Student Activity: KNN Classification & Model Evaluation (40 Minutes)

Scenario:
You are a data scientist at a renowned winery, where you are tasked with predicting wine quality based on various chemical properties. Your goal is to develop a classification model, evaluate its performance, and derive meaningful insights that can help winemakers improve wine quality.

Part 1: Data Exploration & Preprocessing (10 min)
- Load the dataset and inspect its features and target classes.
- Identify any data imbalances or patterns in the feature distribution.
- Apply Z-score normalization (Standardization) to ensure fair distance calculations.
- Split the dataset into 80% training & 20% testing.

Think & Answer:
- What story does the dataset tell about different wine quality classes?
- Are there certain chemical components that appear more frequently in high-quality wines?
- If a dataset has imbalanced class distribution, how might this affect our model's predictions?
- Imagine two features: alcohol content ranging from 8 to 15 and citric acid ranging from 0.1 to 0.8. How could this impact distance calculations in KNN if unscaled?
- How does train-test splitting ensure fairness in model evaluation?

Part 2: Model Building & Optimization (15 min)

- Train a KNN model with different `K` values (try `K=1 to 20`).
- Use cross-validation to determine the best `K`.
- Predict wine quality on the test set.
- Compute the confusion matrix, classification report, and accuracy score.

Think & Answer:
- Suppose you train a KNN model with `K=1`. How might the model behave?
- If `K` is too large, what kind of bias might the model develop?
- What does the best `K` value tell us about decision boundaries?

- You observe that the model has high precision but low recall for certain wine classes. What does this mean?
- How could a winemaker use this information to make decisions about wine quality classification?

Part 3: Model Performance Insights (15 min)
- Convert the multiclass labels into a binary format using `label_binarize()`.
- Plot ROC curves and compute AUC for each class.

Think & Answer:
- If a class has an AUC score close to 0.5, what does this indicate about the model's ability to classify it?
- Why might some classes have a higher AUC score than others?
- What external factors could make certain wine qualities harder to predict?

- Suppose you discover that alcohol content is a strong predictor of high-quality wine. If you were a winemaker, how could you use this insight?
- Can any of these features be manipulated to improve wine quality, or are they naturally occurring?
- If you had access to more data, what additional features would you want to include in this analysis?