Problem Answers and Instructions for Coding Solutions:

1. What is the main disadvantage when applying kNN to large datasets or datasets involving a large number of features?

KNN doesn't work well with large datasets because the accuracy of KNN reduces with high dimension data as there would be very little distance between nearest and farthest node. For every test data, distance between all test data and data points would be calculated which in turn would lead to long timeframe and a larger space.

2. Condensing is a technique to reduce the complexity of the kNN by reducing which term of the O(nd) complexity?

The term n can be reduced to reduce the complexity. We use condensing technique inorder to remove the unwanted prototypes during the training phase. Comparison is done between every test data in each training instance (i.e., n total comparisons). Therefore, lesser number of comparisons (n), lesser is the complexity.

3. Implementing knn for dataset 1.

   Solution (Code -- 3_Knn.ipynb) is attached.
   Instruction to run the code: copy the dataset into a path and change the path in the code and then execute the code. Change value of K (from 1 to 11) which is at line 90.

   Table :

K	Predicted	Actual	Accuracy (%)
1	1	2	75.17948717948717
2	3	0	70.12820512820513
3	0	4	73.74358974358974
4	3	7	75.82051282051283
5	8	4	81.356
6	4	2	84.56410256410255
7	3	5	87.87179487179487
8	3	1	82.9743574358974
9	5	9	79.998765465
10	1	8	84.61538461538462
11	4	3	71.79487179487179

   K=7 is the best K.

4. **It is handwritten and has been merged in the same pdf at the end.**

5. copy the dataset into a path and change the path in the code and then execute the code(5_NaiveBayes.ipynb).

6. By increasing the learning rate, the model converges to a suboptimal solution.
   The solution is attached (code -- 6_GradientDescAB.ipynb).
   Instructions to run the code :
   a. run the code for the part a -- **GradientDesc_A(4.5, 4, 0.05,0.001)**
   b. run the code for the part b -- **GradientDesc_B(2.4, 2, 0.05,0.001)**

7. Why applying linear regression to a classification problem may not be adequate? What is the effect of outliers in this type of classifier?

Linear regression technique is unbounded and predicts a continuous variable and therefore may not be adequate for classification problem.
While performing linear regression on binary classification problems, absolute numbers (0,1) are achieved. Due to the concept of overfitting, there are wrongs instances that get predicted during adding new instances and training them.
Outliers are data points that are numerically far away from other points in a dataset. As they affect the slope of line, predictions are wrong sometimes, data is distorted and also there are errors in the classification.

8. copy the dataset into a path and change the path in the code and then execute the code (8_logisticRegression.ipynb)

Assignment

Problem 4

	Cancer	Pulmonar infection
Radiotherapy	0	20
Medication	10	0

$\lambda_{11} = 0, \ \lambda_{12} = 20, \ \lambda_{21} = 10, \ \lambda_{22} = 0$

Risk, $R(\alpha_j | x) = \sum_{j=1}^{c} \lambda_{ij} P(\omega_j | x)$

$P(\omega_1) = 0.6$

$P(\omega_2) = 0.4$

$$R(\alpha_1) = \lambda_{11} P(\omega_1) + \lambda_{12}(P(\omega_2))$$
$$= 0(0.6) + 20(0.4)$$
$$= 0 + 8$$
$$= 8$$

$$R(\alpha_2) = 10(0.6) + 0(0.4)$$
$$= 6$$

$R(\alpha_1) > R(\alpha_2)$