COSC 4335 *“Data Mining”* Fall 2018

Assignment1: (Exploratory) Data Analysis

for the Pima Indians Diabetes Dataset

Individual Assignment



Due: Sunday, Sep 16, midnight (electronic Submission)

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using Python 2 or 3 & Jupyter notebook)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, domain knowledge, and also practical experience.

Download *the Pima Indians Diabetes[[1]](#footnote-1)* dataset

it has 8 numerical attributes and a binary class variable (1 indicates that the person is assumed to have diabetes), indicating the following information:

1. Number of times pregnant   
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test   
3. Diastolic blood pressure (mm Hg)   
4. Triceps skin fold thickness (mm)   
5. 2-Hour serum insulin (mu U/ml)   
6. Body mass index (weight in kg/(height in m)^2)   
7. Diabetes pedigree function   
8. Age (years)   
9. Class variable (0 or 1)

5 Examples in the Pima Indians Diabetes Dataset:

6,148,72,35,0,33.6,0.627,50,1

1,85,66,29,0,26.6,0.351,31,0

8,183,64,0,0,23.3,0.672,32,1

1,89,66,23,94,28.1,0.167,21,0

0,137,40,35,168,43.1,2.288,33,1

Assignment1 Tasks:

Apply the following exploratory data analysis techniques **using Python 2 or 3** to the dataset:

1. Compute the mean value and standard deviation for attributes 2-6**.** Remove 0’s that do not make sense prior to computing these statistics.
2. Compute the covariance matrix for attributes 2-6 next, compute the correlations for each of the 10 pairs of the 5 attributes. Interpret the statistical findings! Remove 0’s that do not make sense prior to computing the covariance matrix and correlations.
3. Create a scatter plot for attributes 3 and 6 of your dataset and a second scatter plot for attributes 2 and 7. Interpret the two scatter plots**!**
4. Create histograms for attributes 2, 3 and 6. Then create the same histograms for the 3 attributes for the instances of class 1 and for the instances of class 0; interpret the obtained 9 histograms.
5. Create box plots for the 2nd , 7th and 8th attribute; one for the whole dataset and one each for the instances of the two classes. Remove 0’s that do not make sense prior to computing the box plots.Interpret and compare the obtained 9 boxplots!
6. Create supervised scatter plots[[2]](#footnote-2) for all pairs of attributes 2-6—these are 10 plots. Next create two 3D-scatterplots: one for attributes 2, 3, 6 and one for attributes 2, 4, 6. Interpret the obtained scatter plots; in particular address what can be said about the difficulty in predicting diabetes. Assess the usefulness of the 3D scatterplot compared to the 2D plots!
7. Create a Star plot for the first 10 instances[[3]](#footnote-3) of class 0 and the first 10 instances of class 1 for attributes 1, 2, 3, 8 in the *cleaned dataset*. Interpret the obtained plots.
8. Fit a linear model that predicts the class attribute using the 8 z-scored, continuous attributes of the *cleaned dataset* as independent variables[[4]](#footnote-4). Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. Next,drop the two attributes, whose coefficients are the closest to 0, and obtain a linear model using the six remaining attributes as independent variable. Do the coefficients tell you anything about the importance of the attribute in predicting diabetes? What about negative and positive coefficients? Also compare the two regression functions!
9. Create 3 decision tree models with 20 or less nodes[[5]](#footnote-5) using the *cleaned dataset* (total number of nodes should be less than 21 do not submit models with more than 20 nodes!); Explain how the 3 decision tree models were obtained. Report the training accuracy and the testing accuracy for each decision tree; interpret the learnt decision tree—what do they tell you about the importance of the 8 continuous attributes for the classification problem? **12 points**
10. Write a conclusion (at most 18 sentences!) summarizing the most important findings of the assignment—what did we learn about the dataset? In particular, address the findings obtained related to predicting the class attribute.

Remarks:

* About 27-36% of the Assignment1 points will be allocated to interpreting statistical findings and visualizations!
* The dataset contains some missing attribute values (values of 0 that do not make sense); therefore, some kind of data cleaning has to be done. For task 6-8 use the *cleaned dataset*, whereas for the other tasks use the original dataset and follow the instructions given in the respective task descriptions!
* Name the submission uhid\_lastname.zip, including two files: the .ipynb file and the .pdf report.

1. We will also use a “cleaned version” of this dataset, for some tasks of Assignment1 (see below!). [↑](#footnote-ref-1)
2. Supervised density plots could also be used instead! [↑](#footnote-ref-2)
3. Based on the ordering of the examples in the original dataset. [↑](#footnote-ref-3)
4. That is, you have to transform attributes 1-8 into z-scores before fitting the linear model! [↑](#footnote-ref-4)
5. Intermediate nodes count! [↑](#footnote-ref-5)