Homework 1 **CSCI 5622**
Due: 11.59pm on September 23, 2024

Instructions for homework submission
b) Please write a brief report and *include your code.*
c) Create a **single pdf** and submit it on **CANVAS**. Please do not submit .zip files or colab notebooks.
d) Please start early :)
e) The maximum grade for this homework is **7 points** (out of 100 total for the class).

The data included in this homework ('Breast_Cancer.csv') contains information of breast cancer patients from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institute of Cancer (2017). The data involved female patients with infiltrating duct and lobular carcinoma breast cancer diagnosed in 2006-2010. In total, there are 4024 samples, with 85% pertaining to the Alive class. The dataset contains the following variables, where variables 1-14 correspond to the features and variables 15-16 correspond to the outcomes:

1. **Age**: Age
2. **Race:** Race
3. **Marital status:** Marital status
4. **T Stage:** Tumor stage based on American Joint Committee on Cancer (AJCC) Tumor (T), Nodal (N) and Metastatic (M) staging
5. **N Stage:** Nodal stage based on American Joint Committee on Cancer (AJCC) Tumor (T), Nodal (N) and Metastatic (M) staging
6. **6th Stage:** 6th stage based on American Joint Committee on Cancer (AJCC) Tumor (T), Nodal (N) and Metastatic (M) staging
7. **differentiate:** Degree of differentiation of cancer cells from normal cells
8. **Grade:** The grade of cancer
9. **Stage:** The stage of cancer, Regional – A neoplasm that has extended, Distant – A neoplasm that has spread to parts of the body remote from the primary tumor either by direct extension or by discontinuous metastasis
10. **Tumor size:** The size of the tumor in millimeters
11. **Estrogen status:** Whether the breast cancer cells are estrogen receptor (ER) positive or negative
12. **Progestrone status:** Whether the breast cancer cells are progestrone receptor (PR) positive or negative
13. **Regional Node Examined:** The amount of regional node examined to assess the spread from the original tumor to the nearest lymph nodes
14. **Regional Node Positive:** Lymph nodes with metastases greater than 0.2mm (micrometastases or larger)

15. **Survival Months:** Patient survival in months
16. **Status:** Whether the patient is alive

(a) (1 point) Data exploration: Identify the categorical and continuous variables. Plot the histogram of each variable (i.e., 16 histograms). How are the variables distributed (e.g., unimodal, bimodal, uniform distributions)?

(b) (1 point) Data exploration: Plot scatter plots between each continuous feature and the 'Survival months' outcome. Following that, compute the Pearson's correlation between each continuous feature and the survival months outcome. What associations do you observe between features and outcome?

(c) (1 point) Data exploration: Plot grouped bar charts for each categorical feature and the 'Status' outcome. Each cluster of bars should correspond to one value of the categorical feature. Within each cluster, there should be two bars – one bar corresponding to the 'alive' and one bar corresponding to the 'dead' outcome.

(d) (3 points) Classification: Randomly split the data samples into training, validation / development, and testing sets (e.g., 70-15-15% split). **Implement** a K-Nearest Neighbor classifier (K-NN) to classify in terms of patient 'Status' (i.e., 'alive' / 'dead' outcomes). Use the euclidean distance ($l2$-norm) as a distance measure for continuous or clearly ordered variables and the Hamming distance as a distance measure to categorical non-ordered variables (e.g., marital status). Explore different values of $K = 1, 3, 5, 7, 9, 11, \ldots$. You will train one model for each K value using the train data and compute the classification accuracy (Acc), balanced classification accuracy ($BAcc$), and F1-score (the latter calculated based on the 'dead' class) ($F1$) of the model on the validation set. Plot the Acc, $BAcc$, and $F1$-score metrics on the validation set against the different values of K. Please report the best hyper-parameter K^* based on the Acc metric, the best hyper-parameter K^{**} based on the $BAcc$ metric, and the best hyper-parameter $K^\dagger$ based on the $F1$ metric. Finally, report the Acc, $BAcc$, and $F1$ metrics on the test set using K^*, K^{**}, and $K^\dagger$.

Note: Please **implement** the K-NN and the hyper-parameter tuning process. You can use available libraries that implement basic operations, such as vector/matrix operations, but you cannot use libraries that implement the K-NN, randomly split the data, or calculate performance metrics. Please provide your code with comments in the report.

(e) (0.5 points) ML deployment: Assume that the University of Colorado Anschutz Hospital is planning to deploy this system over the next months in order to predict breast cancer survival for their patients. What would be your thoughts / questions / concerns regarding this?
Note: CU Anschutz predominantly serves patients in Denver, CO (31% Latino/Hispanic, 10% Black/African American) and Aurora, CO (29% Hispanic/Latino, 16% Black/African American, 6% Asian).

(f) (0.5 points) ML deployment: Using the optimal $K^\dagger$ that resulted from the hyper-parameter tuning process in terms of $F1$-score in question (d), report the $F1$ metric separately for each race that is present in the data (i.e., White, Black, Other).
Note: The above categorization does not adhere with the U.S. Census Bureau standards on race and ethnicity, that require five minimum categories: White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander.

.