Homework 2 **CSCE 5622**
Due: 5.00pm on March 7, 2025

Instructions for homework submission
a) Please write a brief report and *include your code.*
b) Create a **single pdf** and submit it on **CANVAS**. Please do not submit .zip files or colab notebooks.
c) Please start early :)
d) The maximum grade for this homework is **8 points** (out of 100 total for the class).

Analyzing life expectancy across the world

Since the 1950s, there has been substantial global progress in life expectancy at birth. From 1950 to 2015, the number of years that a newborn is expected to live, on average, increased worldwide by 24 years, or by about 3.6 years per decade (WHO, 2015). Despite this impressive global progress, large disparities remain in the levels of mortality observed across countries and regions. These differentials result from uneven progress in public health, inequalities in access to food and safe drinking water, sanitation, medical care, and behavioral patterns and societal contexts that affect the survival of individuals. The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status and related factors for all countries in the world. The datasets are made available to public for the purpose of data analysis for informing public health policies.

Figure 1: Countries and regions in the world (Source: nationsonline.org)

In this homework, our goal is to design a machine learning model that estimates life expectancy per country per year. Data can be found in 'hw2.csv' and contain 2,938 samples measured between 2000-2015 across 193 countries, including:

1. *Country*: country considered
2. *Year*: year of measurement
3. *Status*: status of the country (low/middle-income, upper-/high-income)
4. *Life expectancy*: life expectancy (age) **(outcome)**

5. *Adult Mortality*: adult mortality rate (probability of dying between 15 and 60 years per 1000 population)
6. $Infant\,deaths$: number of Infant Deaths per 1000 population
7. *Alcohol*: alcohol recorded per capita (15+) consumption (litres of pure alcohol)
8. $Percentage\,expenditure$: expenditure on health per capita
9. $Hepatitis\,B$: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
10. $Measles$: number of reported cases of measles per 1000 population
11. BMI: average body mass index of the population (kg/m^2)
12. *Polio*: number of reported cases of poliomyelitis
13. $Total\,expenditure$: total health expenditure expressed as a percentage of the country's gross domestic product (GDP) (%)
14. GDP: country's gross domestic product (GDP) (US dollars)
15. *Population*: country's population
16. $Thinness\,10-19\,years$: prevalence of thinness among adolescents aged 10-19 years, defined as individuals with BMI less than two standard deviations below the median (%)
17. $Thinness\,5-9\,years$: prevalence of thinness among children aged 5-9 years, defined as individuals with BMI less than two standard deviations below the median (%)
18. $Income\,composition\,of\,resources$: Income composition of resources (ICOR) measures how good a country is at utilizing its resources, graded between 0 to 1, higher ICOR indicates optimal utilization of available resources

The rows in 'hw2.csv' refer to the data samples, while the columns denote the corresponding variables for each data sample. The *training data* will include samples from Afghanistan to South Africa (rows 2-2410). The *development data* will include samples from South Sudan to Tuvalu (rows 2411-2715). Finally, the *testing data* will include samples from Uganda to Zimbabwe (rows 2716-2939).

(i) (1 point) Data cleaning: Please clean the dataset. You can identify and substitute any missing data, identify and correct any incorrect values, etc.

(ii) (1 point) Data exploration: Compute the correlation matrix $C \in \mathbb{R}^{15\times 15}$ that contains the Pearson's correlation coefficients between all pairs of numerical variables (except the year, i.e., from $Life\,expectancy$ to $Income\,composition\,of\,resources$). The element $C(i,j)$ of the matrix will include the Pearson's correlation between variable i and feature j. Visualize the matrix C using a heatmap. Which variable are the most correlated to each other? Which variables are the most correlated with the $Life\,expectancy$ outcome? Do these results align with your initial expectations? Please discuss your observations.
Note: You can use the *matshow* function from *matplotlib.pyplot*.

(iii) (2 points) Predicting life expectancy: The goal of this question is to predict life expectancy using economic, social, and health factors. **Implement** a linear regression model using

the ordinary least squares (OLS) solution. The output of the model is $Life\,expectancy$. The input features include: $Year$, $Status$, $Alcohol$, $Percentage\,expenditure$, $Hepatitis\,B$, $Measles$, BMI, $Polio$, $Total\,expenditure$, GDP, $Population$, $Thinness\,10-19\,years$, $Thinness\,5-9\,years$, $Income\,composition\,of\,resources$. Report the coefficient of determination R^2, Pearson's correlation r, and mean absolute error MAE between the actual and predicted life expectancy values on the development and testing sets.
Hint: You will build the data matrix $\mathbf{X} \in \mathcal{R}^{N_{train} \times D}$, whose rows correspond to the training samples $\mathbf{x_1}, \ldots, \mathbf{x_{N_{train}}} \in \mathcal{R}^{D \times 1}$ and columns to the D features (including the constant 1 for the intercept): $\mathbf{X} = \begin{bmatrix} 1,\ \mathbf{x_1^T} \\ \vdots \\ 1,\ \mathbf{x_N^T} \end{bmatrix} \in \mathcal{R}^{N_{train} \times D}$. Then use the ordinary least squares solution that we learned in class: $\mathbf{w^*} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}$.
Note: You can use libraries for matrix operations, random sampling, and model evaluation criteria, but please **implement** the linear regression algorithm and **add short comments to your code**. You can use the pseudoinverse function to compute the closed-form solution. In case the computation of the closed-form solution takes a long time, you can sub-sample your training data and compute an estimate of the closed-form solution using a random subset of the training data.

(iv) (1 point) Please discuss the estimated coefficients of the linear regression model. How might these findings support public health officials in making informed decisions? Furthermore, in what ways could this model and its outcomes be utilized to educate the public?

(v) (1 point) Based on your findings from question (ii), experiment with different feature combinations using linear (and non-linear) regression models. Please use the development data for hyper-parameter tuning (i.e., to assess the feature selection and the linear/non-linear regression models) based on the mean absolute error MAE between the actual and predicted life expectancy values. Please report and discuss the MAE results from the experiments on the development data. Using the model that gave the best MAE in the development data, please report the MAE on the test set.
Note: You can use publicly available libraries for this question. We would like to have an informative but non-redundant feature set, i.e., the features should be predictive of the outcome of interest but not too highly correlated with each other.

(vi) (1 point) Use the sample mean of the $Life\,expectancy$ outcome to binarize the data (i.e., assign samples with life expectancy larger than the mean to class 1 and samples with life expectancy lower than the mean to class -1). Run a logistic regression algorithm to classify between class 1 and -1. Use the development data for hyper-parameter tuning (i.e., to determine the regularization strength, regularization penalty term, etc.) based on the classification accuracy metric. After hyper-parameter tuning, report the accuracy of the classifier on the test data using the best combination that resulted from the development data.
Note: You can use publicly available libraries for this question.

(vii) (1 point) Conduct a separate analysis for low/medium-income countries (dataset 1) and upper/medium-income countries (dataset 2). Specifically, replicate the analyses from questions (ii) and (v) for each dataset independently. Please provide the correlation matrix and the MAE for both low/medium-income countries and upper/medium-income countries. What are your key observations? *Note:* You can use publicly available libraries for this question. You can maintain the same split for the training, development, and test sets (e.g., the train set for dataset 1 will include samples from all low/medium-income countries from Afghanistan to

South Africa (rows 2-2410)).

References

World Health Organization. (2015). World Mortality Report. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/mortality_theme_wmr2015_highlights.pdf