The following results have been obtained by applying 5-fold cross-validation on the given dataset of 669 rows where each row contained value of 8 features and values ranged from 1 to 10. In each case, whole dataset was divided into 5 parts and 4 parts were used as training data and 1 part as test data.

Part ID of Test Data	TPR	TNR	PPV	NPV	FPR	FNR	FDR	F1 SCORE	ACCURACY (%)
1	0.82759	1.0	1.0	0.88372	0.0	0.17241	0.0	0.90566	92.53731
2	0.95522	0.94029	0.94118	0.95455	0.05970	0.04478	0.05882	0.94815	94.77612
3	0.93181	0.88888	0.80392	0.96385	0.11111	0.06818	0.19608	0.86315	90.29850
4	0.92857	0.98113	0.92857	0.98113	0.01886	0.07142	0.07142	0.92857	97.01493
5	0.97142	0.97959	0.94444	0.98969	0.02040	0.02857	0.05555	0.95774	97.74436

Taking accuracy as a measure of performance evaluation , using last one-fifth data as test data yielded the best performance.

TPR avg=0.9229256538623497

TNR avg=0.9579822617115935

PPV avg=0.9236227824463119

NPV avg=0.9545889207171884

FPR avg=0.042017738288406506

FNR avg=0.0770743461376503

FDR avg=0.07637721755368813

F1 SCORE avg=0.9206568655376298

ACCURACY avg=94.47424531477948 %

Why are you using cross validation? Do the dataset justify it?

Answer: Cross validation is a technique mainly used in such scenerios where the goal is to predict something and we need to estimate how accurately a predictive model will perform in practice. It involves partitioning data into a subset called training data and testing data and fitting a function by using training data and predict test data using it. But simply portioning produces greater variance, so to reduce variance to repeat this process taking different subsets of training and testing data and averaging the result. . Every data point gets the chance to be in a test set exactly once, and gets to be in a training set *k-1* times. So variance reduces as k increases.

This dataset justifies it because the amount of data is limited, and simply dividing into training and test sets will make evaluations become sensitive to how the division was made.

So taking an average by applying 5-fold cross validation is justifiable in this case..[5]

b. Besides accuracy, which of the criteria mentioned above should be used in cross validation for the given data set? Explain.

Answer: Besides accuracy F1 score should be used in cross validation in this dataset.

The reason is that F1 score the harmonic mean of precision and recall. Since precision increases if recall decreases and the vice versa occurs, to ensure contribution of both of these performance measures, we can use F1 score as it conveys the balance between the precision and the recall.