CIS 635 Data Mining

Project

Description

The project this semester consists of building a pipeline for periodic analysis of (generated) employee medical data. You will need to prepare the data by combining two data sources, cleaning the data and preparing it for processing. You will also need to choose the best classification technique. The end result is that you will create a pipeline to prepare and analyze the data and write a report of your work.

Data

The files that you are given contain data about 6,781 employees. You will get weekly training and test data from 2 sources. All files have an id for each employee which uniquely identifies the employee. Thus employee 3473 in the A file is the same employee as 3473 in the B file. The problem is that the id numbers are not in order. Also, there are some missing values that need to be estimated.

Source A files contain some vital measurements taken weekly:

variable	type	description	acceptable values
id	unique key		
temp	numeric	patient's temperature	90 - 106
bpSys	numeric	blood pressure (systolic)	90 - 150
vo2	numeric	VO^2 max	10 - 70
throat	numeric	throat culture	80 - 120
atRisk	factor	virus test	0=no virus, 1=has virus

Source B files contain the results of weekly surveys:

variable		description	acceptable values
id	unique key		
headA	factor		0 to 9
bodyA	factor		0 to 9
cough	factor		0=no, 1=yes
runny	factor		0=no, 1=yes
nausea	factor		0=no, 1=yes
diarrhea	factor		0=no, 1=yes

Procedure

Follow the steps below to process the data files:

1. do some preliminary analysis and clean up any outliers or missing data
2. run tests to compare classifiers and make a choice based performance
3. experiment with different plots (histogram, scatter plot, line plot) to determine one that clearly illustrates something interesting about the data
4. write an R script that can be run weekly to
   a) clean and merge the data sources (two files for training and two for test)
   b) build classification model with the training data
   c) predict atRisk for employees using the test data and write predictions to a text file
   d) generate the chosen plot and save to a file (.png)
5. analyze the data further (plotting, clustering, statistics) to describe the data and to support your claim of which classifier is best.

Hand in

- the weekly R script
- a maximum 4 page report that discusses the decisions you made in the procedure above and the results that you received. Also, most importantly, include the conclusions you reached and the support for the classifier that was chosen as best. Use charts and plots to help illustrate. In writing your report, you can use the ideas in chapters 1 and 7 of the book Storytelling with Data by Cole Nussbaumer Knaflic (free for GVSU students at https://www.gvsu.edu/library/).

Rubric :

description	points
R script	
cleaning	5
merging	5
classification	15
generate plot	5
report	
analysis	10
writeup	10
total	50