1) Given the following dataset with two features and real-valued output, which of the following models will be most suitable in case of i) linear regression ii) ridge regression, iii) LASSO regression?

X1 1.5 -8.4 4.1 -2.3 5.7
X2 2.7 2.2 2.0 -5.5 -0.5
Y 6.3 -3.7 7.9 -12.9 4.9
Model 1: Y = X1 + 2*X2 Model 2: Y = 2.5*X2 Model 3: Y = X1 + 2*X2 – 0.3

Solution sketch:
For each of the three models, the loss function is different

i) Linear Regression: $1/N\Sigma_i(y\_i- y\_pred,i)^2$
ii) Ridge Regression: $1/N\Sigma_i(y\_i- y\_pred,i)^2 + 0.5*||W||^2\_2$
iii) LASSO regression: $1/N\Sigma_i(y\_i- y\_pred,i)^2 + ||W||\_1$

Where y_pred = W.X = w1*x1 + w2*x2 + w0

In each case, compare the loss value of the 3 models. The model for which this value is least, is the best. The best model need not be the same for all the loss functions.

2) Two time-series X and Y are provided. Check for (lagged) auto and cross-correlation between them. Using Granger causality, identify if there is any causal relationship between them. Consider a maximum lag of 2.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
X	2	4	0	4	4	2	-4	-6	-2
Y	1	-2	2	0	-1	-3	-1	2	0

Solution Sketch:
Correlation coefficient between two time-series A and B: Cov(A,B)/sqrt(Var(A)*Var(B)).
$Cov(A,B) = 1/T(\Sigma_t A(t)B(t)) - (1/T(\Sigma_t A(t)))* (1/T(\Sigma_t B(t)))$ where T is the number of observations.
Set A = X.
To calculate lagged auto-correlation with a lag of Δ, Consider B(t) = X(t- Δ)
To calculate lagged cross-correlation with a lag of Δ, Consider B(t) = Y(t- Δ). Repeat with B(t) = Y(t+Δ)
The number of observations T will decrease as the lag Δ increases. However, to calculate $(1/T(\Sigma_t A(t)))$ and $(1/T(\Sigma_t B(t)))$ we can use all observations.

In case of Granger causality:
Model 1: X(t) = a1*X(t-1) + a2*X(t-2); Model 2: X(t) = a1*X(t-1) + a2*X(t-2) + b1*Y(t-1) + b2*Y(t-2)
Estimate the coefficients for both models using least square regression, based on a subset of the observations (maybe 5).
See which model has less squared error in predicting the remaining observations.
If Model 2 has less error, then we can say that Y Granger-causes X. Else, Y doesn't Granger-cause X.

Repeat:
Model 1: Y(t) = a1*Y(t-1) + a2*Y(t-2); Model 2: Y(t) = a1*Y(t-1) + a2*Y(t-2) + b1*X(t-1) + b2*X(t-2)
Estimate the coefficients for both models using least square regression, based on a subset of the observations (maybe 5).
See which model has less squared error in predicting the remaining observations.
If Model 2 has less error, then we can say that X Granger-causes Y. Else, X doesn't Granger-cause Y.

3) Our aim is to find if Z is causally dependent on Y. Based on observations, we have estimated the following structural causal model:
W~N(10, 20), X~Ber(0.8), Y=5*X+W*(1-X), Z=N(3*W+5*Y,20)
Draw the graphical model corresponding to this SCM. What is the answer to our causal dependence question? Is there any confounder?

Solution sketch:
The edges are as follows: W->Y, X->Y, W->Z, Y->Z
Clearly, Z depends on Y, so there is a causal dependence. W influences both Y and Z, so it is a confounder.

4) We are trying to predict if the Masters' Degree marks M of a person is a causal factor for their annual income I. Consider the following dataset which records their Masters' Degree marks M, inherited income Y and age A. Considering only linear relations, what are your observations regarding the causal relationshipin question? Can you see any confounders in the dataset?

M	0	0	65	78	55	0	85	50	70
I	1	2	0	5	2	0	10	3	5
A	40	42	27	31	40	55	25	50	28
Y	2	3	3	5	6	8	8	10	15

Solution sketch:
This problem should be approached using the Double-ML formulation. Form two linear regression models: one for the target variable "I" based on intervention variable "M", using "A, Y" as possible confounders. In another model, regress "M" on "A, Y". Estimate the parameters as follows:

1) Regress I on A, Y (I=bA +cY), get residuals W and coefficients b, c by least squares
2) Regress M on A, Y (M=dA + eY), get residuals V and coefficients d, e by least squares
3) Regress W on V (W=fV), to get the coefficients d by least squares

Interpretation: how significant is the coefficient f compared to b and c? If high, then M has a causal relation on I. Also, do A or Y have a strong impact on both I and M (I.e. are the residuals W and V low)? If so, that (A or Y) is a confounder that impacts both I and M.

5) We want to check if a particular scholarship can prevent children from dropping out of School. Accordingly, we collected data of several students, including their various attributes. Accordingly, we attempt to carry out a randomized control trial by dividing the population into a control and an intervention group. Discuss how you will do it and what are your observations.

Scholarship received?	N	N	N	Y	Y	Y	N	Y	Y
Parents educated?	N	N	Y	Y	Y	N	N	Y	Y
Parents working?	N	N	Y	N	Y	N	Y	Y	Y
Gender	F	M	F	F	M	M	F	M	M
Dropped out?	Y	Y	Y	Y	N	N	N	N	N

Solution sketch:
Target variable: Dropped Out. Intervention: Scholarship received?
Divide the population into two groups (control and intervention), based on who received the intervention (scholarship) and who didn't.
Check out: are the two groups identical w.r.t. the other 3 attributes, I.e. is the fraction of "Y" equal in both groups for each attribute?
If yes: condition right for RCT. Check if fraction of drop-out is less in intervention group compared to control group.
If no, inspect if any of the other attributes seem to be strongly related to the drop-out outcome

6) The monthly sales Y of a particular item (represented by a real number) is known to depend on 3 factors: the season S (represented by a binary variable), the per-capita GDP of people G (whether they can afford to buy the item, represented by a real number) and the amount of investment A in advertising the item. Given the following observations of the variables, in each case, can we estimate which features have positive and negative contributions on the sales in each case?

ID	S	G	A	Y
1	0	6.5	1.2	23.7
2	0	7.1	2.6	27.5
3	0	7.8	2.1	27.6
4	0	7.6	1.5	21.8
5	0	8.9	0.7	20.6
6	0	8.4	1.6	21.9
7	0	8.8	2.2	28.8
8	0	6.9	2.8	27.9
9	0	5.1	2.2	24.5
10	0	4.8	0.9	18.4

ID	S	G	A	Y
11	1	4.9	1.8	26.7
12	1	4.3	1.9	26.5
13	1	5.5	2.3	31.8
14	1	7.8	2.0	39.5
15	1	11.8	1.7	42.8
16	1	9.2	1.4	38.6
17	1	6.8	0.8	31.3
18	1	6.4	0.9	32.1
19	1	6.9	1.1	33.2
20	1	5.1	1.5	32.9

Solution Sketch:

Since S is binary, consider the average value of the target Y for both values of S. Note that S=1 causes Y to rise, when marginalized over G and A. Calculate the mean values of G, A and Y, and for each example see their deviations from their respective mean values (irrespective of S). Find the relation between deviation of G and deviation of Y, irrespective of S and A (marginalizing them). It can be seen that there is a positive relation between them, i.e. increase of G causes increase of Y etc. Perform similar analysis for A. Now, for each example, we know which factor had a positive influence (i.e. tended to increase Y) and which factor had negative influence (i.e. tended to decrease Y).

7) In the previous case, since we do not know the structural causal model involving Y, S, G, A, so we need to train a highly accurate Machine Learning-based predictive model f, such that Y = f(S,G,A). Using this, can you estimate the Shapley Values of the features for each observation?

Solution Sketch:

This is basically a coding exercise. Fit a decision tree classifier and use SHAP package of Python.

8) In a small country, the chance of any person finding employment in a quarter is found to be related to the GDP growth rate G of the country, which has two states – "rising (R)" and "falling

(F)". The employment status ES(t) of any person ("employed E" or "unemployed U") in quarter t follows a probability distribution conditioned on this national GDP growth rate G(t), and also their employment status ES(t-1) in the previous quarter. The state transition distribution of GDP and the employment status distribution are provided below. A sequence of the GDP growth status over 5 quarters is provided, along with the number of employed NE(1) and unemployed NU(1) people in the first quarter (t=1). Estimate the number of employed NE(t) and unemployed NU(t) people in the remaining quarters.

GDP sequence (quarters 1 to 5): R R F F R

NE(1) =120, NU(1) =80

TRANSITION	ES(t) = E	ES(t) = U
G(t)=R, ES(t-1) = E	0.9	0.1
G(t)=R, ES(t-1) = U	0.6	0.4
G(t)=F, ES(t-1) = E	0.7	0.3
G(t)=F, ES(t-1) = U	0.1	0.9

Solution sketch:
At each quarter, estimate how many people change from employed to unemployed and vice versa using the probability table. For example, at t=2, G(t)=R and there are 120 employed people at t=1. According to the probability table, 90% of them (108) are expected to remain employed in t=2. Out of the 80 unemployed people, 60% of them (48) are likely to find employment. So NE(2)= 108+48=156, NU(2)=200-156=44. In t=3, G(3)=F, so according to the table 30% of the employed persons (156*0.3=47) may turn unemployed, while 90% of the unemployed people (44*0.9=40) are expected to remain unemployed. So after t=3, NU(3)=87, NE(3)=200-87=113 etc

9) Consider the quarterly time-series of sales of 2 different products that depends on the overall economic condition of the country. We have a time-series of length 20, consisting of the country's GDP growth rate and the sales of the 2 products. Using this data, I want to model the situation with a Hidden Markov Model, by representing the economic condition by a discrete state variable Z(t) which takes 3 possible values (strong, medium, weak). The emission distribution is Gaussian for each of the products, though with different parameters. Using empirical approach, how can we estimate the parameters of the HMM?

Quarter	1	2	3	4	5	6	7	8	9	10
GDP growth rate	6.1	6.3	6.9	5.4	5.2	5.7	2.4	5.5	5.9	7.2
X1	35	34	39	30	28	30	19	28	29	36
X2	61	63	75	58	55	57	38	55	57	75

Quarter	11	12	13	14	15	16	17	18	19	20
GDP growth rate	7.5	6.8	5.5	5.1	5.7	3.8	5.3	5.6	6.7	7.0
X1	35	33	31	30	32	17	29	31	37	38
X2	80	64	68	62	66	39	54	58	69	68

Solution Sketch:
Estimate the value of Z(t) at each step based on GDP growth rate. Discretize it appropriately, choosing

thresholds. Then estimate the state transition parameters using relative frequencies, and emission distributions by maximum-likelihood, i.e. to calculate parameters of X1 given Z=1, choose all examples where Z=1, and calculate the sample mean and sample variance of X1 based on those examples.

Example: let us say Z=1 if GDP growth rate is below 5.5, Z=2 if it is above 6.5 and Z=3 if it is between 5.5 and 6.5. So, the new dataset is like:

Quarter	1	2	3	4	5	6	7	8	9	10
GDP growth rate	2	2	3	1	1	2	1	1	2	3
X1	35	34	39	30	28	30	19	28	29	36
X2	61	63	75	58	55	57	38	55	57	75

Quarter	11	12	13	14	15	16	17	18	19	20
GDP growth rate	3	3	1	1	2	1	1	2	3	3
X1	35	33	31	30	32	17	29	31	37	38
X2	80	64	68	62	66	39	54	58	69	68

Estimate the transition parameters as follows: p(Z=1|Z=1) =4/8 = 0.5 (8 times we have Z=1, out of them 4 times it is followed by another 1). Similarly, P(Z=2|Z=1) = 4/8 = 0.5, p(Z=3|Z=1)=0 (as 3 never follows 1) This way, we can also calculate p(Z=i|Z=2) and p(Z=j|Z=3).

Estimation of transition distribution: for Z=1, the observations of X1 are (30, 28, 19, 28, 31, 30, 17, 29). Calculate their sample mean 'm11' and sample variance 's11'. So we can say: p(X1|Z=1) ~ N(m11, s11) Similarly, the corresponding values of X2 are (58, 55, 38, 55, 68, 62, 39, 54) from which we can calculate m12 and s12. Thus, p(X2|Z=1) ~ N(m12, s12). Now repeat same analysis for Z=2 and Z=3.

10) Now consider the same data as above, but our goal is slightly changed. I know that the economic condition of the country depends not only on the immediate GDP growth rate, but on the GDP itself which is a function of the past values of the GDP. Furthermore, your aim is now to predict the quarterly sales of X2 based on X1, and the GDP growth rate. Consider you want to represent this situation using an RNN, where the hidden state represents the economic condition as a 2D vector variable, and discuss how we can estimate the parameters of the RNN.

Solution Sketch:
RNN hidden state Z(t) is 2D vector, and has two inputs: G(t) and X1(t) which we together make a vector Y(t).
The RNN state transition equation is Z(t) = A*Z(t-1) + B*Y(t) where A, B are 2x2 matrices, and the output equation is X2(t) = C*Z(t) where C is 1x2 vector. Consider Z(0)=[0,0]. Using the values provided, formulate the equations in terms of the variables (A11, A12, … B11, B12 etc) and look for least-square solution. We will find that there will be many terms involving products of these parameters, so they cannot be solved easily using normal analytical approaches. So the approach will be gradient-descent by iterative approach. Here, we make initial estimates of all parameters. Then one parameter (say A11) is taken, and the derivative dL/dp of the total loss L is calculated WRT that parameter p. Accordingly, the parameter is updated ($p_{new} = p_{old} + a*dL/dp$) where 'a' is the learning rate.

11) The economic condition of a country is denoted by GDP growth rate G(t), the investment in industry development by I(t), and the average greenhouse gas emission rate by E(t). At any given time, we can either increase the investment or decrease it, but this has repercussions on the GHG emission. My aim is to simultaneously improve the economic condition and also maintain the environmental balance.

t	G(t)	I(t)	E(t)
0			2
1	5.0	1	3.0
2	6.1	1	3.5
3	6.5	0	3.3
4	5.6	1	3.6
5	6.1	1	4.1
6	7.0	0	4.0
7	6.3	0	3.9
8	5.6	1	4.1
9	6.5	0	3.9
10	5.8	0	3.4

t	G(t)	I(t)	E(t)
11	5.1	0	3.1
12	4.3	0	2.6
13	3.1	1	2.7
14	3.8	1	3.2
15	4.3	0	2.8
16	3.5	0	2.6
17	2.2	0	2.3
18	1.6	0	1.9
19	0.3	0	1.6
20	-0.9	1	2.1
21	-0.4		

i) From the given data, estimate how I impacts E and G using the following model:
$G(t+1) = (1+a*I(t))*G(t) + b*(1-I(t))$ if $G(t)>0$, $G(t+1) = G(t) + a*I(t)$ if $G(t)<0$
$E(t) = E(t-1)+c*I(t)*G(t) + d*(1-I(t))$

ii) Suppose the original economic condition X(t=0) and GHG concentration Y(t=0) is <100,60>, and the growth rates of both GDP and GHG emissions have been reduced to G(0)=0 and E(0)=0 due to Covid lockdown. Further, consider that I(t) can now be [0,1,2]. Now, use A* algorithm to optimize the sequence of investment decisions for 4 time-steps, so that economy doesn't shrink (i.e. X(t+1) should be >=X(t)), but GHG concentration does not rise above threshold of 80. The edge cost is the GHG emission. Note the X(t+1) = X(t) + G(t), Y(t+1) = Y(t) + E(t).

Solution Sketch:

First part is basically linear regression. Estimate all the variables using the least-square approach, separately for I(t)=1 and I(t)=0. You can actually estimate a, b, c, d etc at each step, and then take their means as the estimates.

In part 2, the state-space is <X(t), Y(t)>. Note that at each step you have 3 possibilities: I=0, I=1, I=2. Using the provided formulae, calculate the state-space at each time-step, in case of each decision. As mentioned in the question, the emission rate is the edge cost. The heuristic function can be the minimum of the children edge costs. The nodes that violate any of the two mentioned criteria, will have no children (STOP). The leaf nodes will be reached after 4 time-steps, and those leaf nodes which satisfy the given criteria are the goal nodes. Now apply A* as we know it.