1. **Project Overview:**

This project carries out the following tasks on a breast cancer dataset that is available on sklearn.datsets:

* Load and convert the raw data into a structured dataframe
* Standardize the dataframe to have all features on the same scale and improve mathematical evaluations
* Use Principal Component Analysis (PCA) to transform the original features to uncorrelated variables; these are 30 components
* Apply PCA to reduce the 30 components into 2 PCA components
* Save raw and transformed data to external Excel files
* Train a Logistics Regression model using data that was reduced by PCA
* Predict target classes on the test data
* Evaluate the performance of the model and save the prediction results in an Excel file

1. **Perquisites:** Tools used

|  |  |
| --- | --- |
| **Tool** | **Purpose** |
| Excel | Visual review of datasets (raw, transformed, predictions) |
| pandas | Functionalities for reading, manipulating and exporting data |
| numpy | Powers numerical operations and array handling |
| matplotlib | For creating standard plots |
| sklearn | Supplies data, tools to prepare and analyze the data, and algorithms to build and evaluate predictive model |
| openpyxl | For reading and writing Excel |

1. **Instructions:**
2. Packages including openpyxl, scikit-learn, matplotlib, numpy, pandas have to be installed manually as they are not built-in on Python
3. The code develops and exports files to Desktop directory; use discretion to edit underlisted code lines to alternative directories (or file names) that may be preferred:

|  |  |
| --- | --- |
| **Code Line** | **Syntax** |
| Line 22 | output\_path = "C:/Users/HP/Desktop/Raw\_Breast\_Cancer\_Data.xlsx" |
| Line 39 | loadings\_path = "C:/Users/HP/Desktop/PCA\_Loadings.xlsx" |
| Line 52 | output\_path = "C:/Users/HP/Desktop/PCA\_Results.xlsx" |
| Line 79 | prediction\_output\_path = "C:/Users/HP/Desktop/PCA\_Prediction\_Results.xlsx" |
| Line 109 | plot\_path = "C:/Users/HP/Desktop/PCA\_Cancer\_Plot.png" |

1. **Code Features**:
2. Data loading and conversion:

Load data, converts it to dataframe and save raw dataset to external Excel file

1. PCA operations:

* Converts raw data to essential components
* Reduce components to 2 PCA components
* Save both sets of datasets above to separate external Excel files

1. Logistic Regression:

* Splits the PCA-transformed data into train and test sets (80/20 split)
* Train a logistic regression model on the training set
* Predicts the target classes on the test set

1. Model Evaluation

* Evaluates the model using:
  + Accuracy
  + Confusion Matrix
  + Classification Report

1. Save Predictions to an external Excel file