Practical Data Science
(Resampling Techniques)

Solve the following problems.

Problem 1: Resampling schemes for ML programs

Given the following data with Age as predictor variable and Survived as target variable:

Id	Age	Survived
1	25	1
2	23	0
3	30	1
4	35	1
5	32	0
6	28	1
7	13	0
8	12	0

a. Build knn model with k=3 for above dataset. Findout the default resampling strategy used by train method of caret package.

b. Find out the observations used for train and test data in each iteration and also confusion matrix for each iteration:

i. Repeated Holdout with 3 iterations and 75% train data. Is stratification used?
ii. 4-fold Cross Validation. Is stratification used?
iii. 4-fold Cross Validation with 3 repeats. Is stratification used?
iv. Leave one out cross validation.
v. Bootstrapping with 3 iterations.

c. Supply the parameter grid with k = 3,5,7,9,11 and 4-fold cross validation scheme(repeated 3 times). Understand how train function selects the optimal model. Is the final model built with entire data?

d. Use the optimal knn model build in above question to predict the class of following passengers:

Id	Age
9	26
10	36
11	9
12	24

Problem 2: Weighted KNN Learning

Given the following training data, predict the class of the following new example using K-Nearest Neighbour for k=5: age<=30, income=medium, student=yes, credit-rating=fair. For similarity measure use a simple match of attribute values:

Practical Data Science
(Resampling Techniques)

$$\text{Similarity}(A,B) = \sum_{i=1}^{4} w_i * \partial(a_i, b_i) \Big/ 4 \quad \text{where } \partial(a_i, b_i)$$

is 1 if a_i equals b_i and 0 otherwise. a_i and b_i are either *age, income, student* or *credit_rating*. Weights are all 1 except for income it is 2.

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no