Advanced Ensemble Methods

Assignment 1: Basic - Voting Ensemble for AI Engineer Specialization

Context: One Hour AI Solution platform needs to classify which AI specialization area would be most suitable for new engineers joining the platform based on their skills and background.

Task: Implement a basic voting classifier ensemble that combines three different models to predict the most suitable AI specialization area.

Dataset: engineer_skills.csv

Requirements:

1. Split data into features and target variables
2. Create three different classifier models from the ones covered in class:
    - Decision Tree
    - Logistic Regression
    - K-Nearest Neighbors
3. Implement a voting classifier to combine these models
4. Evaluate the performance using accuracy and a confusion matrix
5. Compare the voting ensemble performance with individual models
6. Write a brief summary of your findings (max 5 sentences)

Assignment 2: Intermediate - XGBoost for Engineer Performance Prediction

Context: One Hour AI Solution wants to predict how well an engineer will perform on different types of AI challenges based on their past performance and attributes.

Task: Implement an XGBoost model to predict engineer performance scores and analyze feature importance.

Dataset: engineer_performance.csv

Requirements:

1. Preprocess the data (handle any needed transformations, split into train/test)
2. Implement XGBoost regressor to predict performance_score
3. Tune at least 3 hyperparameters to improve model performance
4. Calculate the feature importance and visualize the top 5 most important features
5. Create at least one interaction feature from existing features and evaluate if it improves the model
6. Compare the performance with a simple decision tree model
7. Write a brief analysis of your findings, particularly focusing on what factors most influence engineer performance

Assignment 3: Advanced - Ensemble Stacking for Session Success Prediction

Context: One Hour AI Solution wants to predict whether an AI problem-solving session will be completely successful (solving the client's problem within the one-hour timeframe) based on various factors.

Task: Build a stacking ensemble model to predict session success probability.

Dataset: session_success.csv

Requirements:

1. Perform data preprocessing:
    - Convert categorical variables appropriately
    - Create any additional features that might be relevant
    - Split data into training and testing sets
2. Create a stacking ensemble:
    - First layer: Implement at least 3 different models from the ones covered (Random Forest, Gradient Boosting, and another of your choice)
    - Second layer: Use a logistic regression model as the meta-learner
3. Implement proper cross-validation to prevent data leakage during stacking
4. Evaluate model performance using appropriate classification metrics (accuracy, precision, recall, F1-score)
5. Compare the stacking ensemble with:
    - Individual base models
    - A simple voting ensemble of the same base models
6. Analyze which factors most strongly predict session success and failure
7. Create a brief report explaining your approach, results, and recommendations for improving session success rates