Class imbalance:

The dataset shows a **slight class imbalance**, with **62.74% benign (B)** and **37.26% malignant (M)** cases, resulting in an **imbalance ratio of 1.68:1**. Since this ratio is below the typical **2:1 threshold for severe imbalance**, standard models should perform reasonably well. However, if **accurate malignant detection** is a priority, techniques like **SMOTE, oversampling, or class weighting** can help improve model performance.

Scatter Plot:

1. **Radius Mean vs. Texture Mean**

* A scattered distribution with no clear linear relationship.
* It indicates that these features may not be strongly correlated.

1. **Perimeter Mean vs. Area Mean**

* A clear **positive correlation**, meaning as the perimeter increases, the area also increases.
* This is expected since larger tumors will generally have higher perimeters and areas.

1. **Smoothness Mean vs. Compactness Mean**

* A scattered pattern, indicating a **weak correlation** between these two features.

1. **Concavity Mean vs. Concave Points Mean**

* A **strong positive correlation** since concavity and concave points are closely related.
* Tumors with a high concavity also tend to have many concave points.

Histogram:

**1. Radius Mean**

* The values are **right-skewed**, meaning most tumors have a small radius, but some have significantly larger values.
* The distribution suggests that a **majority of tumors have a lower radius**, but a few cases show much larger tumor sizes.

**2. Texture Mean**

* The texture values appear more **normally distributed** compared to the other features.
* This suggests that texture varies more evenly across different tumors.

**3. Perimeter Mean**

* Similar to the **radius mean**, the **perimeter values are also right-skewed**.
* Most tumors have a smaller perimeter, while a few cases exhibit much larger perimeters.
* This aligns with the radius distribution, as **larger radii contribute to larger perimeters**.

**4. Area Mean**

* The distribution shows a strong **positive skew** (long tail on the right).
* This indicates that **most tumors are small in area**, but some have **exceptionally large areas**.
* This feature might be useful for distinguishing between **benign (small) and malignant (large) tumors**.

**5. Smoothness Mean**

* The values are clustered more centrally with less skew.
* This suggests that **smoothness does not vary as drastically** compared to features like area or perimeter.

**6. Compactness Mean**

* This feature has a **right-skewed distribution**, meaning most tumors have low compactness, but some have high compactness values.
* Higher compactness values might be more indicative of **malignant tumors**.

**7. Concavity Mean**

* The distribution suggests that **most tumors have low concavity**, but some have much higher values.
* Since concavity refers to how indented the tumor's surface is, **higher concavity values are often linked to malignancy**.

**8. Concave Points Mean**

* The histogram shows a similar pattern to concavity, with a right-skewed distribution.
* Tumors with many **concave points** tend to be more irregular in shape, which is an indicator of malignancy.

Application of Preprocessing Techniques with Justifications:

* Variable Transformation:

Logarithmic transformation was applied to area\_mean and perimeter\_mean to reduce skewness and normalize data distribution for better model performance.

* Discretization:

Continuous features like radius\_mean were binned into categories (e.g., small, medium, large) to simplify patterns and enhance interpretability.

* Removing Noise:

Outliers in features like radius\_mean a were identified using the mean method. These outliers could be misleading by introducing extreme values that do not represent the general trend of the data. Removing them ensured a more stable learning process.

* Feature Selection:

Using SelectKBest with ANOVA F-test, the top 5 most relevant features (radius\_mean, texture\_mean, perimeter\_mean, area\_mean, and smoothness\_mean) were selected. This helped reduce dimensionality, improve computational efficiency, and enhance model interpretability while retaining the most predictive attributes for distinguishing between malignant and benign tumors.

* Normalization:

The dataset contained features with varying scales, such as radius\_mean (small values) and area\_mean (larger values). Min-Max Scaling was applied to bring all numerical features to the same range (0 to 1), preventing certain features from dominating others in the learning process. Additionally, Standard Scaling (Z-score normalization) was used to center the data around a mean of 0 and a standard deviation of 1, ensuring better convergence in gradient-based models.

Results of Preprocessing Techniques

This preprocessing ensures the dataset is optimized for machine learning models by improving feature relevance, handling categorical values, and preparing data for better performance during training by:

* Handling missing values ensured data completeness, preventing potential issues during model training.
* Encoding categorical variables transformed non-numeric data into a machine-readable format, improving model compatibility.
* Feature selection reduced dimensionality, enhancing computational efficiency and improving model interpretability.
* Scaling standardized feature values, ensuring fair weighting and improving model convergence speed.

**Raw vs. Preprocessed Data**

|  |  |  |
| --- | --- | --- |
|  | Raw Data | Preprocessed Data |
| Feature reduction | Contained more features, including those with potentially less relevance to classification. | Retained only the top 5 most significant features (radius\_mean, texture\_mean, perimeter\_mean, area\_mean, smoothness\_mean), improving computational efficiency. |
| Categorical Encoding | The diagnosis column was categorical (M and B). | This column was converted into numerical values (0 for benign, 1 for malignant), making it compatible with machine learning algorithms. |
| Scaling and Normalization | Feature values were in their original ranges, leading to potential discrepancies in model weight assignments. | Applied both Min-Max Scaling (range 0-1) and Standard Scaling (mean 0, std 1) to standardize features, ensuring fair weighting and improving model convergence. |
| Impact on Model Performance | Higher-dimensional space with unscaled features could result in inefficient learning and longer training times. | Optimized feature selection and scaling helped in faster convergence, reduced overfitting, and improved model interpretability. |