Supervised Methods in Machine Learning
2nd report guideline: "kNN, Linear regression, and multilinear regression"
Author: Andrés Cifuentes

Please use this document as a guideline for your team submission of the ***second*** assignment. This assignment has a total mark of ***30*** points. This report will be used for assessing your understanding of the basics of machine learning. Specifically, it is intended to assess your ability in the following aspects:

- Understanding of the problem. This includes your ability to perform an adequate data exploration and to obtain reasonable hypothesis/conclusions.
- Presentation of possible solutions. This includes your ability to present reasonable and supported ideas and arguments. You should introduce more than one possible solution since in real life a single problem can be solved by different means.
- Correct implementation of the possible solutions, including all reasonable and required steps from data acquisition to outcome of the solutions.
- Discussion of best outcomes and remarks.

This assignment will use the "diabetes_012_health_indicators_BRFSS2015.csv" dataset. A summary of the dataset can be found at the end of this document.

The submission can be in Spanish. However, it is expected that packages and functions documentations are in English.

Part 0: Datasets and codes are ***publicly available*** in a GitHub repository named as following the pattern *your_name_A2 (e.g. AndresCifuentes_A2)*. Report was created by using R markdown. Both R markdown and pdf are in the main page of the GitHub repository. **(2 points)**

Part 1: Data exploration and data wrangling (8 points).

To solve this section, you should use the "diabetes_012_health_indicators_BRFSS2015.csv" dataset. Implementation MUST be clearly explained in a PDF document created by using R markdown. Marking will follow the criteria shown below:

- There is a section in the report introducing the overall dataset.
- There is one subsection explaining each of the following tasks:
    - Load the dataset.
    - Exploratory Data Analysis of original dataset: Including (but not limited to) a brief description of the variables, distribution of the variables in the dataset, summary statistics of the numerical variables, and relevant visualizations.
    - Binarization of the class variable (Diabetes_012) by considering no diabetes as 0 and prediabetes / diabetes as 1.
- Explanations includes relevant codes, reasonable visualizations of the datasets, images, and screenshots (if required).

Part 2: KNN (10 points).

Implement predictive models by using the KKN method when considering a) diabetes (binary version), b) HeartDiseaseorAttack, and c) sex as the class variable (1 class variable per model).

- Create adequate versions of the dataset for each model by sub-setting the dataset so that the class variable is balanced, and it corresponds to the 1% of the dataset. (tip: you can balance de dataset by using stratified random sampling.). Use the new sub datasets to further pre-process your data (Feature engineering).
- Using the correct sub dataset, perform and document the following experiments per each class variable:
    - train a KNN model using all variables besides the class variable as predictors. You must find an optimal K. Test the performance of your model by using 10-fold cross-validation.
    - Remove 5 predictors that you consider are not contributing to the model (explain your assumptions) and find a second KNN model. You must find an optimal K. Test the performance of your model by using 5-fold cross validation.
    - Remove 5 further predictors that you consider are not contributing to the model (explain your assumptions) and find a second KNN model. You must find an optimal K. Test the performance of your model by using 3 repeated 10-fold cross validation.
- Discuss all your codes and results.

Part 3: Linear and multilinear regression (10 points).

Implement predictive models by using the linear and multilinear regression when considering a) BMI, b) MentHlth, and c) PhysHlth as the target variable (1 target variable per model).

- Create new versions of the dataset for each model by sub-setting the dataset so that you use a representative sample of the dataset with 1% of the size of the original dataset. Use the new sub datasets to further pre-process your data (Feature engineering).
- Using the correct sub dataset, perform and document the following experiments per each target variable:
    - Train a multilinear regression model that use all variables. Test the performance of your model by using 10-fold cross-validation.
    - Remove all predictors that are not contributing to the model (that are not significant) and find a second multilinear regression model. Test the performance of your model by using 5-fold cross validation.
    - Select a single predictor that you consider is the best predictor of the target variable. Test the performance of your model by using 3 repeated 10-fold cross validation.
- Discuss all your codes and results.

Summary of the variables in the "diabetes_012_health_indicators_BRFSS2015.csv" dataset

Diabetes_012: 0 = no diabetes 1 = prediabetes 2 = diabetes
HighBP:0 = no high BP 1 = high BP
HighChol: 0 = no high cholesterol 1 = high cholesterol
CholCheck: 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
BMI: Body Mass Index
Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
Stroke: (Ever told) you had a stroke. 0 = no 1 = yes
HeartDiseaseorAttack: coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity: physical activity in past 30 days - not including job 0 = no 1 = yes
Fruits: Consume Fruit 1 or more times per day 0 = no 1 = yes
Veggies: Consume Vegetables 1 or more times per day 0 = no 1 = yes
HvyAlcoholConsump: (adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes
AnyHealthcare: Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
NoDocbcCost: Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
GenHlth: Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth: days of poor mental health scale 1-30 days
PhysHlth: physical illness or injury days in past 30 days scale 1-30
DiffWalk: Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Sex: 0 = female 1 = male
Age: 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Education: Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.
Income: Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more