**Assignment #1 — due Tuesday, Sep. 26, 2023**

*Submit your homework on Canvas.

*No late homework will be accepted for credit.

*Append the codes you used to your submission.

## Problem 1: Investigation of Life Expectancy *(100 points)*

**Dataset:** Life_Expectancy_Data.csv
**Description:** The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data-sets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. The final dataset consists of 22 Columns and 2938 rows which meant 20 predicting variables. The meaning of these predictors are as follows.
*Year:* 2000-2015
*Status:* Developed or Developing
*Life expectancy:* Life Expectancy in age
*Adult Mortality:* Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
*infant deaths:* Number of Infant Deaths per 1000 population
*Alcohol:* Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
*percentage expenditure:* Expenditure on health as a percentage of Gross Domestic Product per capita(%)
*Hepatitis B:* Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
*Measles:* Measles - number of reported cases per 1000 population
*BMI:* Average Body Mass Index of entire population
*under-five deaths:* Number of under-five deaths per 1000 population
*Polio:* Polio (Pol3) immunization coverage among 1-year-olds (%)
*Total expenditure:* General government expenditure on health as a percentage of total government expenditure (%)
*Diphtheria:* Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
*HIV/AIDS:* Deaths per 1 000 live births HIV/AIDS (0-4 years)
*GDP:* Gross Domestic Product per capita (in USD)
*Population*: Population of the country
*thinness 1-19 years:* Prevalence of thinness among children and adolescents for Age 10 to 19 (% )
*thinness 5-9 years:* Prevalence of thinness among children for Age 5 to 9(%)
*Income composition of resources:* Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
*Schooling:* Number of years of Schooling(years)

**Goal:** Fit linear regression models to answer the following questions. Note that there are 193 countries which will be treated as factors. Although it is an important factor, please ignore the predictor "Country". Otherwise, there will be around 200 predictors.

1. Report the summary of the linear model. What are the predicting variables actually affecting the life expectancy? Justify your answer based on the outputs of linear regression model.

2. Construct the 95% confidence intervals for the coefficient of "Adult Mortality" and "HIV/AIDS". Are you confident that these predictors have negative impact on the life expectancy? Explain why.

3. Construct the 97% confidence intervals for the coefficient of "Schooling" and "Alcohol". Explain how these predictors impact the life expectancy.

4. Based on the p-values, which are the top-seven most influential predictors? Use these predictors to fit a smaller model and report the summary.

5. Use the smaller model to predict the life expectancy if a new observation is given with *Year=2008*, *Status=Developed*, *Adult Mortality=125*, *infant deaths=94*, *Alcohol=4.1*, *percentage expenditure=100*, *Hepatitis B=20*, *Measles=13*, *BMI=55*, *under-five deaths=2*, *Polio=12*, *Total expenditure=5.9*, *Diphtheria=12*, *HIV/AIDS=0.5*, *GDP=5892*,*Population*=$1.34 \times 10^6$, *Income composition of resources=0.9*, *Schooling=18*. Report the 99% confidence interval for your prediction.

6. Use AIC to compare the full model and the smaller model.

## Problem 2: Predicting Breast Cancer *(100 points)*

**Dataset:** BreastCancer_train.csv and BreastCancer_test.csv
**Description:** The objective is to identify each of a number of benign or malignant classes. Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself. Each variable except for the first was converted into 11 primitive numerical attributes with values ranging from 0 through 10. There are 16 missing attribute values. See cited below for more details. The meaning of these predictors are as follows.
*ID:* Sample code number
*Cl.thickness:* Clump Thickness
*Cell.size:* Uniformity of Cell Size
*Cell.shape:* Uniformity of Cell Shape
*Marg.adhesion:* Marginal Adhesion
*Epith.c.size:* Single Epithelial Cell Size
*Bare.nuclei:* Bare Nuclei
*Bl.cromatin:* Bland Chromatin
*Normal.nucleoli:* Normal Nucleoli
*Mitoses:* Mitoses
*Class:* Class

**Goal:** Learn a classifier based on the training dataset and test its performance on test dataset.

1. Use all the predictors to fit a logistic regression model and report the summary. Plot the ROC curve on the test dataset.

2. Use the predictors *Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin* to fit a logistic model and report the summary. Plot the ROC curve on the test dataset.

3. Use all the predictors to fit an LDA model and report the summary. Plot the ROC curve on the test dataset.

4. Use the predictors *Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin* to fit an LDA model and report the summary. Plot the ROC curve on the test dataset.

5. Use all the predictors to fit a QDA model and report the summary. Plot the ROC curve on the test dataset.

6. Compare all the above models by AUC.

## Problem 3: Implementing KNN classification *(100 points)*

**Dataset:** BreastCancer_train.csv and BreastCancer_test.csv
**Goal:** use any programming language to implement the KNN classifier. Your implemented codes should provide the following options:

(1) the option to specify the number of neighbours: $K$

(2) the option to specify a Gaussian kernel $k(u) = e^{-u^2/2}$ with a bandwidth $h$ to for weighting

$$w(\text{benign}|x_{test}) = \sum_{(x_i,y_i)\in\text{train set}} k\Big(\frac{\|x_{test} - x_i\|}{h}\Big) \cdot \mathbb{1}(y_i = \text{benign})$$

$$w(\text{malignant}|x_{test}) = \sum_{(x_i,y_i)\in\text{train set}} k\Big(\frac{\|x_{test} - x_i\|}{h}\Big) \cdot \mathbb{1}(y_i = \text{malignant})$$

so that the estimated probabilities of labels for the test observation are

$$\hat{p}(\text{benign}|x_{test}) = \frac{w(\text{benign}|x_{test})}{w(\text{benign}|x_{test}) + w(\text{malignant}|x_{test})} \quad \text{and} \quad \hat{p}(\text{malignant}|x_{test}) = 1 - \hat{p}(\text{benign}|x_{test})$$

(3) the option to output an ROC curve and AUC.

Submit the *source code with a detailed instructions on how to use your code.* Then report

1. The performance of KNN classification (with the best tuned $K$) on test data. Report the best tuned $K$-value, plot the ROC curve, and report the AUC. Then, report the running time (after the best $K$ is specified).

2. The performance of KNN Gaussian kernel weighting classification (with the best tuned bandwidth $h$) on test data. Report the best tuned bandwidth $h$, plot the ROC curve, and report the AUC. Then, report the running time (after the best $h$ is specified).

3. Compare with the achievements from *Problem 2*.