به نام خدا

دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)
دانشکده برق
یادگیری ماشین – نیم‌سال دوم 1401-1402

تمرین عملی بیزین، سری دوم، درس یادگیری ماشین

Code for all programming assignments should be well documented. **A working program with no comments will receive only partial credit**. Documentation entails writing a description of each function/method, class/structure, as well as **comments throughout the code** to explain the program flow. Programming language for the assignment is **Python**.

Following libraries can be used when necessary:

- Matplotlib, NumPy, SciPy, Pandas and other basic libraries.
- libraries for calculating mean, variance and covariance.

Following libraries **mustn't** be used in any way:

- Libraries for making classifiers (You must implement classifiers from scratch).

Collaboration Policy

You are to complete this assignment individually. However, you may discuss the general algorithms and ideas with classmates, TAs, peer mentors and instructors in order to help you answer the questions. But we require you to:

• not explicitly tell each other the answers
• not to copy answers or code fragments from anyone or anywhere
• not allow your answers to be copied
• not to get any code from the Web

If you have any questions regarding this assignment, please contact Mr. Alireza Rahmati.

Telegram ID: @AliReza_AR

Submit by 18th Shahrivar, 1402, 11.59pm.

In this project, you must detect diabetes people. There is a dataset (link). This dataset has 8 features in 2 classes. The Bayesian analysis must be Implemented for this dataset.

- **Data vitalization and Preprocessing:**
  - Visualize the relationship between variables using scatter plot (According to Figure 1 in the appendix).
  - Split dataset for Train and Test, 70% and 30% of data respectively.
- **Modeling:**
  - Implement Gaussian Naïve Bayes classifier by Train dataset.
- **Evaluation:**
  - Evaluate the model by Train and Test dataset.
  - Evaluate the model based on 4-fold cross validation.
  - Polt two confusion matrix for the classifier by Train and Test dataset.
- **Model efficiency improvement (Extra credit):**
  - Select the best four features according to the Data vitalization part then make a new dataset.
  - Say, why these features are the best.
  - Retrain the model with the new dataset.
  - Evaluate the new model with the new dataset.
- **Apply thresholding method (Extra credit):**
  - Add a parameter to the model for adjusting the classifier.
  - Set this parameter for detecting diabetes when the probability of diabetes is more than 40%.

What to Submit:

• Code with comments.

• A short write-up about your implementation with results and your observations from each result.

Good luck 😊

Appendix:

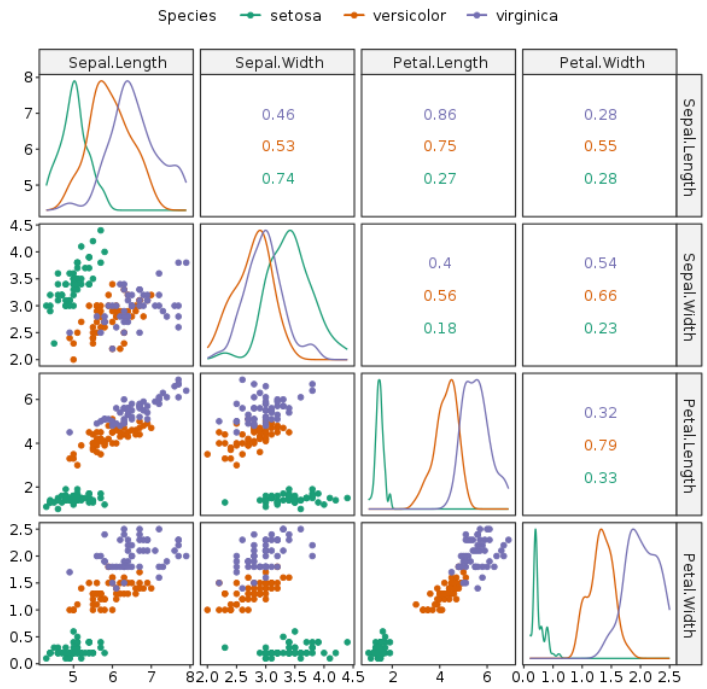

Figure 1: Scotter plots and line charts are used in descriptive statistics to show the relationship between Iris dataset features.