Exploratory data analysis (EDA)

Overall Conclusions of Used Approaches
I implemented four different data processing and feature selection/engineering strategies to develop a predictive model. For each approach, I performed exploratory data analysis (EDA). Below is a comprehensive summary of the key findings from each method, along with overarching insights.

1. Baseline (Approach 1)

- **Purpose:**
  - This approach serves as an unprocessed benchmark, providing a baseline for comparison with more advanced methods.
  - No data cleaning, feature selection, or outlier removal was performed.
  - The baseline model highlights the raw performance of the dataset without any preprocessing, which is useful for evaluating the impact of subsequent feature engineering and selection steps.

2. Correlation-Based Selection & Outlier Removal (Approach 2)

- **Steps:**
  - Removed 359 highly correlated features ($r > 0.8$) to address multicollinearity and overfitting.
  - Identified and removed 5 outliers using the Isolation Forest algorithm.
- **Key Insights:**
  - **Feature Importance:** The top 10 features identified are primarily related to:
    1. **Tumor Enhancement Texture:** Features like *WashinRate_map_Cluster_Prominence_tumor* and *WashinRate_map_information_measure_correlation2_tumor* highlight the importance of texture-based metrics.
    2. **Tumor Size and Morphology:** Features such as *Volume_cu_mm_Tumor* and *TumorMajorAxisLength* emphasize the role of tumor size and shape in distinguishing ER status.
    3. **Tumor Enhancement Variation:** Metrics like *WashinRate_map_skewness_tumor* suggest that variations in tumor enhancement patterns are also significant.
    4. **Tumor Enhancement:** *Grouping_based_proportion_of_tumor_voxels_3D_tumor_Group_1*.
  - **Visualizations:**
    - **t-SNE:** Shows slightly better clustering compared to UMAP, with localized regions where one class dominates. However, significant overlap remains, indicating that ER status is not easily separable.
    - **UMAP:** The classes are more evenly mixed, with no clear separation.
- **Conclusion:** While this approach reduces redundancy and identifies key features, the moderate overlap in visualizations suggests that additional feature engineering or non-linear modeling techniques may be necessary.

3. Statistical Distribution-Based Selection (Approach 3)

- **Steps:**
  - Applied the two-sample Kolmogorov-Smirnov (KS) test to identify features with significant distributional differences between ER status groups.
  - Selected the top 10 features based on the lowest p-values (< 0.00001).

- **Key Insights:**
    - **Feature Importance:** The top 10 features are consistent with Approach 2 in highlighting:
        1. **Tumor Enhancement:** Features like *SER_Washout_tumor_vol_cu_mm* and *SER_map_mean_tumor* emphasize the importance of SER-derived volumetric and intensity metrics.
        2. **Tumor Size and Morphology:** Metrics such as *Volume_cu_mm_Tumor* and *TumorMajorAxisLength* again appear as significant.
        3. **Tumor Enhancement Texture:** Features like *WashinRate_map_Cluster_Prominence_tumor* and *WashinRate_map_inverse_difference_moment_normalized_tumor* reinforce the role of texture-based descriptors.
    - **Visualizations:**
        - **Boxplots and KDE Plots:** Show moderate separation between ER status groups, but significant overlap remains.
        - **t-SNE and UMAP:** Both reveal overlapping clusters, with no clear separation between classes.
- **Conclusion:** The KS test effectively identifies features with significant distributional differences, but the lack of clear separation in visualizations suggests that these features alone are insufficient for classification. More complex interactions or non-linear relationships may need to be explored.

4. Non-Linear Dimensionality Reduction + Robust Scaling (Approach 4)

- **Steps:**
    - Applied Robust Scaling to handle outliers and non-normal distributions.
    - Used UMAP for non-linear dimensionality reduction, achieving a trustworthiness score of 0.865.
    - Identified the top 10 UMAP components using the Point-Biserial Correlation Coefficient.
- **Key Insights:**
    - **UMAP Performance:** The high trustworthiness score indicates that UMAP effectively preserves local and global data structures, making it suitable for high-dimensional datasets.
    - **Feature Importance:** The top 10 UMAP components show weak linear correlations with ER status (ranging from 0.0730 to 0.0441), suggesting that linear methods may not capture the underlying relationships effectively.
    - **Visualizations:**
        - **UMAP and t-SNE** reveal complex, non-linear structures in the data, but no clear separation between ER status groups is observed in 2D visualizations.
- **Conclusion:** While UMAP does not yield striking 2D separation, it may still be valuable for advanced non-linear classification algorithms (e.g., Random Forests, Gradient Boosting, or Neural Networks) that can exploit the preserved local and global structures.

General Observations and Recommendations

1. **Feature Redundancy:**
    - The high correlation among features (341 out of 529 with $r > 0.8$) underscores the need for dimensionality reduction or feature selection to improve model performance and interpretability.
2. **Class Separation:**

    - While some features (e.g., tumor size, texture, and enhancement metrics) show moderate separation between ER status groups, the classes are not easily separable using the current feature set.
3. **Outliers:**
    - Isolation Forest effectively identifies outliers, but their impact on model performance should be further investigated.
4. **Non-Linear Relationships:**
    - UMAP and t-SNE reveal complex, non-linear structures in the data, suggesting that non-linear models may outperform linear ones.
5. **Top-10 Features:**
    - Both Approaches 2 and 3 consistently identify **tumor size** (*Volume_cu_mm_Tumor*) and **morphological metrics** (*TumorMajorAxisLength*) as important features.
    - **Texture and enhancement features** (e.g., *WashinRate_map_Cluster_Prominence_tumor* and *Grouping_based_proportion_of_tumor_voxels_3D_tumor_Group_1*) also appear in both lists, indicating their discriminative power for ER status.
    - Approach 2 emphasizes *WashinRate_map-based features* related to texture and contrast, while Approach 3 highlights *SER-based volumetric and mean intensity measurements*. This reflects how different feature-selection criteria (correlation thresholds vs. statistical distribution tests) can surface complementary sets of important radiomic features.

Summary

The analysis highlights the importance of tumor size, morphology, and enhancement texture features in distinguishing ER status. However, the moderate class separation and weak linear correlations suggest that more sophisticated modeling approaches are needed to fully exploit the dataset's potential. By leveraging non-linear algorithms and advanced feature engineering techniques, it may be possible to achieve better classification performance and gain deeper insights into the relationship between MRI-derived radiomic features and ER status.