“kNN, Linear regression, and multilinear regression”V

Alejandro Vergara Rincon 81190

2023-10-08

Introduction

In this paper we will analyze the results after using the KNN, linear regression and Multilinear regression models, using the “diabetes_012_health_indicators_BRFSS2015.csv” dataset, in order to choose the best model according to the indications or requirements of the exercises.

In this regard, we begin by briefly explaining the data set variables:

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

For more information about the dataset, please refer to: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

First part Data exploration and data wrangling

The first thing that was done was to load the dataset that had previously been downloaded to the computer.

```
folder<-dirname(rstudioapi::getSourceEditorContext()$path)
parentFolder <-dirname(folder)
data <-
  read.csv(paste0(parentFolder,"/Dataset/diabetes_012.csv"))
```

so:

1. **`folder <- dirname(rstudioapi::getSourceEditorContext()$path)`**: Retrieves the current script's directory location.
2. **`parentFolder <- dirname(folder)`**: Obtains the parent directory of the current directory.
3. **`data <- read.csv(paste0(parentFolder,"/Dataset/diabetes_012.csv"))`**: Reads a CSV file named "diabetes_012.csv" located in the "Dataset" directory, which is the parent directory, and loads the data into the **`data`** variable.