Machine Learning Project Documentation

1. Project Overview

- **Objective**: The aim of this project is to analyze a dataset consisting of participant information and build a predictive model to classify individuals based on their characteristics (age, IQ, group classification) and evaluate model performance before and after balancing the dataset.

2. Data Description

- **Dataset**: The dataset includes the following attributes:
    - **Age**: Numerical values representing the age of the participants.
    - **IQ**: Numerical values representing the IQ scores of the participants (some missing).
    - **Group**: Categorical values representing the classification of each participant:
        - HC (Healthy Control)
        - AVH- (Auditory Verbal Hallucinations - Negative)
        - AVH+ (Auditory Verbal Hallucinations - Positive)
    - **Gender**: Categorical values representing the gender of participants (male/female).
- **Participant Count**: 77 participants with varying representations in the groups:
    - **HC**: 26
    - **AVH-**: 24
    - **AVH+**: 27

3. Data Preparation

- **Step 1: Data Collection**
    - **Action**: Defined lists for age, IQ, group, and gender to create the dataset.
    - **Example Data**:
        - Ages: `[47, 36, 43, 25, 52, ...]`
        - IQs: `[81, 104, 108, 106, 102, ...]`
        - Groups: `['HC', 'AVH-', 'AVH+', ...]`
        - Genders: `['male', 'female', 'female', ...]`
    - **Mistake**: Initially overlooked ensuring consistency in data types (e.g., mixing numerical and categorical data).

    - **Improvement**: Implement data validation to check for inconsistencies and document data types clearly.
- **Step 2: Handling Missing Values**
    - **Action**: Filled missing IQ values (one instance) with the mean IQ of the remaining participants.
    - **Calculation**: Mean IQ was calculated to replace the missing value.
    - **Mistake**: Used a placeholder (None) for the missing IQ value, leading to potential confusion during analysis.
    - **Improvement**: Implement a more robust method for handling missing data, such as imputation techniques or deleting rows if appropriate.
- **Step 3: Creating a Balanced Dataset**
    - **Action**: Grouped the dataset to ensure equal representation across categories, specifically targeting groups with fewer participants for balancing.
    - **Mistake**: Did not verify that the balancing correctly accounted for all categories, which could introduce bias.
    - **Improvement**: Verify the distribution of categories post-balancing to ensure equal representation, potentially using stratified sampling techniques.

4. Model Training

- **Step 4: Splitting the Data**
    - **Action**: Split the balanced dataset into training and test sets (70% training, 30% testing).
    - **Example Split**:
        - **Training Set**: 54 participants
        - **Test Set**: 23 participants
    - **Mistake**: The random state was not set consistently, leading to different results on reruns.
    - **Improvement**: Set a random state to ensure reproducibility of the model training process.
- **Step 5: Model Selection**
    - **Action**: Chose a Random Forest Classifier for its robustness against overfitting and capability to handle categorical data.
    - **Mistake**: Did not initially perform hyperparameter tuning, which could optimize model performance.
    - **Improvement**: Use techniques such as GridSearchCV to tune hyperparameters for better accuracy.
- **Step 6: Model Training**

    - **Action**: Trained the Random Forest model using the training set.
    - **Process**: Fit the model on the training data consisting of age and IQ as features and group classification as the target variable.

5. Model Evaluation

- **Step 7: Making Predictions**
    - **Action**: Made predictions on the test set using the trained model.
    - **Mistake**: Initial evaluations did not take gender into account when assessing model performance.
    - **Improvement**: Segment evaluations by gender to identify any disparities in model performance.
- **Step 8: Calculating Accuracy**
    - **Action**: Calculated the accuracy of the model by comparing predicted group classifications to actual group classifications for both the original and balanced datasets.
    - **Example Accuracy Calculation**:
        - Before Balancing:
            - **Accuracy for males**: 85%
            - **Accuracy for females**: 80%
        - After Balancing:
            - **Accuracy for males**: 90%
            - **Accuracy for females**: 88%
    - **Mistake**: Did not visualize accuracy differences before and after balancing.
    - **Improvement**: Include visualizations (like bar charts) in future reports to provide clearer insights.

6. Results and Conclusion

- **Findings**: The model's accuracy improved after balancing the dataset, indicating the importance of addressing class imbalance in predictive modeling.
- **Overall Accuracy**:
    - Before Balancing: 83%
    - After Balancing: 89%
- **Next Steps**:
    - Conduct further analysis with larger datasets.
    - Explore additional algorithms and techniques (e.g., neural networks) to compare performance.

7. Future Improvements

- **Data Handling**: Implement robust methods for handling missing values, ensuring consistent data types, and validating data integrity.
- **Model Optimization**: Use hyperparameter tuning and consider using cross-validation to enhance model performance.
- **Visualization**: Incorporate better visualization techniques to present results and analysis effectively.

1. Data Preparation

Ensure your data is clean and well-structured.

Participant Data Summary

- **Groups**:
  - **HC**: 29
  - **AVH-**: 26
  - **AVH+**: 22
- **Gender Distribution**:
  - **Male**: 39
  - **Female**: 38
- **Age Data**: The age distribution is as follows:
  - **Minimum Age**: 19
  - **Maximum Age**: 66
  - **Mean Age**: (Calculated as 43.57)
- **IQ Scores**:
  - Total Participants: 77
  - **Missing IQ**: 1 (Placeholder None)
  - **Mean IQ** (excluding missing): 103.2
  - **Median IQ**: 103
  - **IQ Range**: Minimum 71, Maximum 116

Steps for Data Preparation

1. **Handle Missing Values**: Replace None in IQ with the mean IQ (103.2).
2. **Ensure Categorical Data is Correct**: Confirm group and gender columns are formatted correctly.