Carnegie Mellon University

Introduction to Machine Learning

Multiple Choice Questions & Project Work: Week 2

Due date: Monday 7/17/23 @ 11:59 P.M.

INDIVIDUAL WORK:
Total Points: 100 (Please see breakdown in Section 3 Overall Requirements)

1 Homework Assignment Topic: Classification.

Predicting the risk of disease using existing clinical variables is an important topic in preventative medicine. The major goal of this assignment is to develop a model that can accurately detect diabetes in women of Pima Indian tribe by carefully considering the different implications of false positives and false negatives. The dataset is available for download at this link.

2 Dataset Information

The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information of 768 women from a population near Phoenix, Arizona, USA. All patients here are females at least 21 years old of Pima Indian heritage. The outcome tested was Diabetes, 258 tested positive and 500 tested negative. Therefore, there is one target/class (dependent) variable and the following attributes:

1. Pregnancies (number of times pregnant),
2. Oral glucose tolerance test - OGTT (two-hour plasma glucose concentration after 75g anhydrous glucose in mg/dl),
3. Blood Pressure (Diastolic Blood Pressure in mmHg),
4. Skin Thickness (Triceps skin fold thickness in mm),
5. Insulin (2 h serum insulin in mu U/ml),
6. BMI (Body Mass Index in kg/m2),
7. Age (years),
8. Pedigree Diabetes Function ('function that represents how likely they are to get the disease by extrapolating from their ancestor's history')

3 Overall Requirements

1. **[20 Points]**: Perform data exploration & visualization (both univariate and multivariate analysis) using three different techniques that have been studied. The analysis should include the diabetes variable and display how other variables relate to it.
2. **[20 Points]** Because the data contains several physical impossibilities including blood pressures of zero, it is critical to handle/replace these and missing values. Determine and execute your approach to mitigating the risk of using the data as

is and submit a write-up on your reasoning and outcomes of the actions you have taken. The write-up may include a small section (7-10 sentences). Include the visualization and revised summary statistics comparing them to the original data.

3. **[40 Points]** Explore 4 different supervised classification models (refer to the list below). Draw a conclusion on the best performing one. Provide confusion matrices and other performance metrics in the form of tables to compare and contrast the performance of the models.
4. **[10 Points]** Write up a paragraph containing 3-7 sentences explaining your rationale for selecting the best performing model among the four.
5. **[10 Points]** 3-5 sentences relating your model to the analysis goal (specifically addressing false positives and false negatives)

4 List of Supervised Classification Algorithms to work on:

1. Logistic Regression
2. K-nearest neighbors
3. Decision Trees
4. Random forests

5 Homework Submission Requirements:

1. **Report:** Please submit a pdf document containing the detailed description and/or figures corresponding to each of the items listed in the overall requirements section above.
2. **Code:** Submit your code for grading. Make sure you comment each section of the Jupyter notebook with a meaningful description of what the section is attempting to do. Perform these steps prior to submission to ensure we can evaluate and grade your submission:
    a. Kernel->Restart & Run All to execute the notebook from a blank slate
    b. Double-check text, math, code, outputs, figures. Re-run if needed
    c. File->Download as->HTML (.html) to make A3_Classification_<<yourname>>.html
    d. Open A3_Classification_<<yourname>>.html in your web browser
    e. File->Print (Save as PDF) in your browser to make A3_Classification_<<yourname>>.pdf
    f. **Submit two files**, namely,
        i. A3_Classification_<<yourname>>.pdf and your notebook
        ii. A3_Classification_<<yourname>>.ipynb.

        Do not submit your HTML file.

Good luck with your assignment!