Assignment #9

Assignment Overview

In this assignment you will demonstrate your knowledge of building a classifier for the Pima Indians diabetes dataset.

Background

The Pima Indians diabetes dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The dataset consists of 8 medical predictor variables (the number of pregnancies the patient has had, glucose level, blood pressure, skin thickness, BMI, insulin level, age, and a "diabetes pedigree function") and one target variable, *Outcome*. It contains 768 rows.

See: https://www.kaggle.com/uciml/pima-indians-diabetes-database and https://git.io/J8zj1 for more.

Project Specification

This is a group assignment.

Students are encouraged (but not required) to work in groups of max 3 students. Please follow the same rules as stated in previous assignments.

Requirements

In this assignment you will follow the same steps as in Chapter 10 of the textbook, with the obvious exception of the problem/dataset.

You are required to:

1. Implement and test the 5 main functions and 4 main data structures from Chapter 10 (with the proper changes, explained in the notebook itself).
2. Organize your solution / notebook into meaningful cells, using headings, and with proper text, code, plots, figures, links, etc.
3. Prepare a **Conclusion** cell with a summary of your insights and lessons learned.

You don't need to show the intermediate steps in the development.

The final solution will be similar to Code Listing 10.15 (which can be used as a starting point), but adapted to the problem at hand and to a Notebook format.

You are allowed to use matplotlib, seaborn, pylab, or any other plotting library for Python.

You are **NOT** allowed to use pandas, scikit-learn, or any other "data science" library for Python.

Deliverables

You must submit (via Canvas):

- The **link**[1] to a Jupyter notebook on Google Colab containing your entire solution. It must include:
    - Header:
        - Team members' names, date, course name + code, assignment number
    - Your source code
    - Results (of multiple runs) + meaningful comments
    - Plots
    - Figures
    - References (including your "sources of inspiration" for the code)
    - Comments (README-like): installation instructions, dependencies, etc.
    - Project notes (describing what my TA and I cannot see by looking at your source code and/or running your program).
        - Examples: design decisions, documented limitations, future improvements, etc.
    - Your **Conclusion** with a summary of your insights and lessons learned.

Bonus opportunities:

This is an odd-numbered assignment.
There are <u>no</u> bonus opportunities (unless, of course, you guess my zoom background in related lectures). ☺

[1] When sharing the link to your Google Colab notebook, choose the 'anyone with the link can open it' option, i.e., **don't make it specific to a domain** (such as fau.edu) **or individual** (instructor or TA).