Instruction: Completed homework should be typed (e.g., using LaTeX or word document) or hand-written clearly and scanned and uploaded into Moodle. You can discuss about how to use certain tools for data collection and analysis, but no collaboration (e.g., sharing code or using "automated code" generation tools) is permitted to solve the problems.

1. **Learning objective:** Working example of model inversion attack.
    The IWPC Warfarin dosage equation is provided below:

$$Dosage \left(\frac{mg}{week}\right)$$
$$= [5.6044 - 0.2014 \times \text{age (in decades)} + 0.0087 \times \text{height (cm)}$$
$$+ 0.0128 \times \text{weight (kg)} - 0.8677 \times \text{VKORC1 A/G} - 1.6974 \times \text{VKORC1 A/A}$$
$$- 0.4854 \times \text{VKORC1 genotype unknown} - 0.5211 \times \text{CYP2C9} * 1/* 2$$
$$- 0.9357 \times \text{CYP2C9} * 1/* 3 - 1.0616 \times \text{CYP2C9} * 2/* 2 - 1.9206 \times \text{CYP2C9} * 2/* 3$$
$$- 2.3312 \times \text{CYP2C9} * 3/* 3 - 0.2188 \times \text{CYP2C9 genotype unknown}$$
$$- 0.1092 \times \text{Asian race} - 0.2760 \times \text{Black or African-American}$$
$$- 0.1032 \times \text{unknown or mixed race} + 1.1816 \times \text{enzyme} - \text{inducer status}$$
$$- 0.5503 \times \text{amiodarone status}]^2$$

*   Age in decades should be entered as 1 for 10–19 years, 2 for 20–29 years, 3 for 30–39 years
*   Enzyme-inducer status = 1 if patient is taking carbamazepine, phenytoin, rifampin, or rifampicin, otherwise 0
*   For VKORC1, CYP2C9, race and amiodarone status enter 1 if present, otherwise 0. For example, if the patient has VKORC1 A/G genotype then only the coefficient for VKORC1 A/G is 1 and the coefficient for all other VKORC1 genotype is 0. Similarly, for any **nonmatching coefficient assign such coefficient to 0** (e.g., if the race of a person is Caucasian then you simply assign 'Asian race' =0 and 'Black or African–American' =0). For any missing value simply assume its contribution is **zero** to the dosage (e.g., VKORC1 G/G contributes nothing to the dosage). Note $* 3/* 3$ is a type of CYP2C9 genotype (don't confuse '*' for multiplication and '/' for division operation in the formula, multiplication is represented by '×' and there is not division operation in the formula).

Variable	Units or Allowed Values
Age	Years
Height	Centimeters (cm)
Weight	Kilograms (kg)
VKORC1 genotype	A/A A/G G/G U ((for Unknown)
CYP2C9 genotype	*1/*1 *1/*2 *1/*3 *2/*2 *2/*3 *3/*3 U (for Unknown)
Race	A (for Asian) B (for Black or African American) C (for Caucasian or White) U (for Unknown or Mixed Race)
Taking Enzyme Inducer	Y (for Yes) N (for No or not known)
Taking Amiodarone	Y (for Yes) N (for No or not known)

Assume Robin is your neighbor who is 74-year-old Asian male. Assume you are gym buddies with Bob and hence you know his height (5 feet 6 inch) and weight (around 80 kg). Also let us assume that Robin has bipolar disorder (i.e., takes carbamazepine) and heart rhythm problem (i.e., takes amiodarone). One day while visiting Robin's home you saw Robin's prescription for Warfarin dose to be around 24.8 mg/week then what is his VKORC1 and CYP2C9 genotypes? Write a code to compute the genotype using all the auxiliary information you have.
Submit your **code** and README if there are any special instructions. **[points 20+10]** [code + answer]

2. **Learning objective:** How is Federated Learning impacted by the participation of worker nodes/clients and also how the number of training epoch impacts accuracy.

You are given a skeleton code for testing Federated Learning technique. Your goal is to determine the impact of number of clients participating (NUM_CLIENTS) and the number of training epochs (NUM_EPOCHS) each client execute locally.

a) Vary the value of number of clients NUM_CLIENTS as 5, 25, 50 (keeping the number of epochs NUM_EPOCHS fixed to 5) and report the test accuracy for different values of NUM_CLIENTS as shown in the table below. **[points 15]**

b) Vary number of training epochs NUM_EPOCHS as 5, 50, 100 (keeping NUM_CLIENTS fixed to 5) and report the test accuracy for different values of NUM_EPOCHS as shown in the table below. **[points 15]**

Remember to change the seed for random number generation to your student ID number. **(see TODO instruction within the code)**

Part (a) table

NUM_CLIENTS	Accuracy (when NUM_EPOCHS =5)
5	
25	
50	

Part (b) table

NUM_EPOCHS	Accuracy (when NUM_CLIENTS =5)
5	
25	
50	

c) Also comment on what you see when you vary NUM_CLIENTS and NUM_EPOCHS from the tables listed above. **[points 10]**

The sample code only provides how to initialize the model and train and test the model for one value of NUM_CLIENTS and NUM_EPOCHS. You need to write a **nested** loop for part 'a' and 'b' where you vary NUM_CLIENTS and NUM_EPOCHS, respectively. Remember to initialize the model in each iteration. The reinitialization part is given below-

```
# initialize the parameters of the ML model (you need to initialize this each time
you change the client number or epoch numer)
state = iterative_process.initialize()
```

Submit your code and README if there are any special instructions. If want to use separate code for the different parts you can do so, but your README should tell us how to generate the results shown tables.

Setting up environment: Try doing it in **Google Collab** and run the sample notebook provided. If you see error in the execution of the third block (as shown below), rerun the same block again.

```
Rerun Block twice

Once new numpy version is installed you'll see a 'Restart Service' request … restart
before proceeding
```

3. **Learning objective:** Computing statistical parity to see if ML models are fair across different demographics.
   You are given the 'adult' dataset which will be used to predicts if a given adult has an income greater than >$50K. Determine whether the training data has statistical parity for the following protected groups:
   a) Gender: Female **[points 15]**
   b) Race: Black **[points 15]**
   If the difference in probability for the protected and non-protected group is less than 0.05 then you can assume they are very close and thus not biased (i.e., fair) towards the protected group. You are given a skeleton code (in python3) to read and parse the input data. Please look at the code comment to understand the structure of the data. Categorical values have been flattened to generate one-hot encoding.
   Submit your code and README if there are any special instructions.

Submission:

You have to submit three files:

1. Merge all the written parts into a single pdf file named <your unity id>_HW4.**pdf**.
2. Rename the program file you used for as <your unity id>_HW4_QX.extension (e.g., .c/.cpp/.java/.py for question X).
3. Add a README file regarding how to run your code.

Zip all files into <your unity id>_HW4.zip and submit the zip file on Moodle.