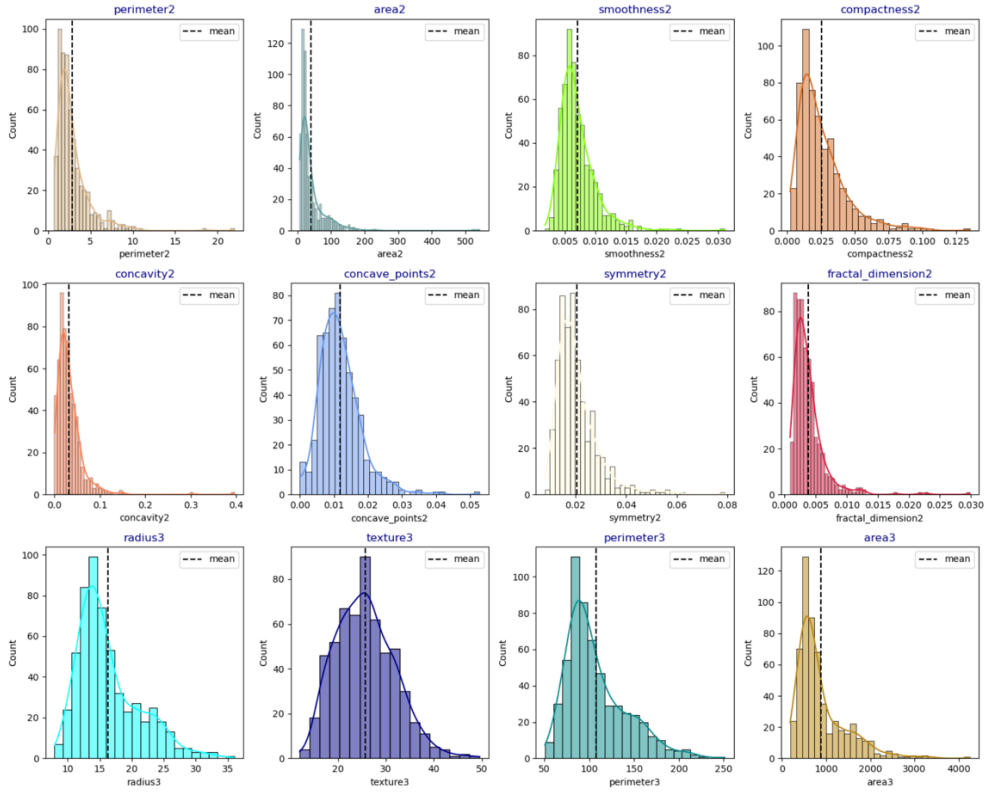

Figure 3. Histograms of Breast Cancer Wisconsin Feature Distributions (2 of 3).

perimeter2 - Right-skewed distribution – perimeter2 is skewed to higher values.

area2 - Strongly right-skewed distribution – area2 is highly skewed to higher values.

smoothness2 - Moderately right-skewed distribution – smoothness2 is highly skewed to higher values.

compactness2 - Right-skewed distribution – compactness2 is skewed to higher values.

concavity2 – Moderately right-skewed distribution – concavity2 is skewed to higher values.

concave_points2 - Moderately right-skewed distribution -- concave_points2 is skewed to higher values.

symmetry2 - Moderately right-skewed distribution – symmetry2 is skewed to higher values.

fractal_dimension2 – Moderately right-skewed distribution – fractal_dimension2 is skewed to higher values.

radius3 - Right-skewed distribution – radius3 is skewed to higher values.

texture3 – Approximately normal distribution – texture3 is symmetrically distributed around the mean.

perimeter3 - Right-skewed distribution – perimeter3 is skewed to higher values.

area3 - Strongly right-skewed distribution – area3 is highly skewed to higher values.

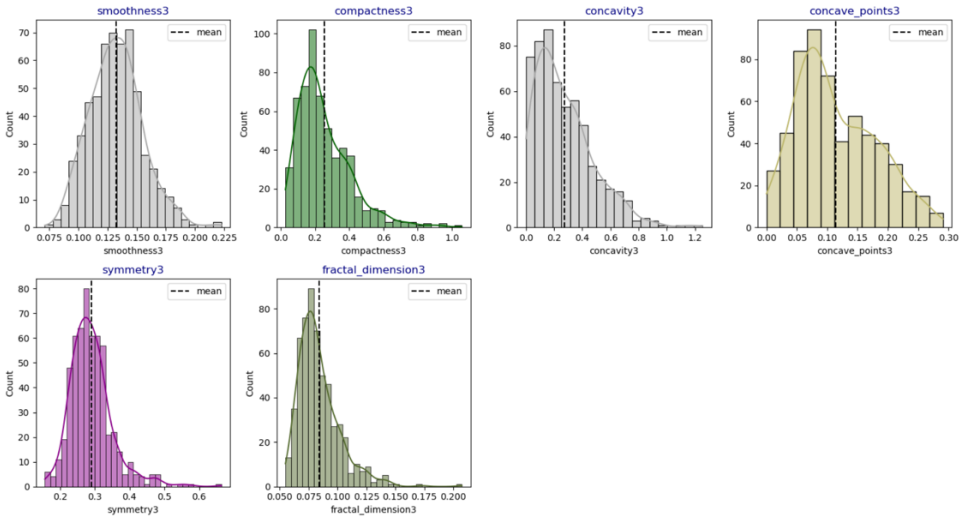

Figure 4. Histograms of Breast Cancer Wisconsin Feature Distributions (3 of 3).

smoothness3 - Approximately normal distribution – smoothness3 is symmetrically distributed around the mean.

compactness3 - Right-skewed distribution – compactness3 is skewed to higher values.

concavity3 – Right-skewed distribution – concavity2 is skewed to higher values.

concave_points3 - Right-skewed distribution -- concave_points3 is skewed to higher values.

symmetry3 - Moderately right-skewed distribution – symmetry3 is skewed to higher values.

fractal_dimension3 – Right-skewed distribution – fractal_dimension3 is skewed to higher values.

2.2.2. Categorical Variable

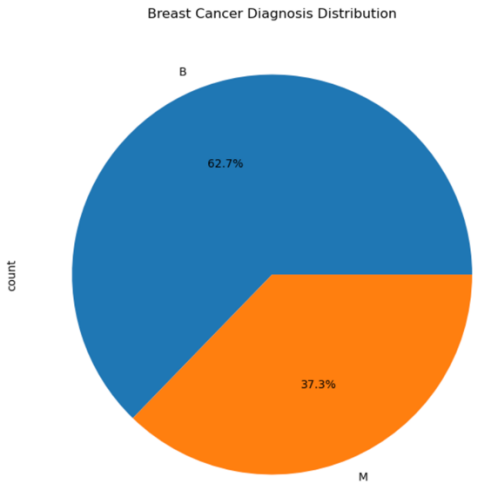

Figure 5. Breast Cancer Diagnosis Distribution.

Diagnosis with two classes: Benign (B) and Malignant (M). The distribution shows: 62.7% Benign (B) and 37.3% Malignant (M)

Benign cases are more frequent, but malignant cases still make up a significant portion, ensuring a balanced analysis.

2.3. Summary Statistics

[19]:

	count	mean	std	min	25%	50%	75%	max
radius1	569.0	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.11000
texture1	569.0	19.289649	4.301036	9.710000	16.170000	18.840000	21.800000	39.28000
perimeter1	569.0	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.50000
area1	569.0	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.00000
smoothness1	569.0	0.096360	0.014064	0.052630	0.086370	0.095870	0.105300	0.16340
compactness1	569.0	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.34540
concavity1	569.0	0.088799	0.079720	0.000000	0.029560	0.061540	0.130700	0.42680
concave_points1	569.0	0.048919	0.038803	0.000000	0.020310	0.033500	0.074000	0.20120
symmetry1	569.0	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.30400
fractal_dimension1	569.0	0.062798	0.007060	0.049960	0.057700	0.061540	0.066120	0.09744
radius2	569.0	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.87300
texture2	569.0	1.216853	0.551648	0.360200	0.833900	1.108000	1.474000	4.88500
perimeter2	569.0	2.866059	2.021855	0.757000	1.606000	2.287000	3.357000	21.98000
area2	569.0	40.337079	45.491006	6.802000	17.850000	24.530000	45.190000	542.20000
smoothness2	569.0	0.007041	0.003003	0.001713	0.005169	0.006380	0.008146	0.03113
compactness2	569.0	0.025478	0.017908	0.002252	0.013080	0.020450	0.032450	0.13540
concavity2	569.0	0.031894	0.030186	0.000000	0.015090	0.025890	0.042050	0.39600
concave_points2	569.0	0.011796	0.006170	0.000000	0.007638	0.010930	0.014710	0.05279
symmetry2	569.0	0.020542	0.008266	0.007882	0.015160	0.018730	0.023480	0.07895
fractal_dimension2	569.0	0.003795	0.002646	0.000895	0.002248	0.003187	0.004558	0.02984
radius3	569.0	16.269190	4.833242	7.930000	13.010000	14.970000	18.790000	36.04000
texture3	569.0	25.677223	6.146258	12.020000	21.080000	25.410000	29.720000	49.54000
perimeter3	569.0	107.261213	33.602542	50.410000	84.110000	97.660000	125.400000	251.20000
area3	569.0	880.583128	569.356993	185.200000	515.300000	686.500000	1084.000000	4254.00000
smoothness3	569.0	0.132369	0.022832	0.071170	0.116600	0.131300	0.146000	0.22260
compactness3	569.0	0.254265	0.157336	0.027290	0.147200	0.211900	0.339100	1.05800
concavity3	569.0	0.272188	0.208624	0.000000	0.114500	0.226700	0.382900	1.25200
concave_points3	569.0	0.114606	0.065732	0.000000	0.064930	0.099930	0.161400	0.29100
symmetry3	569.0	0.290076	0.061867	0.156500	0.250400	0.282200	0.317900	0.66380
fractal_dimension3	569.0	0.083946	0.018061	0.055040	0.071460	0.080040	0.092080	0.20750

Figure 6. Breast Cancer Wisconsin Dataset Statistical Summary.

Sample size: 569 observations for all features

Size measurements:

Areas (area1, area2, area3) have the largest values, with area3 having the highest mean (880.58) and maximum (4254.0).

Perimeters (perimeter1, perimeter2, perimeter3) are also substantial, with means ranging from about 2.87 to 107.26.

Radius (radius1, radius2, radius3) show interesting differences, with radius1 and radius3 having similar magnitudes (means of 14.12 and 16.27), but radius2 is much smaller (mean of 0.41).

Shape descriptors:

Texture features (texture1, texture2, texture3) show significant variation between groups.

Smoothness, compactness, concavity, and concave_points generally have small values below 1.0.

The "2" features (smoothness2, compactness2, etc.) generally have smaller values than their "1" and "3" counterparts.

Symmetry and fractal_dimension features are relatively small across all three sets.

Variability:

Area measurements show high standard deviations, indicating wide dispersion.

Several concavity features have minimum values of 0, suggesting some observations lack concavity entirely.

Distribution characteristics:

For most features, the median (50%) is closer to the 25% quartile than the 75% quartile, suggesting right-skewed distributions.

The maximum values are often substantially higher than the 75% quartiles, indicating potential outliers.

2.4. Multivariate Analysis

2.4.1. Introduction and Feature Organization

The Breast Cancer Wisconsin Diagnostic dataset contains 30 distinct features derived from digitized images of fine needle aspirate (FNA) of breast masses. These features represent various cellular characteristics that can help differentiate between malignant and benign tumours. Due to the large number of features, this analysis organizes them into three meaningful groups:

- **Mean Features** (suffix "1"): Average values of cell nucleus characteristics
- **Standard Error Features** (suffix "2"): Standard error of measurements
- **Worst Features** (suffix "3"): The "worst" or most extreme values observed

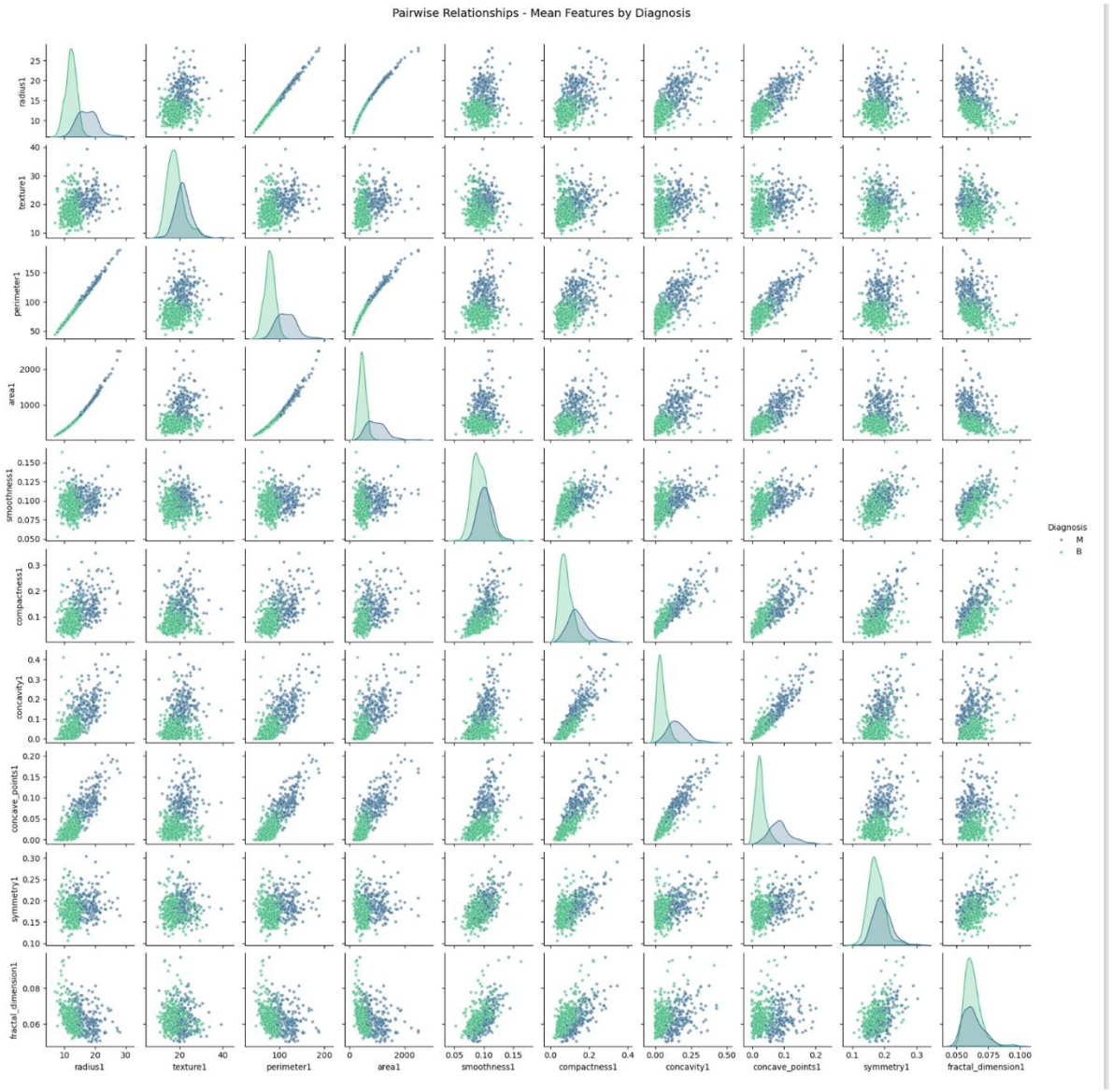

Figure 7. Pairwise Relationships - Mean Features by Diagnosis.

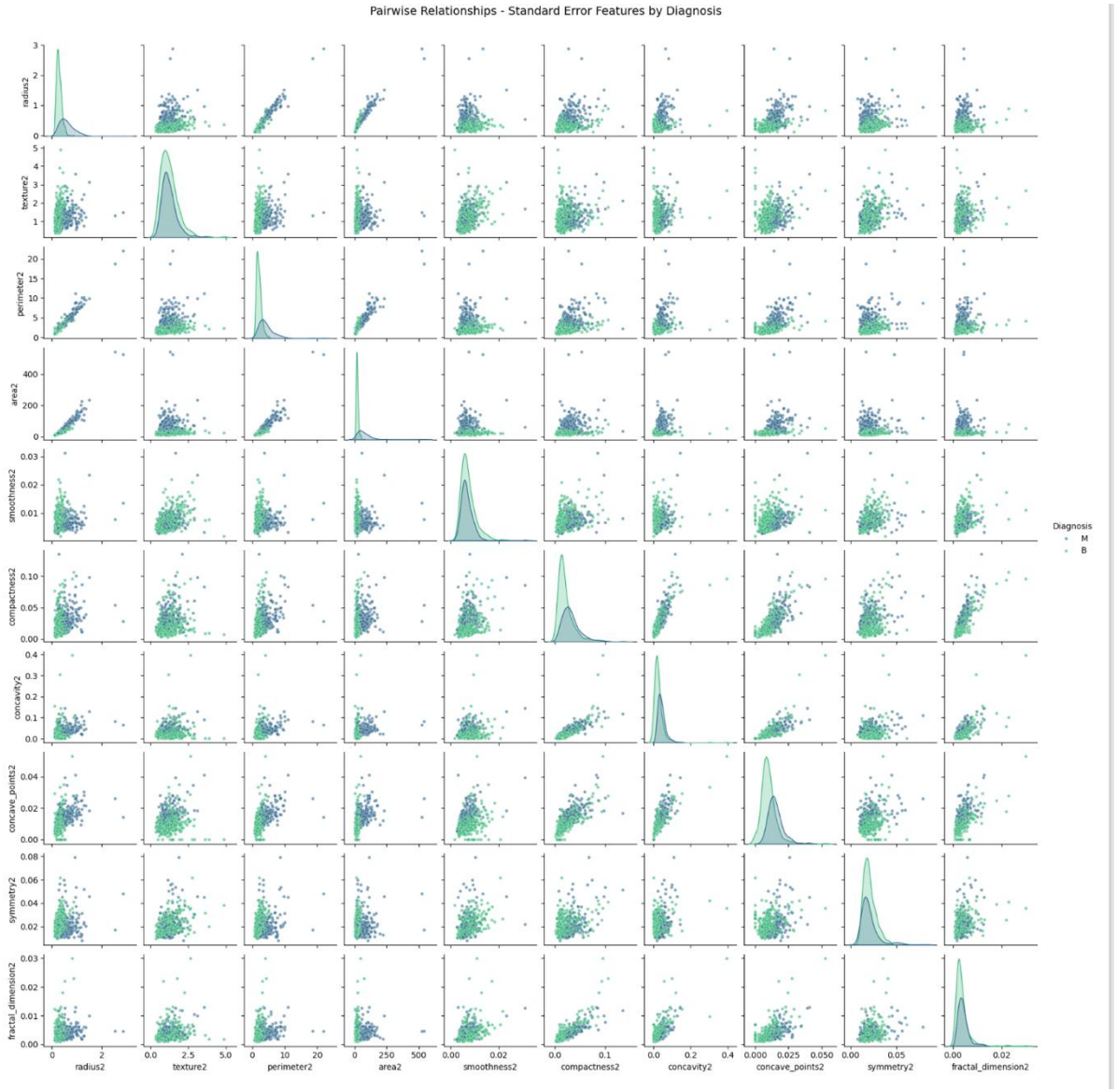

Figure 8. Pairwise Relationships – Standard Error Features by Diagnosis.

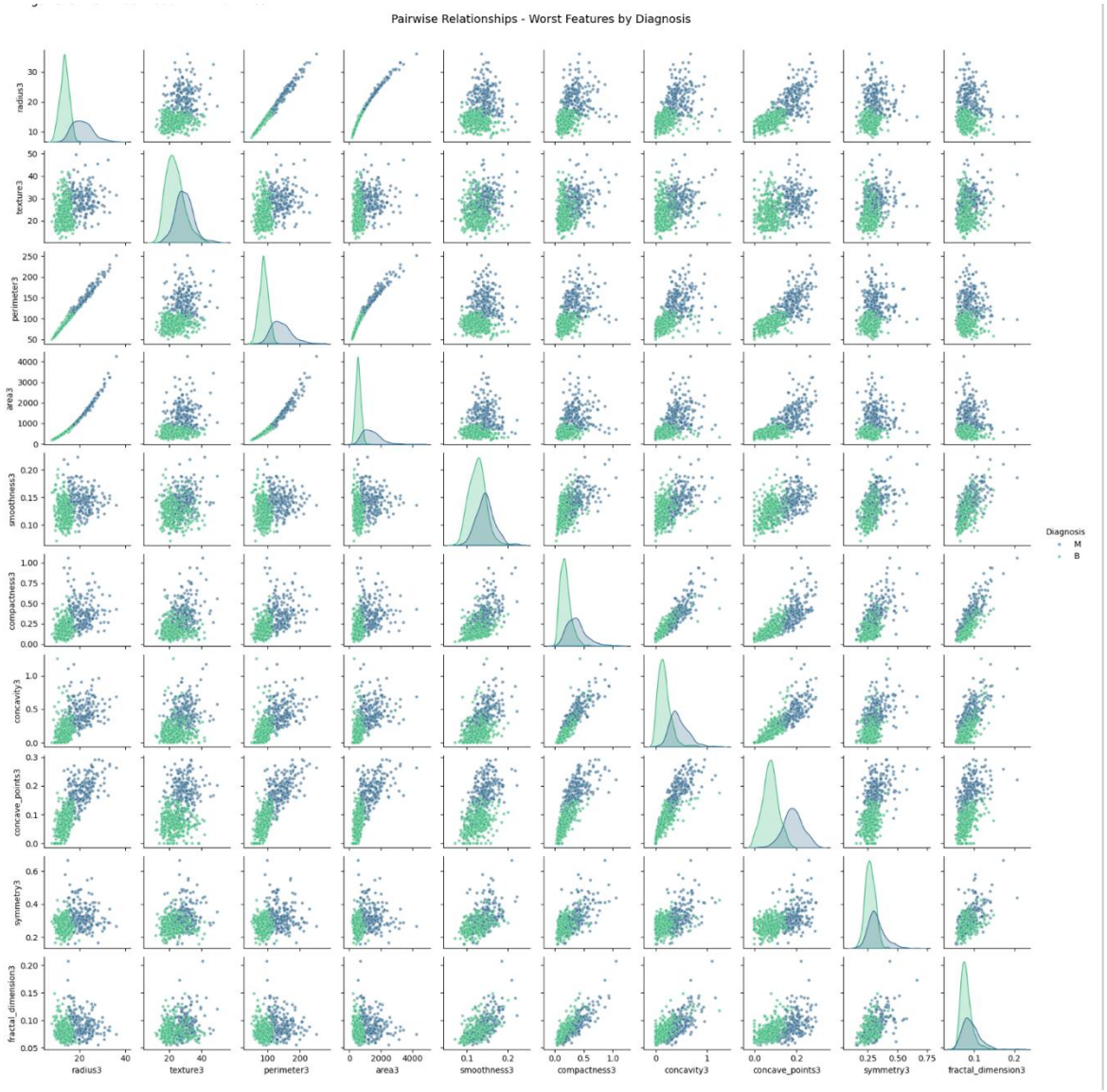

Figure 9. Pairwise Relationships - Worst Features by Diagnosis.

This grouping strategy allows for more effective visualization and interpretation of relationships between features and their diagnostic value. The following analysis examines each feature group separately, identifying patterns that can inform feature selection and engineering for breast cancer diagnosis.

2.4.2. Diagonal Analysis

The KDE diagonal plots across all three feature types (mean, standard error, worst) show consistent bimodal distributions in most features, particularly pronounced in the "worst" features.

The separation between malignant (blue) and benign (green) classes is most distinct in size-related measures (radius, perimeter, area) and boundary irregularity features (concavity, concave_points) across all three measurement types.

The "worst" features show the clearest separation between diagnostic classes, followed by mean features, while standard error features generally show more overlap between classes.

Features like smoothness, symmetry, and fractal_dimension demonstrate less distinct separation between diagnostic classes in their diagonal distributions.

2.4.3. Off Diagnoal Analysis: Relationships between Independent Features

Mean Features

- Strong positive linear relationships exist between radius1, perimeter1, and area1 (r-value near 1), indicating high redundancy among these size measurements.
- Concavity1 and concave_points1 show strong correlation with each other and moderate correlation with size features.
- Texture1, smoothness1, symmetry1, and fractal_dimension1 display weaker correlations with other features, suggesting they capture different cell characteristics.

Standard Error Features

- Strong correlations exist between radius2, perimeter2, and area2, indicating measurement redundancy.
- The standard error features generally show more spread and less clear separation between diagnostic classes compared to mean and worst features.
- The relationships between different standard error measurements appear more scattered, suggesting higher variability in these measurements.

Worst Features

- The strongest correlations again appear among radius3, perimeter3, and area3, with near-perfect linear relationships.
- Compactness3, concavity3, and concave_points3 show moderate to strong correlations with each other.
- The worst features demonstrate the most pronounced separation between malignant and benign cases, particularly for concave_points3, concavity3, and the size-related measures.
- Texture3 shows better separation between classes than its mean and standard error counterparts.

2.4.4. Diagnosis Relationship: Patterns across Feature Types

Mean Features vs Diagnosis

- Concave_points1, concavity1, radius1, perimeter1, and area1 show the strongest separation between malignant and benign cases.
- Smoothness1 and fractal_dimension1 show the least separation among mean features.

Standard Error Features vs Diagnosis

- Standard error features generally show less clear separation between diagnostic classes.
- Radius2, area2, and concave_points2 offer the best, though modest, separation among standard error features.

Worst Features vs Diagnosis

- The worst features provide the clearest distinction between malignant and benign cases.

- Concave_points3, concavity3, perimeter3, area3, and radius3 show excellent separation between classes.
- Compactness3 and texture3 demonstrate better diagnostic value compared to their counterparts in mean and standard error measurements.

2.4.5. Cross-Feature Type Observations

For each characteristic, the "worst" measurement consistently provides better diagnostic separation than mean or standard error, making worst features the strongest predictors of diagnosis.

Standard error measurements generally provide the least distinct separation between diagnostic classes and are therefore less reliable predictors.

The correlation patterns between features remain consistent across the three feature types, with similar relationships observed in mean, standard error, and worst measurements.

Among all measurements, the best predictors of diagnosis include:

- from the worst features group - concave_points3, concavity3, radius3, perimeter3, and area3.
- from the mean features group - concave_points1, concavity1, radius1, perimeter1, and area1.
- from the standard error group - radius2, area2, and concave_points2, though these are less effective than their counterparts in the other groups.

2.5. Correlation

2.5.1. Overall Correlation Heatmap

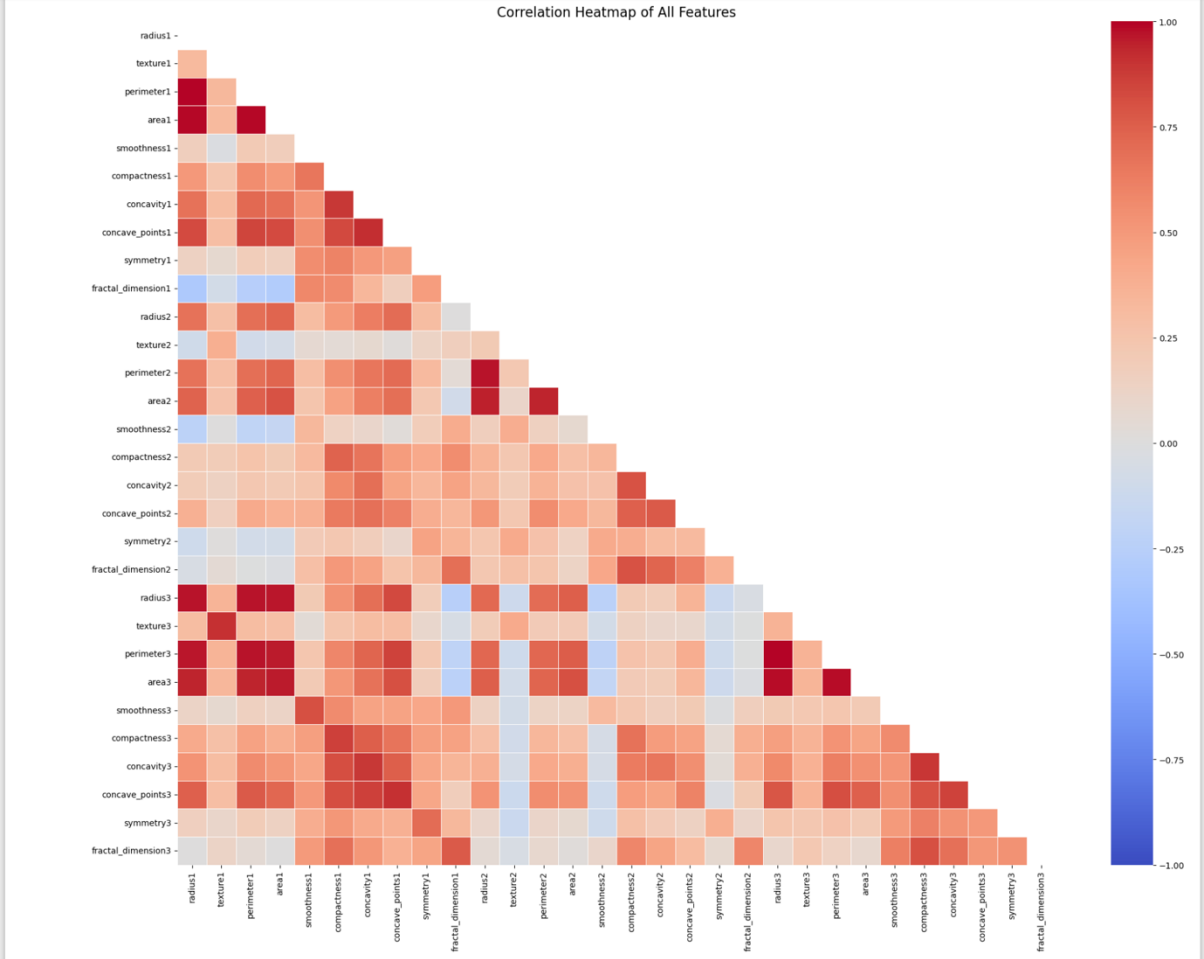

Figure 10. Correlation Heatmap of All Features.

Strong positive correlations (dark red) exist between size-related measurements across all three feature types (radius, perimeter, area)

Concavity and concave_points show strong positive correlations with each other across all feature types

Features within the same measurement type (mean, SE, worst) tend to correlate more strongly with each other than with features from other types

2.5.2. Mean Feature Correlation Heatmap

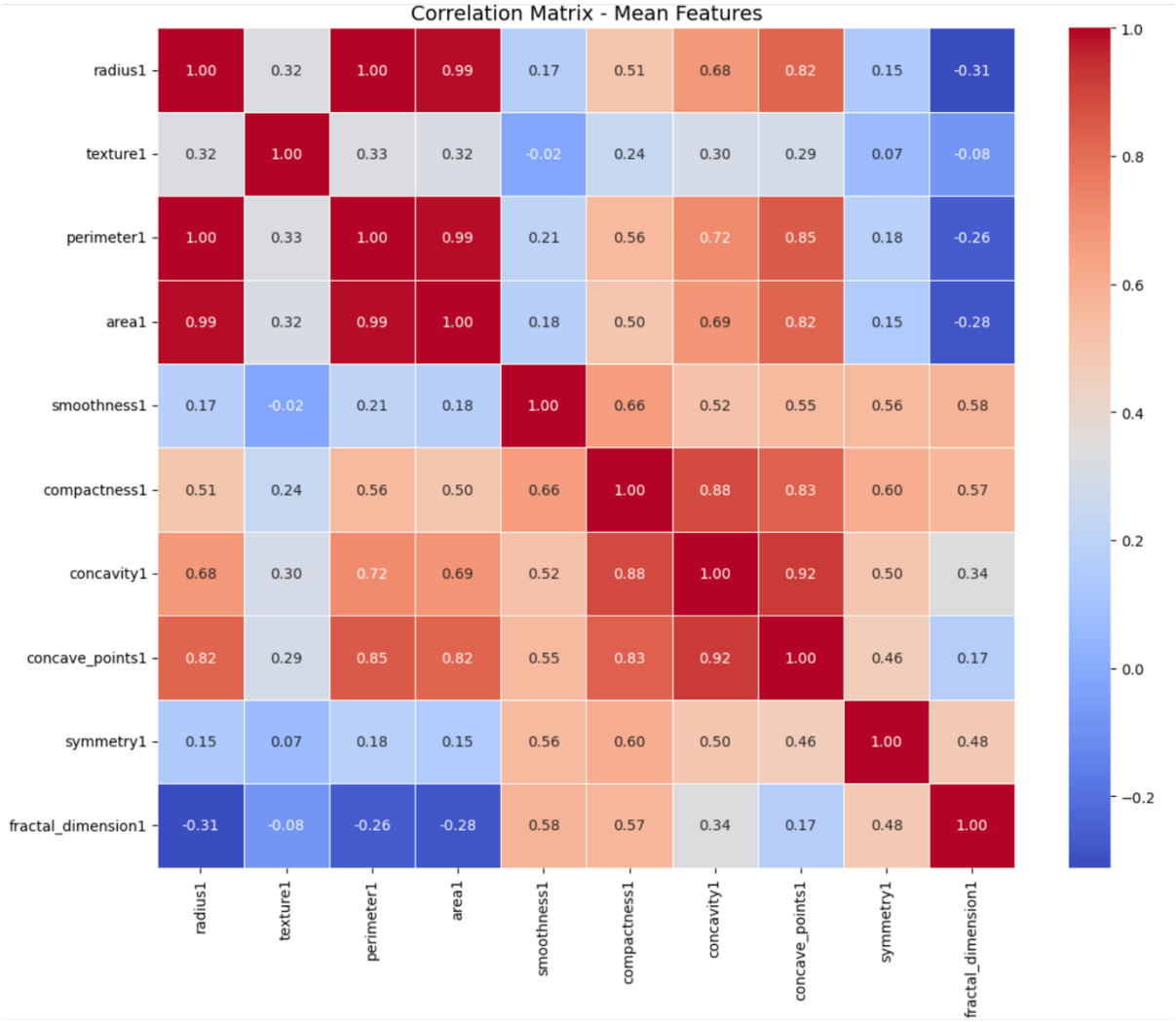

Figure 11. Correlation Matrix – Mean Features.

Radius1, perimeter1, and area1 have extremely high positive correlations (0.99-1.00) with each other, indicating redundancy

Concavity1 and concave_points1 show very strong positive correlation (0.92), suggesting they measure similar aspects of cell nuclei

Compactness1 shows strong positive correlation with concavity1 (0.88) and concave_points1 (0.83)

Fractal_dimension1 has moderate negative correlation with radius1 (-0.31), perimeter1 (-0.26), and area1 (-0.28)

Texture1 shows relatively weak correlations with other features, suggesting it captures unique information

2.5.3. Standard Error Feature Correlation Heatmap

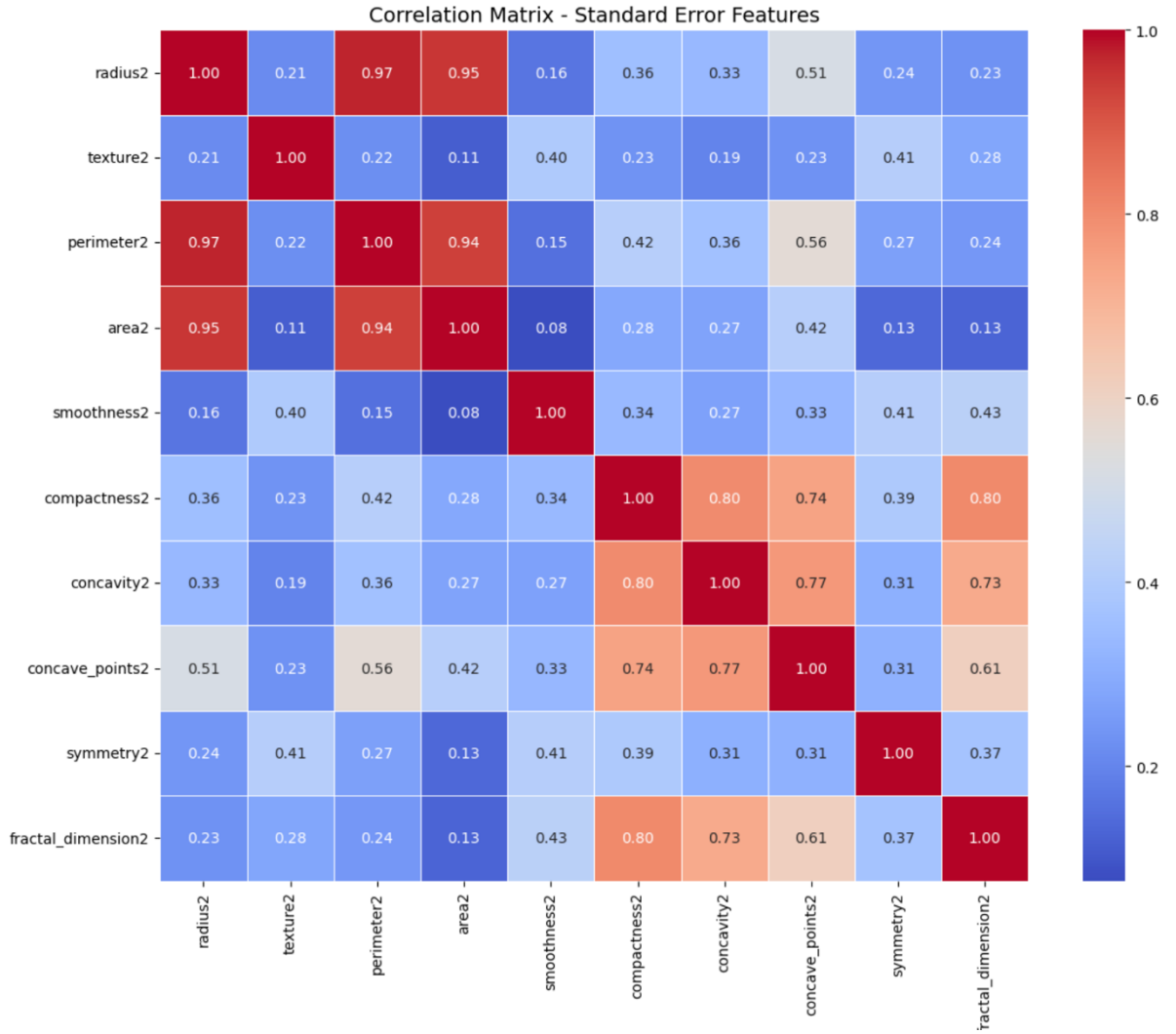

Figure 12. Correlation Matrix – Standard Error Features.

Radius2, perimeter2, and area2 show very strong positive correlations (0.94-0.97) with each other

Compactness2 has strong positive correlations with concavity2 (0.80) and fractal_dimension2 (0.80)

Concavity2 and concave_points2 show strong positive correlation (0.77)

Texture2 generally shows weak correlations with other standard error features

2.5.4. Worst Error Feature Correlation Heatmap

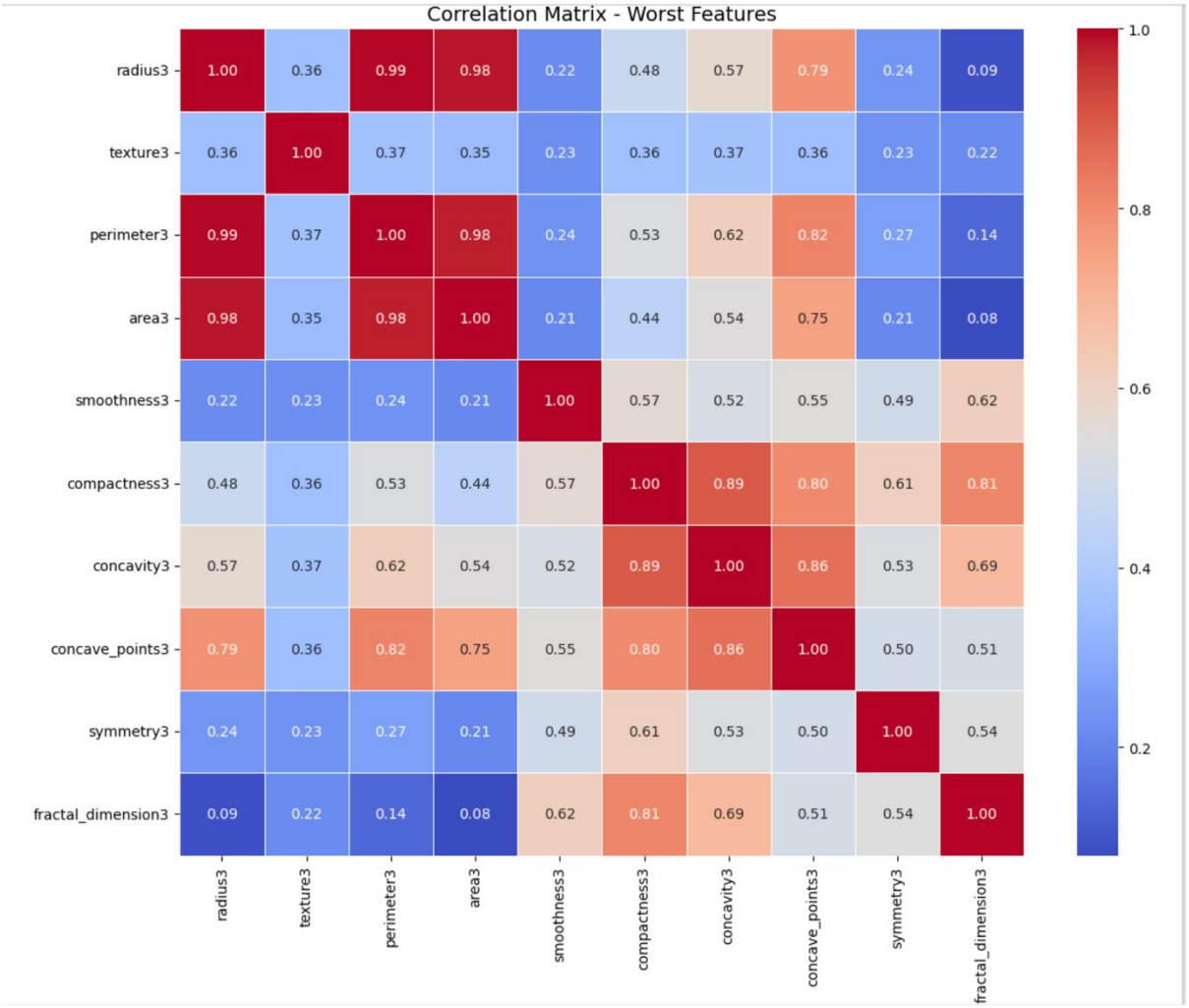

Figure 13. Correlation Matrix – Worst Error Features.

Radius3, perimeter3, and area3 show extremely high positive correlations (0.98-0.99) with each other

Concavity3 and concave_points3 show strong positive correlation (0.86)

Compactness3 shows strong positive correlations with concavity3 (0.89), concave_points3 (0.80), and fractal_dimension3 (0.81)

Concave_points3 has strong positive correlations with radius3 (0.79) and perimeter3 (0.82)

Fractal_dimension3 shows very low correlation with size measurements (radius3, perimeter3, area3)

2.5.5. Key Insights from Correlation Analysis:

The size measurements (radius, perimeter, area) are highly redundant across all feature types and could be reduced to a single representative feature

Boundary irregularity features (compactness, concavity, concave_points) form another highly correlated group

Texture features generally have weaker correlations with other features, suggesting they provide complementary information

Fractal dimension shows negative correlation with size measurements but positive correlation with boundary irregularities

3. Class Labeling For Target Variable / Developing Ground True Data

3.1. Approach to Multi-class Classification

For the Breast Cancer Wisconsin Diagnostic dataset, the original target variable is binary, classifying tumours as either malignant (M) or benign (B). To develop a more nuanced classification system, I transformed this binary classification into a 5-class risk assessment scale using a composite score approach based on key predictive features identified in the EDA.

The composite score method used the following steps:

1. Selected the most discriminative features based on correlation analysis and multivariate visualization:
    - concave_points3
    - concavity3
    - radius3
    - perimeter3
    - area3
2. Normalized these features using MinMaxScaler to ensure equal contribution to the composite score
3. Created a composite risk score by calculating the mean of these normalized features
4. Divided the population into quintiles based on this composite score
5. Assigned risk categories to each quintile:
    - Very Low Risk
    - Low Risk
    - Moderate Risk
    - High Risk
    - Very High Risk

3.2. Result of Class Labelling

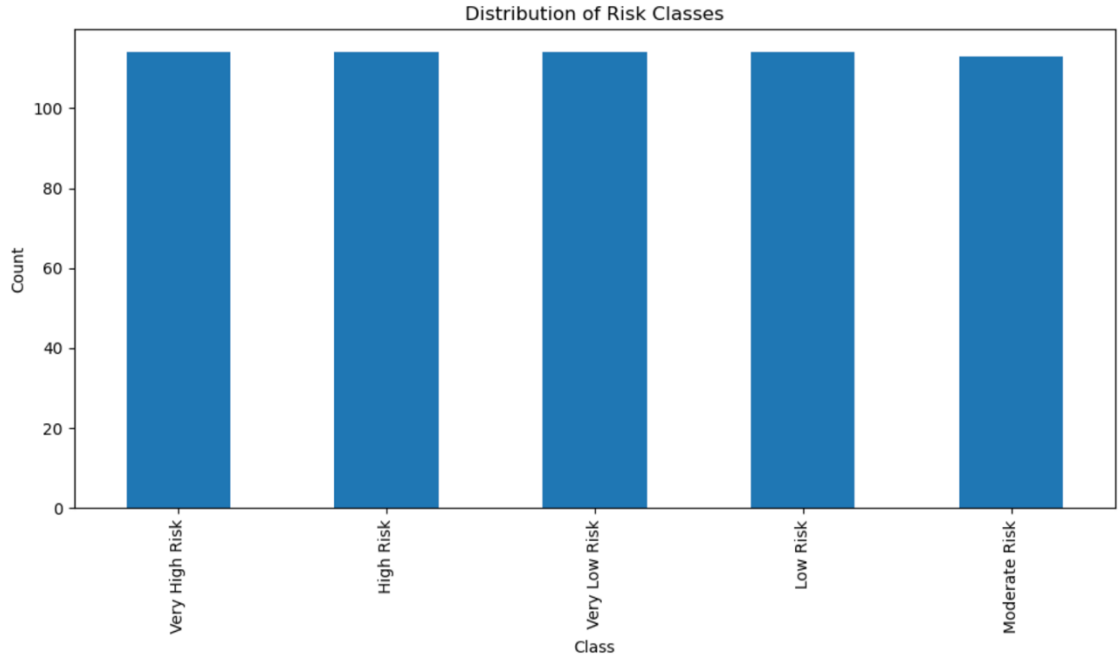

Figure 14. Distribution of Risk Classes

The implementation resulted in a nearly perfect balance among the five classes:

Risk Class	Sample Count
Very High Risk	114
High Risk	114
Very Low Risk	114
Low Risk	114
Moderate Risk	113

3.3 Correlation with Original Diagnosis

Note: Results might vary slightly between runs due to normalization processes and percentile-based division methods.

The newly created risk classes show strong alignment with the original diagnosis, with a clear progression of malignancy rates:

Risk Class	Malignant %	Benign%
Very High Risk	100.00%	0.00%
High Risk	78.07%	21.93%

Moderate Risk	6.19%	93.81%
Low Risk	1.75%	98.25%
Very Low Risk	0.00%	100.00%

This distribution demonstrates that the risk classification effectively captures the underlying diagnostic information while providing a more graduated assessment scale.

4. Feature Engineering and Feature Selection

4.1 Feature Engineering

Several new features were engineered based on insights from exploratory data analysis:

1. **Ratio features**: Created ratios of worst to mean measurements (e.g., radius3/radius1) to quantify the degree of cellular abnormality by measuring how extreme the worst measurements are compared to average values.
2. **Difference features**: Calculated the absolute difference between worst and mean measurements to capture the range of variation in cell characteristics.
3. **Composite scores**: Combined related features into meaningful group scores:
    - Irregularity scores: Combined concavity, concave points, and compactness measures
    - Size scores: Integrated radius, perimeter, and area measurements

This engineering process created 26 new features that potentially capture important patterns in the data that individual features might miss.

4.2 Feature Selection

Five distinct feature sets were created for model comparison:

1. **Feature Set 1**: All 30 original features (normalized) - provides a baseline using the complete feature set
2. **Feature Set 2**: 10 worst features only (those ending with '3') - focuses on the measurement type that showed best class separation
3. **Feature Set 3**: 10 top correlated features with diagnosis - includes the most predictive features identified through correlation analysis
4. **Feature Set 4**: 7 representative features from different correlation groups - maintains diversity while minimizing redundancy:
    a. radius3 (representing size measures)
    b. texture3
    c. concave_points3 (representing boundary irregularity)
    d. symmetry3
    e. fractal_dimension3

f. radius1 (mean size representation)

g. concave_points1 (mean boundary irregularity)

5. **Feature Set 5**: 26 engineered features only - tests whether derived features outperform original measurements

5. Training And Decision Tree Model Development

5.1. Model Training Process

The model training process followed a structured approach to evaluate how different feature engineering and selection strategies affected breast cancer diagnostic performance:

1. **Data Preparation**: First, the code cleaned all feature sets using a custom function that replaced infinity values with NaN and then filled these with column means to ensure model stability.
2. **Dataset Organization**: Five distinct feature sets were prepared:
   - All Original Features (feature set 1)
   - Worst Features Only (feature set 2)
   - Top Correlated Features (feature set 3)
   - Representative Features (feature set 4)
   - Engineered Features (feature set 5)
3. **Train-Test Split**: For each feature set, the data was split into training (70%) and testing (30%) sets using train_test_split with stratification to maintain the same class distribution in both sets.
4. **Model Initialization**: A DecisionTreeClassifier was initialized with a fixed random_state=42 to ensure reproducibility across all experiments.
5. **Model Training**: The classifier was fitted to the training data using the fit () method with the feature set and corresponding target labels.
6. **Prediction**: The trained model generated predictions on the test set using the predict () method.

5.2. Evaluation Process

After training each model, a comprehensive evaluation was performed:

1. **Metric Calculation**: Four key performance metrics were calculated for each model:
   - **Accuracy**: Measures the overall percentage of correct predictions, indicating how often the model correctly classified both malignant and benign cases.
   - **Precision**: Determines the proportion of positive predictions that were correct, reflecting the model's ability to avoid false positives.
   - **Recall**: Identifies the proportion of actual positive cases correctly identified, showing the model's ability to find all malignant cases.

- **F1-Score**: Provides the harmonic mean of precision and recall, balancing the model's ability to be both precise and complete in its diagnosis.

2. **Results Storage**: All metrics along with feature counts and confusion matrices were stored in a dictionary for each feature set.
3. **Classification Reports**: Detailed classification reports were generated showing class-specific performance metrics (precision, recall, F1-score) for both benign and malignant classes, as well as macro and weighted averages.

6. Final Table Comparison and Observation

6.1. Model Performance Comparison Table

Feature Set	Feature Count	Accuracy	Precision	Recall	F1 Score
Worst Features Only	10	0.959064	0.959287	0.959064	0.958856
Representative Features	7	0.959064	0.960135	0.959064	0.958700
Top Correlated Features	10	0.918129	0.917855	0.918129	0.917858
Engineered Features	26	0.906433	0.906201	0.906433	0.905782
All Original Features	30	0.900585	0.900166	0.900585	0.900079

Figure 15. Comparison Tabel for All Feature Set.

	Accuracy	Precision	Recall	F1 score
Worst Features Only	0.959064	0.959287	0.959064	0.958856
Representative Features	0.959064	0.960135	9.959064	0.958700
Top Correlated Features	0.918129	0.917855	0.918129	0.917858
Engineered Features	0.906433	0.906201	0.906433	0.905782
All Original Features	0.900585	0.900166	0.900585	0.900079

Note: Results might vary slightly between runs due to normalization processes and percentile-based division methods.

6.2. Key Observations from Model Comparison

"Worst Features Only" and "Representative Features" both achieved the highest accuracy of 95.91%, demonstrating that carefully selected features outperform using all original features.

The "Representative Features" set achieved top performance with only 7 features, showing that a small set of well-chosen features can be just as effective as larger feature sets. This set also achieved the highest precision (0.960135).

Despite having all 30 original features, the "All Original Features" set performed worst with just 90.06% accuracy, confirming that more features don't necessarily lead to better performance.

The "Engineered Features" set performed only slightly better than using all original features, suggesting that the specific engineering approach used may not have captured additional discriminative information.

The "Top Correlated Features" set performed well with 91.81% accuracy, but not as well as the worst or representative feature sets.

7. Appendix: Source Code Repository

The complete source code for this project is available at the following link:

https://github.com/ThuanDanchoi/COS40007/blob/main/Portfolio%20Assessment%201/code/portfolio_assessment1.ipynb

8. References

Shanawad, V. (2023). Random Forest with Bootstrap Sampling for Beginners. *Kaggle*. Retrieved from https://www.kaggle.com/code/vinayakshanawad/random-forest-with-bootstrap-sampling-for-beginner

Wolberg, W.H., Street, W.N., & Mangasarian, O.L. (1995). Breast Cancer Wisconsin (Diagnostic) Dataset. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/dataset/17/breast%2Bcancer%2Bwisconsin%2Bdiagnostic

GeeksforGeeks. (2023). Decision Tree Implementation in Python. Retrieved from https://www.geeksforgeeks.org/decision-tree-implementation-python/