

Data Preparation:

To prepare the data, I first read in the dataset and examined its structure. The dataset contained 386 entries and 10 columns, with a mix of categorical and numerical data. I identified and removed 11 duplicate rows. There were also missing values in the 'tumor-size' and 'inv-nodes' columns, which I addressed by imputing values based on the mean of respective breast-quadrant groups. Additionally, I performed one-hot encoding on categorical variables to ensure they could be used effectively in machine learning models. These preprocessing steps helped convert the dataset into a structured format suitable for training predictive models.

Insights from Data Preparation:

During the data preparation phase, I observed that some categorical variables had significant class imbalances, meaning that certain categories had a much higher frequency than others. This imbalance could affect model performance by biasing predictions towards the majority class. The 'deg-malig' column was the only numeric variable without modification, while variables such as 'tumor-size' and 'inv-nodes' had to be transformed for consistency, as they were initially represented as categorical ranges. The presence of missing values and duplicates indicated potential data collection inconsistencies that required careful handling. Addressing these issues ensured that the dataset was clean, structured, and ready for machine learning model training.

Model Training Procedure:

I implemented three machine-learning models:

1. K-Nearest Neighbor (KNN) Classifier
2. K-Nearest Neighbor (KNN) Classifier using Grid Search CV
3. Linear Classification

To ensure reliable model evaluation, I split the dataset into training (70%) and testing (30%) sets. The training set was used to fit the models, while the testing set was used to evaluate their performance.

For the **Linear Classification** model, I used an **SGDClassifier** with a learning rate of 0.1 and a perceptron loss function. The model was trained iteratively using gradient descent to minimize classification errors. This approach allows for efficient learning on large datasets but is sensitive to hyperparameters like the learning rate and regularization strength.

For the **K-Nearest Neighbor (KNN) Classifier**, I initially set $k=3$ to determine its baseline performance. KNN relies on computing distances between data points, so proper feature scaling was essential. To improve accuracy, I performed **hyperparameter tuning** using **Grid Search CV**, testing multiple values of k (from 1 to 100) to find the optimal value that minimizes classification error.

Model Performance:

My models performed as expected, with the Linear classifier achieving relatively high accuracy, while the KNN model showed signs of overfitting. The overfitting in KNN could be attributed to the chosen grid parameters ranging from 1 to 100, as lower values of K tend to fit the training data too closely. Increasing the starting value of the grid from 1 to 2, or selecting a more optimal range, might have helped reduce overfitting and improved the model's generalization to unseen data. Additionally, tuning other hyperparameters, such as distance metrics or weighting strategies, could have further enhanced performance.

Which metric is most important for this problem?

In this problem, recall is the most important metric because it focuses on identifying as many actual recurrence cases as possible. Since the dataset deals with breast cancer recurrence classification, missing a true recurrence, or having a false negative, could have serious consequences, such as delaying necessary treatment. A high recall ensures that fewer recurrence cases go undetected, which is critical in medical diagnoses where missing a positive case could be life-threatening. While precision, which measures how many predicted positive cases are actually correct, is also valuable, it's generally less important in this context. False positives might lead to extra tests or treatments, but false negatives could mean a missed diagnosis, putting a patient at serious risk. Because of this, recall should be the top priority when evaluating model performance for this problem.

Observations of Plots:

The data reveals key patterns in tumor characteristics and progression. Tumor size distribution suggests certain sizes are more common, which may impact early detection and treatment. A skewed distribution could indicate tumors are often diagnosed at specific stages, highlighting gaps in screening. Similarly, the histogram of tumor location suggests an uneven distribution across breast quadrants, possibly due to anatomical or biological factors, which may inform targeted screening efforts. The box plot of inv-nodes shows variability in lymph node involvement, a crucial factor in cancer severity. A wide spread or outliers suggest differences in tumor aggressiveness, affecting prognosis and treatment planning. Overall, these insights could help analyze recurrence rates, patient demographics, and risk factors for better cancer management.

Model Confidence:

Overall, the models struggled with classification, likely due to high variance or weak predictive features in the data. The Linear Classifier performed better than KNN, but both models had low accuracy and recall, making their predictions unreliable. The overfitting in KNN further reduces confidence in the results. To improve performance, techniques like feature engineering or using more advanced models, such as decision trees or ensemble methods, could be explored.