DA5401: A2 Dimensionality Reduction, Visualization, and Classification Performance

Objective: This assignment will challenge you to apply concepts of vector spaces, dimensionality reduction, and feature engineering to a real-world classification problem. You will use Principal Component Analysis (PCA) to reduce the data's dimensionality and then evaluate how this reduction affects the performance of a Logistic Regression classifier.

1. Problem Statement

You have been given the **Mushroom Dataset**, which contains various attributes of mushrooms, classified as either edible or poisonous. This dataset is an excellent example of a high-dimensional, categorical dataset with significant feature redundancy. Your task is to apply PCA to this dataset and evaluate its effectiveness by comparing the performance of a Logistic Regression classifier trained on the original data versus the PCA-transformed data.

You will submit a Jupyter Notebook with your complete code, visualizations, and a plausible story that explains your findings. The notebook should be well-commented, reproducible, and easy to follow.

2. Tasks

Part A: Exploratory Data Analysis (EDA) & Preprocessing [10 points]

1. **Load and Prepare the Dataset: [3]**
    - Load the Mushroom dataset. Note that it is composed entirely of categorical features. The target variable is the 'class' column (`e` for edible, `p` for poisonous).
    - Perform **one-hot encoding** on all categorical features. Explain why this is a necessary step before applying PCA.
    - Separate the features (X) from the target variable (y).
2. **Initial Analysis: [2]** Print the dimensions of the dataset after one-hot encoding. Observe the significant increase in the number of features.
3. **Standardization [5]:** Explain why **standardizing** the one-hot encoded features is still a good practice before PCA, even though they are binary. Implement `StandardScaler` from scikit-learn.

Part B: Principal Component Analysis (PCA) [20 points]

Ensure you build a plausible story alongside the visualizations.

1. **Apply PCA [2]:** Perform PCA on the standardized, one-hot encoded dataset. Do not specify the number of components initially.
2. **Scree Plot & Optimal Components [8]:**
    - Create a **scree plot** showing the explained variance ratio and the cumulative explained variance ratio.

    - Determine the optimal number of principal components to retain. Justify your choice based on the scree plot (e.g., aiming to retain 95% of the variance).
3. **Visualization [10]:**
    - Project the data onto the first two principal components. Create a 2D scatter plot, using different colors to distinguish between edible and poisonous mushrooms.
    - If you have more than 2 PC dimensions, visualize other pair plots.
    - Discuss what this visualization tells you about the separability of the two classes in the new, reduced feature space.

Part C: Performance Evaluation with Logistic Regression [20 points]

1. **Baseline Model [5]:**
    - Split your **original, standardized** data into training and testing sets.
    - Train a **Logistic Regression** classifier on the training data.
    - Evaluate its performance on the test set using a classification report (including precision, recall, and F1-score) and the classification accuracy.
2. **PCA-Transformed Model [5]:**
    - Transform both your training and testing sets using the optimal number of components you determined in Part B.
    - Train a new **Logistic Regression** classifier on the PCA-transformed training data.
    - Evaluate its performance on the PCA-transformed test data using a classification report and accuracy.
3. **Comparison and Analysis [10]:**
    - Compare the performance metrics of the two models (the one on original data and the one on PCA-transformed data).
    - Is there a significant difference in performance? Explain why or why not, considering the trade-off between dimensionality reduction and information loss. Did PCA's ability to handle feature collinearity and redundancy provide a performance benefit?
    - Discuss the usefulness of using Logistic Regression as a **surrogate performance measurement** for evaluating the effectiveness of PCA.

---

3. Submission Guidelines

- Submit a single Jupyter Notebook with all your code, visualizations, storytelling, and answers to the conceptual questions.
- Your notebook should be self-contained and run without errors. Use markdown cells to structure your answers and explanations.
- Ensure your code is clean, readable, and well-commented.

Evaluation Criteria:

- Correct implementation of one-hot encoding, standardization, PCA, and Logistic Regression.
- Quality and clarity of visualizations.
- Storytelling and plausible narratives.
- Insightful analysis and interpretation of the results, especially the performance comparison.
- Demonstrated understanding of the conceptual links between vector spaces, PCA, and model performance.

Good luck!