Q1. a) You have been given a task to perform the data preprocessing of the data retrieved from multiple sources, before you start applying the data mining task. Identify, (atleast 5) data quality issues with the sample data set retrieved from the master data set. Suggest, how do you resolve these quality issues (python code is not required)? **[5]**

					BLOOD	COVID-19
TXN-ID	NAME	AGE	HEIGHT	WEIGHT	GROUP	RESULT
T001	RAMA	45	145	62kg	O+ve	Positive
T002	SEETHA	43	168	45kg	B+ve	Negative
T003	Akbar	38	172	60kg	Iam+ve	Positive
T004	BIRBAL	45	168	52kg	AB+ve	Negative
T005	THenali	22	157	78kg	B-ve	1
T006	Venkat	36	157	54kg	O-ve	Negative
T007	Rajuu	350	132	48kg	O+ve	Positive
T008	HARI	32	180	120lbs	AB-ve	Negative
T009	Inba	25		85kg	O+ve	0
T010	SysUsr789	20	165	68kg	O-ve	Negative

The attribute value SysUsr789 for the Name in the given data (T010 record) is not consistent with other names and it has alpha numeric when compared with other data types. (0.5)

This data quality issue can be resolved by replacing that field with right name/data type for consistency. (0.5)

The Age 350 is the outlier in T007 record and Height for Inba is missing (T009) (0.5)

These data issues can be resolved by filling the mean value of age and height. (0.5)

There is a mismatch in the data type units in T008, the Weight Unit for Hari is 120lbs whereas all other attributes are having Kg values. (0.5)

This is the data type issue and it can be done through data transformation by either manual or automatic edits of erroneous data (0.5)

The blood group has different representation in T004 record, inconsistent format of Iam+ve is being used in the blood group. (0.5)

This can be replaced with either NULL or by applying binning techniques (0.5)

Transaction id T005 has Covid Result-Representation Mismatch as 1 and in T009 it has 0, instead of indicating positive and negative values (0.5).

This data quality issue can be solved by applying data transformation such as data smoothing to make the simple changes as there are only two values which requires replacement. (0.5)

Q1. b) Your friend needs your help. She needs to classify job applications into good/bad categories, and also to detect job applicants who lie in their applications using density estimation to detect outliers. To meet these needs, do you recommend using a discriminative or generative classifier? Why? **[2 marks][1 mark each]**

Ans – Generative classifier [1] . Reason – for density estimation you should calculate P(x|y) [1]

Q2. a) Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Justify your stance. In such a scenario, what steps would you take? Should you increase the regularization hyperparameter, λ or reduce it? Why? [3]

The model is likely underfitting the training dataset which means it has a high bias. [1] You should try reducing hyperparameters. [1]

Higher values of the hyperparameter increase bias and reduce variance, while lower values have the opposite effect. A too-small value might result in overfitting, while a too-large value could lead to underfitting. Cross-validation techniques can be employed to find the optimal value for λ. Striking the right balance is essential for achieving a model that generalizes well to unseen data. [1 mark for justification]

Q2. b) Suppose you have been given a large dataset with n=2000000 instances and m(# of features)=300000 for each instance. You are asked to use multivariate linear regression to fit the θ parameters to our data. Which approach would you prefer, gradient descent or methods of least square and Why? [3]

Gradient descent (1 mark).

Method of least square is very slow, if n is very large. Computing inverse is roughly O(n^3) (2 marks for the explanation)

Q3. Suppose you are building a logistic regression model for the given dataset using gradient descent approach. You managed to identify the theta parameters θ_0, θ_1 such that $J(\theta_0,\theta_1)=0$ where $J(\theta_0, \theta_1)$ is the cost function. Which of the following statements (a-d) must be True? Justify your answer in each case.

[6 marks]

a) The model will work perfectly well for the unseen/new instances without any error. It will predict correct values of the target variable, Y. False, it's an over fitted model.

b) If $J(\theta_0,\theta_1)=0$ for some values of θ_0, and θ_1, then $H_\theta(x(i))=y(i)$ for every training example $(x(i),y(i))$. True

c) *For* $J(\theta_0,\theta_1)$ to be 0, θ_0, and θ_1 must be 0. False, it is not necessary

d) $J(\theta_0,\theta_1)$ cannot be 0. False. It can be 0.

[0.5 marks for True/ false. 1 mark for justification]

2. Explain the importance of feature scaling in learning model parameters, *θ in logistic regression.* *[2 marks]*

When using Gradient descent, you should ensure that all features have a similar scale, otherwise, it will take longer time to converge

Q4. Suppose we train a model to predict whether a credit card transaction is Fraudulent or Not. After training the model, we apply it to a test set of 200 new transactions (also labelled) and the model produces the contingency table below.

		Predicted Class	
		Fraud	Not Fraud
True Class	Fraud	60	0
	Not Fraud	120	20

List your crisp point-wise observations on the classifier with supporting justification. (4 marks)

Metrics with respect to Fraud Class:

		Predicted Class	
		Fraud	Not Fraud
True Class	Fraud	60 (TP)	0 (FN)
	Not Fraud	120 (FP)	20 (TN)

Precision = TP/(TP+FP) = 60/(60+120) = 60/180 = 33.33%
Recall = Sensitivity = TP/P = TP/(TP+FN) = 60/(60+0) = 100%

Metrics with respect to “Not Fraud” Class:

		Predicted Class	
		Fraud	Not Fraud
True Class	Fraud	60 (TP)	0 (FN)
	Not Fraud	120 (FP)	20 (TN)

Precision = TP/(TP+FP) = 20/(20+0) = 100%
Recall = TP/P = TP/(TP+FN) = 20/(20+120) = 14.29%

Precision wrt Fraud class = 33.33%
Precision wrt Fraud class = 100%
Recall wrt Not Fraud Class = 100%
Recall wrt Not Fraud class = 14.29%

Q5. a) Use ID3 decision tree algorithm to train the classifier, find which of the features among {Readership Base, Writer's Reputation spread in other countries} is best suited for "root node" in the tree construction. Pictorially represent complete resultant decision tree. Show all the calculations and round the values to four decimal scale as appropriate. **[4 Marks]**

Use case: Committee of experts convene every year to nominate literary works to become eligible for the awarded of highest category by assessing the works on multiple parameters. Below is one subset of such features. Categorizing the works provides a transparent & streamlined way of nomination process. Quantified values of attributes are discretized in below data. Build a machine learning model to classify if an original literary work of writers has “High” or “Medium” or “Low” chances of nomination by the committee.

Readership Base	**Writer's Reputation spread in other countries**	**Distinctive in Style**	**Chances of Nomination of Literary Work**
Low	High	High	High
Low	High	Low	Medium
Low	Low	High	Low
High	Low	Low	Low
Low	High	High	Medium
High	High	Low	High
Low	Low	Low	Medium
Low	Low	Low	Medium

b) Justify the below statement with any plagiarism free example. **[2 marks]**

"Assessing the model performance of built decision tree classifier using only the training data set is detrimental to the process."

----

a) (Both the below answer key must be accepted by the evaluators)

Answer Key-1 (if Log base “2” is used by students):

Class Entropy : 1.5

Entropy of feature 1”readers base”: 1.887, Gain : 0.3113

Entropy of feature 2 “writer reputation..”: 0.6667, Gain : 0.8333

Inference : “Writer’s reputation spread.... ” has the minimum entropy or maximum info gain and hence it’s the selected root for decision tree building

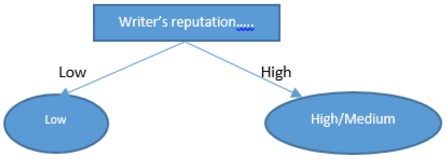
Answer Key-2 (if Log base “3” is used by students for k=3 number of distinct classes for normalization):

Class Entropy : 0.9464

Entropy of feature 1”readers base”: 0.75, Gain : 0.1964

Entropy of feature 2 “writer reputation..”: 0.4206, Gain : 0.5258

Inference : “Writer’s reputation spread.... ” has the minimum entropy or maximum info gain and hence it’s the selected root for decision tree building



Marking Scheme:
1 mark: Entropy of class attribute
1 mark: Entropy or Information Gain of the 1st feature
1 mark: Entropy or Information Gain of the 2nd feature
0.5 mark: Correct choice of the root
0.5 mark: Final decision tree with leaves (labelled with majority voting) in depth 1, built with chosen root.
Partial Marking: If none of the above are correct but correctly tried to implement the algorithm= 1.5m

b) Answer Key:
Generic reason is applicable here. A fully grown decision tree(DT) with complex rules is more prone to learn all the pattern in the training data. At the best case, accuracy of the DT most likely will be 100% and is not the good criteria to measure the perf. Unseen test/validation data is best for evaluation.
Marking Scheme:
1 mark: Correct reason
1 mark: Any sample dataset split to show that training & test (or validation) may be split with diverse data distribution for illustration
Partial Marking: -None-