**Question 1**

Given data:

* **Correct malignant predictions (True Positives, TP):** 15
* **Correct benign predictions (True Negatives, TN):** 75
* **Incorrect malignant predictions (False Positives, FP):** 3
* **Incorrect benign predictions (False Negatives, FN):** 7

Now, let's calculate:

1. **Error rate:**

Error rate=FP+FNTotal predictions=3+715+75+3+7=10100=0.1Error rate=Total predictionsFP+FN​=15+75+3+73+7​=10010​=0.1

1. **Sensitivity (Recall):**

Sensitivity=TPTP+FN=1515+7=1522≈0.682Sensitivity=TP+FNTP​=15+715​=2215​≈0.682

1. **Precision:**

Precision=TPTP+FP=1515+3=1518≈0.833Precision=TP+FPTP​=15+315​=1815​≈0.833

1. **F-measure:**

F-measure=2×Precision×RecallPrecision+Recall=2×0.833×0.6820.833+0.682≈0.75F-measure=Precision+Recall2×Precision×Recall​=0.833+0.6822×0.833×0.682​≈0.75

**Summary:**

* **Error rate:** 0.1
* **Sensitivity:** 0.682
* **Precision:** 0.833
* **F-measure:** 0.75

**Question 2**(a) **Underfitting** occurs when a machine learning model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and test data. The major cause of underfitting is using a model that is too basic (e.g., low complexity, insufficient features).

(b) **Overfitting** happens when a model learns not only the underlying patterns but also the noise and random fluctuations in the training data, leading to poor generalization on unseen data. It usually occurs when the model is too complex relative to the amount of training data.

(c) **Overfitting** occurs when the model is overly tuned to the training dataset, capturing irrelevant patterns (noise) along with the important ones, usually due to high model complexity, insufficient data, or lack of regularization techniques.

**Question 3**

It seems like the question got cut off after "somebody who." Based on what's provided, it's likely asking for the **probability that someone who tests positive is actually resistant to the antibiotic** (i.e., the **posterior probability**, which can be calculated using **Bayes' theorem**).

Here’s the setup:

* **P(R):** Probability of being resistant = 0.02
* **P(NR):** Probability of not being resistant = 0.98
* **False positives:** P(Positive | NR) = 0.01
* **False negatives:** P(Negative | R) = 0.05
* **True positives:** P(Positive | R) = 1 - 0.05 = 0.95

We need to find **P(R | Positive)**, the probability that a person is resistant given a positive test result:

P(R∣Positive)=P(Positive∣R)⋅P(R)P(Positive)P(R∣Positive)=P(Positive)P(Positive∣R)⋅P(R)​

Where:

P(Positive)=P(Positive∣R)⋅P(R)+P(Positive∣NR)⋅P(NR)P(Positive)=P(Positive∣R)⋅P(R)+P(Positive∣NR)⋅P(NR)

Let’s plug in the values:

P(Positive)=(0.95⋅0.02)+(0.01⋅0.98)=0.019+0.0098=0.0288P(Positive)=(0.95⋅0.02)+(0.01⋅0.98)=0.019+0.0098=0.0288

Now, calculate **P(R | Positive):**

P(R∣Positive)=0.95⋅0.020.0288=0.0190.0288≈0.659P(R∣Positive)=0.02880.95⋅0.02​=0.02880.019​≈0.659

So, the probability that someone who tests positive is actually resistant is approximately **65.9%**.

**Question 4**

Class imbalance occurs when one class is significantly more frequent than the other(s). This can have a significant impact on the **confusion matrix** and derived metrics:

1. **Accuracy can be misleading**: In imbalanced datasets, a high accuracy can be achieved by predicting the majority class most of the time, even if the model performs poorly on the minority class. For example, if 95% of cases are benign and the model always predicts "benign," it will have 95% accuracy but fail to detect malignant cases.
2. **Precision and recall**: For the minority class, **recall** (True Positive Rate) may be low, even if accuracy is high. **Precision** might be decent if few positive predictions are made, but it doesn’t reflect the model's overall effectiveness if the class is underrepresented.
3. **F1-score**: Can offer a better measure for imbalanced data by balancing precision and recall, but it too can be sensitive to class imbalance.
4. **ROC-AUC**: This metric can sometimes remain high despite poor performance on the minority class due to imbalance, as the ROC curve can still be dominated by the majority class.

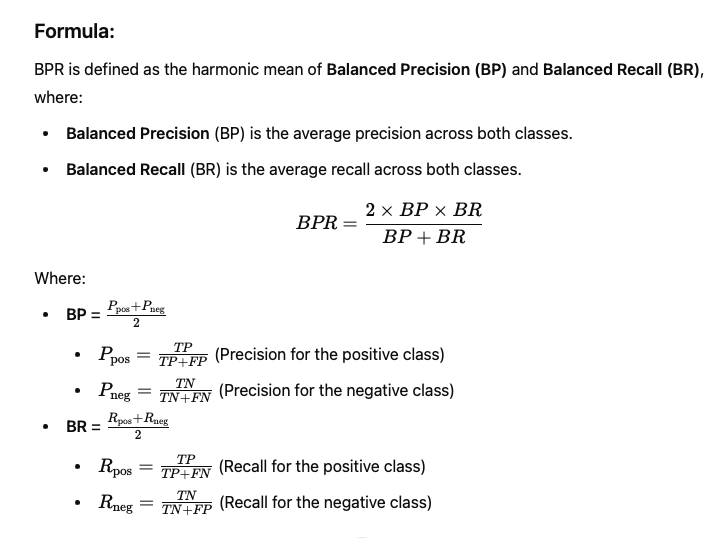
**Question 5**

**New Metric: Balanced Performance Ratio (BPR)**

**Definition:**

The **Balanced Performance Ratio (BPR)** is a new metric that addresses class imbalance by combining precision and recall for both classes (positive and negative) into a single interpretable value. It takes into account the true performance across both classes, unlike traditional metrics like accuracy that can be skewed by class imbalance.

**Formula:**

****

**Question 6**

**Comprehensive Model Evaluation Framework:**

1. **Outer Cross-Validation (K-fold):**
   * **Goal**: Test generalization performance.
   * **Process**: Split the data into **K** folds (e.g., K=5), train on K-1 folds, test on the remaining fold, repeat K times.
   * **Output**: Average performance across folds.
2. **Nested Cross-Validation:**
   * **Goal**: Avoid overfitting during hyperparameter tuning.
   * **Process**: Inside each fold of outer cross-validation, perform **inner cross-validation** (K-fold or grid search) to tune hyperparameters.
   * **Output**: Final model with optimized parameters for each outer fold.
3. **Leave-One-Out Cross-Validation (LOO-CV):**
   * **Goal**: Minimize bias when data is scarce.
   * **Process**: Train on all but one data point, test on that point, repeat for all data points.
   * **Usage**: Use LOO-CV to validate the final selected model to ensure robustness.
4. **Evaluation Metrics**: Use metrics like **F1-score, Precision-Recall AUC, Accuracy**, depending on the dataset and goals.

**Question 7**

**Experiment Design:**

1. **Dataset Selection**: Choose diverse datasets (e.g., balanced, imbalanced, large, small) to evaluate performance across various conditions.
2. **Algorithms**:
   * **Models**: SVM, Random Forest, Neural Networks.
   * **Hyperparameter Tuning**: Perform grid search or random search for each algorithm.
3. **K-Fold Cross-Validation**:
   * **Process**: For each dataset, use k-fold cross-validation (varying **k** from 5, 10, to LOO-CV).
   * **Metrics**: Collect metrics (accuracy, F1-score, AUC) for each algorithm.
4. **Experiment Steps**:
   * Train and evaluate each model on each dataset using different **k** values.
   * Average the results across all folds for each **k** and model.

**Impact of Varying K:**

* **Low k (e.g., k=5)**:
  + **Advantage**: Faster computation.
  + **Disadvantage**: More variance in results; potentially less reliable.
* **Higher k (e.g., k=10)**:
  + **Advantage**: Reduced variance in evaluation; more reliable estimates.
  + **Disadvantage**: Slower training; potential overfitting in small datasets.
* **LOO-CV**:
  + **Advantage**: Maximizes data usage for training.
  + **Disadvantage**: Computationally expensive, especially for large datasets.

**Question 8**

**Experiment Design:**

1. **Dataset**: Select a binary classification dataset (e.g., detecting fraud, tumor classification).
2. **Model**: Train a binary classification model (e.g., Logistic Regression, SVM).
3. **Confusion Matrix**: Generate a confusion matrix by comparing predicted and actual labels at varying classification thresholds (e.g., 0.1, 0.5, 0.9).
4. **Metrics**:
   * **Precision**:
   * **Recall (Sensitivity)**:
   * **F1-Score**: Harmonic mean of precision and recall.
   * **Specificity**:
5. **Steps**:
   * Set varying thresholds (e.g., 0.1, 0.3, 0.5, 0.7, 0.9).
   * For each threshold, compute the confusion matrix and derive metrics (precision, recall, F1-score).

**Insights from Metrics:**

* **Precision**: High precision means fewer false positives (useful when false positives are costly, e.g., in fraud detection).
* **Recall**: High recall indicates fewer false negatives (critical in medical diagnoses where missing positive cases is risky).
* **F1-Score**: Balances precision and recall, useful when there's a tradeoff between the two.
* **Threshold Effect**: Lowering the threshold increases recall but decreases precision, while raising it increases precision but reduces recall.

**Question 9**

Here's a comparison of **confusion matrices**, **ROC curves**, and **precision-recall curves** for evaluating the performance of multi-class classification models:

| **Feature** | **Confusion Matrix** | **ROC Curves** | **Precision-Recall Curves** |
| --- | --- | --- | --- |
| **Overview** | Displays counts of true positives, false positives, true negatives, and false negatives for each class. | Graphs true positive rate against false positive rate at various thresholds. | Graphs precision against recall at various thresholds. |
| **Threshold Dependence** | No threshold; reflects direct predictions. | Sensitive to threshold; can vary widely based on the chosen threshold. | Also sensitive to threshold; can show how changes impact precision and recall. |
| **Multi-Class Capability** | Can show per-class performance, but becomes cumbersome as the number of classes increases. | Can extend to multi-class with techniques like One-vs-Rest, but interpretation is complex. | Works better with binary classification; multi-class requires aggregation methods. |
| **Computation** | Simple to compute and visualize. | Requires calculation of TPR and FPR; may be computationally intensive for multi-class. | Requires calculation of precision and recall at multiple thresholds; less intensive than ROC. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Question 10**

**Ethical Implications of Training Data Collection for Facial Recognition**

1. **Bias**:
   * **Challenge**: Facial recognition models can exhibit biases based on race, gender, and age, leading to misidentification and discrimination against underrepresented groups.
   * **Implication**: Biased systems can perpetuate systemic inequalities and erode trust in technology.
2. **Diversity Representation**:
   * **Challenge**: Training datasets often lack diverse representation, resulting in models that perform poorly on certain demographics.
   * **Implication**: Insufficient diversity in training data can lead to ineffective and harmful outcomes for marginalized populations.
3. **Privacy Considerations**:
   * **Challenge**: Collecting facial data raises concerns about consent and surveillance, particularly when individuals are unaware their images are being used for training.
   * **Implication**: Violating privacy can undermine individual rights and lead to misuse of personal data.

**Strategies for Enhancing Fairness and Accountability**

1. **Diverse Data Collection**:
   * Ensure training datasets include a wide range of demographics to reflect the population accurately.
   * Use data augmentation techniques to enhance representation of underrepresented groups.
2. **Bias Audits**:
   * Conduct regular audits and assessments for bias in models, employing fairness metrics to evaluate performance across different demographics.
   * Implement feedback loops to continuously improve model fairness.
3. **Transparent Practices**:
   * Maintain transparency in data collection methods, ensuring clear documentation of sources, consent, and labeling processes.
   * Engage stakeholders, including community representatives, in discussions about data practices.
4. **Privacy-Enhancing Technologies**:
   * Use techniques like differential privacy during data collection and training to protect individual identities.
   * Establish stringent data governance policies to control access and usage of facial data.
5. **Ethical Guidelines and Regulations**:
   * Develop and adhere to ethical guidelines governing the use of facial recognition technology.
   * Advocate for regulatory frameworks that promote accountability and protect individuals’ rights.