COMP3401 Assignment 2

1. Dimension Reduction [50]

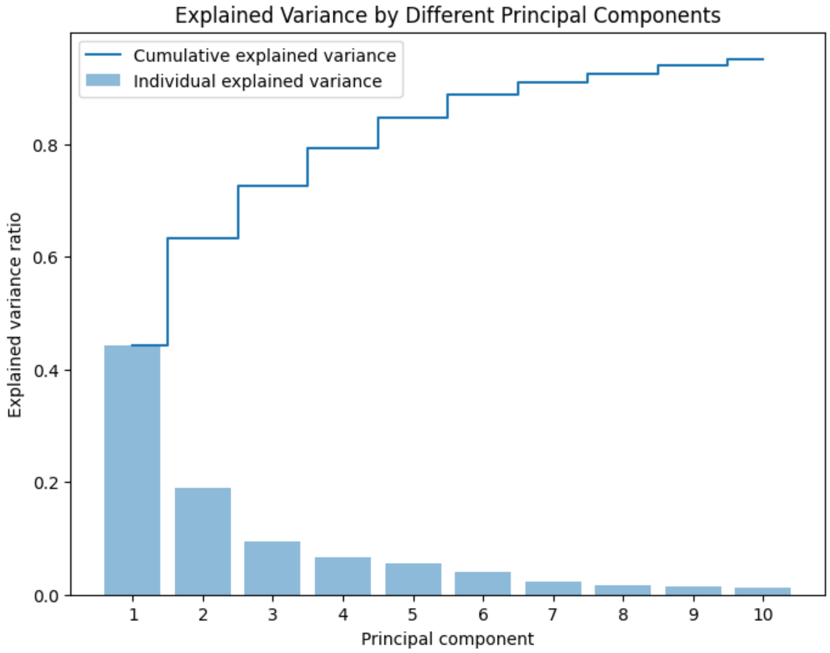
Exercise 1.1: Principal Component Analysis (PCA) on Cancer Data [25]

Tip: How PCA works video

Objective: Apply PCA to the cancer dataset and visualize the separation of classes based on the principal components.

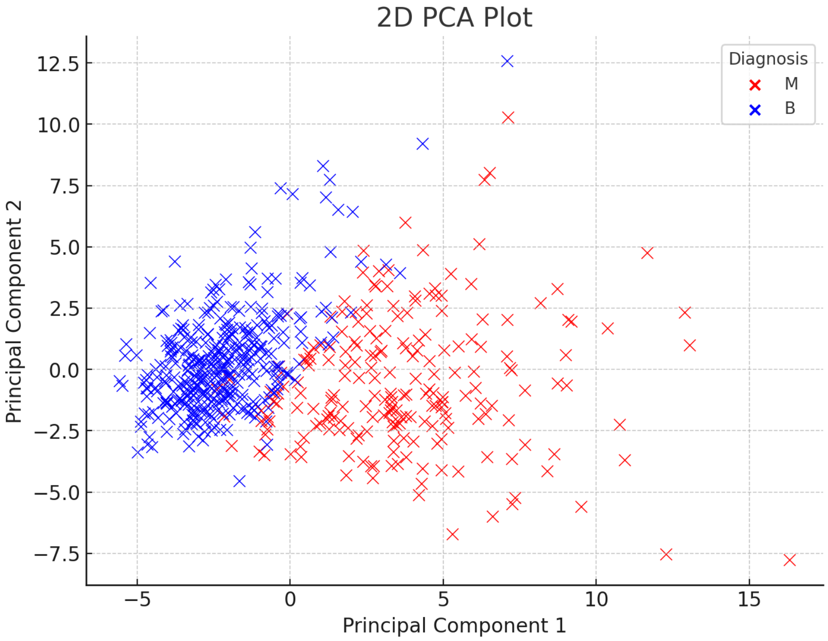
Instructions:

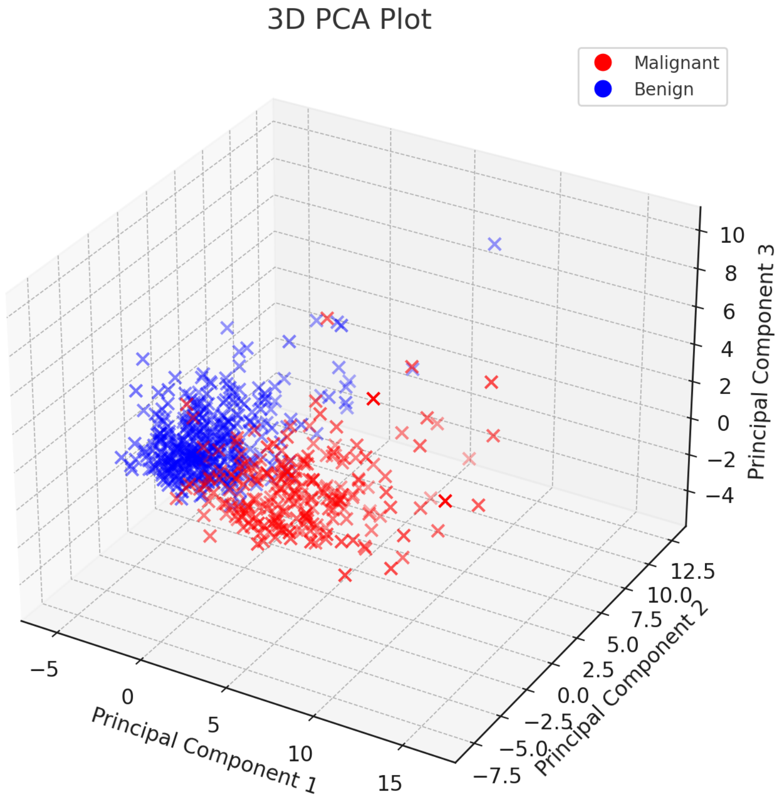
1. **Data Preparation:**
    - Load the cancer dataset.
    - Ensure that you use only numerical data for PCA.
    - Scale the data so that all features have a similar range.
2. **PCA Application:**
    - Perform PCA on the prepared data.
3. **Variance Analysis:**
    - Calculate and display the amount of variance captured by each of the selected principal components.
    - Present the cumulative variance explained by the principal components. See image below for reference.
    - Hint: The `explained_variance_ratio_` property has to be used here.



4. **Visualization:**

- Create a 3D scatter plot using the first three principal components as axes.
- Use the diagnosis as the hue to differentiate between the classes.
- *Optionally*, create a 2D scatter plot using the first two principal components as axes with diagnosis as the hue, as shown in the provided examples. Try to make the 3D plot since you can show more information this way but alternatively, you can make the 2D plot as well without any loss of marks.





5. **Interpretation:**

- Based on the visualizations, do you believe a predictive model could be developed to distinguish between malignant and benign tumors with a reasonable degree of accuracy? Please justify your response.

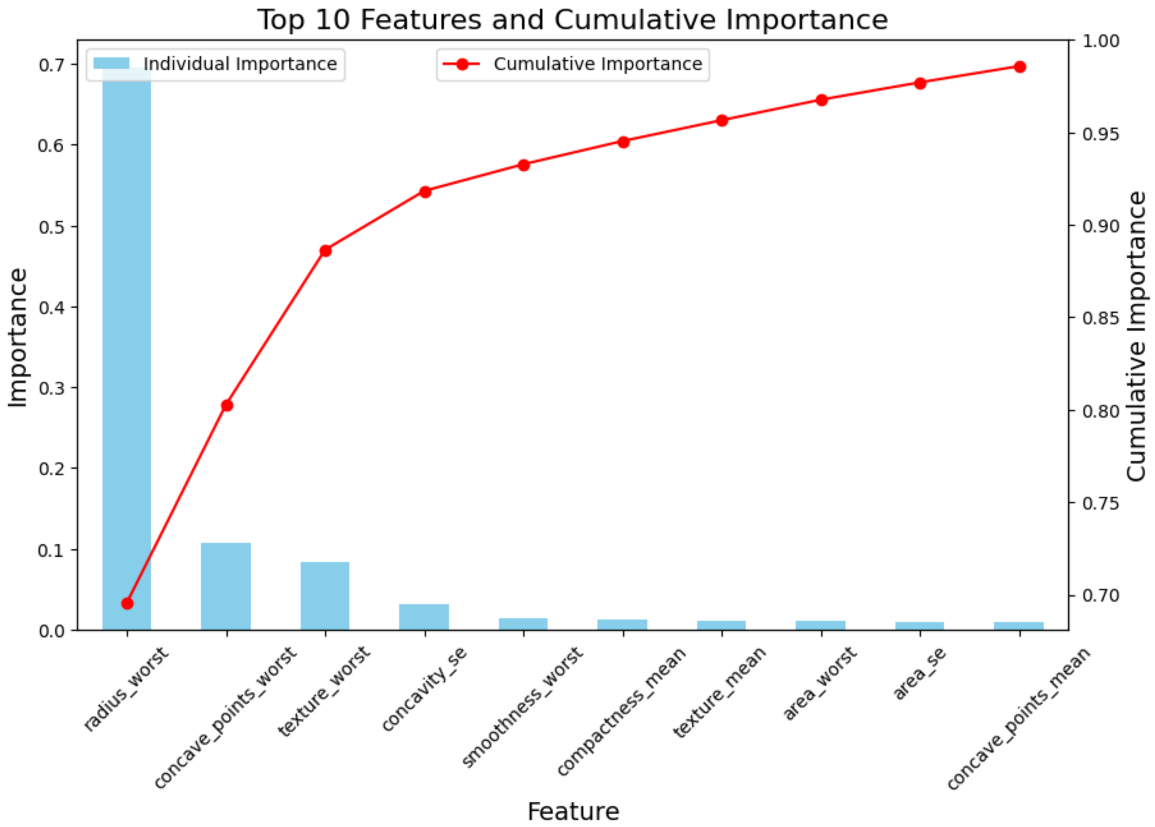
Exercise 1.2: Determining Feature Importance with Decision Trees [25]

Recommended videos (for understanding and smiles)

1. How decision trees work
2. How to use decision trees to select the most important features

Objective: Identify the most significant features for classifying tumors as malignant or benign using a decision tree approach.

1. **Model Construction:**
   - Develop a decision tree classifier. Refer to the 'Feature Selection Notebook' on the Abalone dataset available on D2L for guidance. Although it demonstrates a decision tree regressor for regression tasks, the methodology is analogous and will be beneficial in understanding how to implement a decision tree for classification purposes.
   - Train the model with the dataset provided, focusing on tumor classification.
2. **Feature Importance Evaluation:**
   - Utilize the `feature_importances_` attribute of the decision tree to assess the importance of each feature.
   - Arrange the features in descending order of importance.
3. **Visualization:**
   - Create a bar chart to display the individual importance of the top 10 features.
   - Overlay this with a line graph showing the cumulative importance of these features.
   - Your chart should resemble the example provided in the figure. I made this chart using the `pandas plot` method.
4. **Analysis:**
   - Discuss the results indicated by the feature importance graph.
   - Suggest how these insights could impact the development of more accurate predictive models for tumor classification.



2. Pattern Mining [50]

Exercise 2.1 [25]

A dataset consists of 5 separate transactions. Define minimum support (min_sup) as 60% and minimum confidence (min_conf) as 80%.

The following table presents the transaction ID (TID) alongside the respective items purchased:

TID	items_purchased
T1	{A, B, C, D, E, F}
T2	{G, B, C, D, E, F}
T3	{A, H, D, E}
T4	{A, I, J, D, F}
T5	{J, B, B, D, K, E}

Proceed with the following tasks:

Exercise 2.1.1 [20]

Identify all frequent itemsets by applying both the Apriori and the FP-growth algorithms.

- Solve the problem using a method analogous to the one demonstrated in Example 4.3 for the Apriori algorithm. Include a diagram akin to Figure 4.2 on page 152 and provide all the necessary calculations and steps. [20]

- For the FPGrowth algorithm, refer to Example 4.5 as your guide. Include a representation of the tree similar to the one in Figure 4.7, and construct a table comparable to Table 4.2. Ensure that all the required work is clearly shown. [20]

Exercise 2.1.2 [5]

List all the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g., `A`, `B`):

$$\forall X \in \text{transaction},\ \text{buys}(X, \text{item}_1) \land \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3)\ [s, c]$$

Question 2: Use Apriori to mine for frequent itemsets. [25]

Open the starter.ipynb file provided and write code in the cells that have the `"# Your Code Here"` comment

Submission Instructions

Please follow these instructions for your assignment submissions:

- For Exercises 1.1, 1.2, and 2.2, complete your work in Jupyter Notebooks.
- For Exercise 2.1, you may use MS Word or a similar application and save it as a pdf.
- After completing the exercises, convert each Jupyter Notebook to a PDF file.
- Combine these four PDFs into a single document in the correct order.
- Submit this consolidated PDF along with the original Jupyter Notebook files to the designated Dropbox.

Important: Ensure that the start of each question is clearly indicated in your submissions. Failure to do so will result in a deduction of marks.