Dr. Eick

COSC 6335 *“Data Mining”* Fall 2020 Problem Set1

(Exploratory) Data Analysis, Classification, and Peer Review

Individual Assignment[[1]](#footnote-1)

Fifth Draft

Due: Submit Task1 and 2 by Saturday, September 19, 11p

Due: Peer Reviews Task1 Thursday, September 24, 11p

Due: Submit Task3 and 4 by Monday, Sept. 28, 11p

Due: Peer Reviews Task3 by Thursday, Oct. 1, 11p

Due: Evaluation of Evaluations by Monday, October 5, 11p

Last Updated: Wednesday, September 2, 5p

Total available points for Problem Set1: 93+20+77+20=210 points

Tentative weight of Assignment1: 26-35% of the available points for the 3 problems sets

Remark: points allocated to the ProblemSet1 tasks are subject to change!

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, …
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn how to generate classification models for a given dataset
6. Learn how to evaluate and compare classification models
7. 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.
8. Peer reviewing data analysis and data mining results

1. Exploratory Data Analysis for a Pima Indian Dataset (Graded & Peer Reviewed)



Available points: 58 (+ up to 5 extra points) + 30 points for peer reviewing

Download *the Original Pima Indians Diabetes[[2]](#footnote-2)* dataset <https://www.kaggle.com/kumargh/pimaindiansdiabetescsv>

 which has 8 numerical attributes and a binary class attribute (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

Problem1 Sub-Tasks:

Apply the following exploratory data analysis techniques **using R** (or other tools of your preference) to the original/cleaned Pima Indian Diabetes Dataset; if not otherwise stated, use the original dataset:

1. Compute the mean value and standard deviation for attributes 2, 3, 6, and 7**.** Remove 0’s that do not make sense prior to computing these statistics.Compute the covariance matrix for four attributes next, compute the correlations for each of the 6 pairs of the 4 attributes. Interpret the statistical findings! Remove 0’s that do not make sense prior to computing the covariance matrix and correlations. **6 points**
2. Create histograms for attributes 2 and 4. Then create the same histograms for the 2 attributes for the instances of class 1 and for the instances of class 0; interpret the obtained 6 histograms. **9 points**
3. Create box plots for the 6nd 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 6 boxplots! **6 points**
4. Create supervised scatter plots[[3]](#footnote-3) for all pairs of attributes 1, 2, 6 and 8—these are 6 plots. Next create a single 3D-scatterplot for attributes 2, 6, 8. Interpret the obtained 7 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! **10 points**
5. Fit a linear model that predicts the class attribute using the 8 z-scored, continuous attributes of the *Cleaned Pima Indian Diabetes 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 three attributes, whose coefficients are the closest to 0, and obtain a linear model using the five remaining attributes as independent variables. 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! **11 points**
6. Create 3 decision tree models with 18 or less nodes[[5]](#footnote-5) using the *Cleaned Pima Indian Diabetes Dataset* (total number of nodes should be less than 19 do not submit models with more than 18 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 have in common, and how do they differ; what do the three decision trees tell you about the importance of the 8 continuous attributes for the classification problem? **11 points**
7. Write a conclusion (at most 13 sentences!) summarizing the most important findings of the assignment—what did we learn about the dataset? In particular, discuss the findings obtained related to predicting diabetes. **5 points (and up to 5 extra points)**

Remarks:

* You will use Kritik to peer review the solutions for tasks b-f; each student will review 15 task submissions of other students; Nour will review and grade your submissions of tasks a and g.
* About 27-38% of the task1 points will be allocated to interpreting statistical findings and visualizations!
* The original 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 f and g we provided a *cleaned dataset[[6]](#footnote-6)*, whereas for the other tasks please use the original dataset and make assessing and interpreting presence of missing values a subtask of tasks b-d[[7]](#footnote-7)!

2. Basic Data Science Matter (Not peer reviewed)

Available Points: 20

a. Interpret the supervised scatter plot depicted below; moreover, assess the difficulty of separating males from females using Factor 1 / Factor 2 based on the scatter plot! [5]





b) The boxplot depicted above has been created using the following R-code for an attribute x:

> x<-c (1,2,2,2,8,8,8,10,14,14,18,37)

> boxplot(x)

What is the median for the attribute x? What is the IQR for the attribute x? The higher whisker of the boxplot as at 18; what does this tell you? According the boxplot 18 is not an outlier and 37 as an outlier; why do you believe this is the case?

3. Learning classification Models Using 2 Methods and Model Comparison (Peer Graded and Peer Reviewed)

Available points: 57 (+ 5 extra points) + 15 points for peer reviews

In this task you get the opportunity to use two different classification methods for the cleaned Pima Indians Diabetes dataset; the idea is to apply two different classification techniques to the dataset, to compare the results, to potentially enhance the accuracy of the learnt models via selecting better parameters/preprocessing/using kernels/using “better” distance functions, and byincorporating background knowledge and to summarize your findings in a report.

Choose any 2 from the following 4 approaches to obtain classification models:

1. Neural Networks
2. Support Vector Machines
3. K-nearest Neighbors (up to 5 extra points are available if you choose k-NN as one of your methods for Task3)
4. Decision Trees

Other requirements for Task 3:

* Peer Reviewing: Each student will peer review submissions of three other students
* Accuracy of classification algorithms should be measured using 10-fold cross validation.
* Classification models that achieve higher accuracies will get more points.
* K-NN is relatively easy to implement, but there are also libraries; however, to get good accuracies you might need to spend some time to develop your own distance function[[8]](#footnote-8)! Make sure if you use a library that this library supports importing your own distance function.
* In your report after comparing the experimental results, write a paragraph or two trying to explain/speculate why, in your opinion one classification algorithm outperformed the others.
* Include a brief discussion in your report, how you have selected the parameters of particular data mining algorithms.
* If you applied special approaches to enhance the accuracy of the obtained classifier also describe those.
* In the report also include a brief description of the software you have used in the project.
* Finally, at the end of your report provide a 1-2 paragraphs summary that summarizes your most important findings of
* Your report must contain all the results you obtained for the 2 classification models as only the report will be graded.
* R supports support vector machines, neural networks, and random forests. However, you can use any tool you like for Task3 e.g. scikit-learn is another popular tool.

Deliverables (will be updated on September 16, 2020):

* Please submit ONLY the report using blackboard and/or to Kritik
* Name the report as *<last name>\_StudentID\_*T3.docx (or *<last name>*\_*StudentID\_*P3\_.pdf )

4. Tree and SVM Models and Classification in General(Not peer reviewed)

Available Points: 20

Compute the information-gain for the following decision tree binary split (compute the exact value; just giving the formula will only obtain partial credit)

(12,2,2) (12,0,0)

 (0,2,2)

Meaning: Originally there are 12 examples of C1 and 2 of classes C2 and C3 and after the binary split one subtree contains all examples of class C1 and the second subtree contains all examples of classes C2 and C3.

b. The soft margin support vector machine solves the following optimization problem:



i) What does the second term minimize? ii) What does the first term minimize? iii) What is the purpose of C? iii. Depict all non-zero ξi in Fig, 3 below!

 ****

Fig. 3: SVM Decision Boundaries for a Dataset containing Two Classes.

iv) How many examples are classified incorrectly by the SVM in Fig. 2? v) What is the relationship between ξi and example i being classified correctly? vi) What is the advantage of the soft margin approach over the linear SVM approach?

c) What are the characteristics of overfitting? What can be done to deal with overfitting when learning decision tree models?

1. Collaboration with other students in the course is not allowed! [↑](#footnote-ref-1)
2. We will also use a “cleaned version” of this dataset, for some subtasks of Task1; the dataset can be downloaded from the COSC 6335 webpage. [↑](#footnote-ref-2)
3. Supervised density plots could also be used instead! [↑](#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)
6. You find a link to download the Cleaned Pima Indian Dataset on the COSC 6335 website. [↑](#footnote-ref-6)
7. That is, interpreting missing values and their distribution and assessment of data quality is a subpart of those tasks! [↑](#footnote-ref-7)
8. If you use K-NN in conjunction with an Euclidean distance function, likely the accuracies will not be that great! [↑](#footnote-ref-8)