**Can anyone please help on this?**

Consider the Boston housing dataset uploaded to BB. This small data set provides information about houses in different areas in Boston. The following table provides the dictionary of the variables in the data set.

|  |  |
| --- | --- |
| **Variables** | **Variable description** |
| CRIM | Crime rate |
| ZN | Percentage of residential land zoned for lots over 25,000 ft2 |
| INDUS | Percentage of land occupied by nonretail business |
| CHAS | Does tract bound Charles River (= 1 if tract bounds river, = 0 otherwise) |
| NOX | Nitric oxide concentration (parts per 10 million) |
| RM | Average number of rooms per dwelling |
| AGE | Percentage of owner-occupied units built prior to 1940 |
| TAX | Full-value property tax rate per $10,000 |
| PTRATIO | Pupil-to-teacher ratio by town |
| LSTAT | Percentage of lower status of the population |
| MEDV | Median value of owner-occupied homes in $1000s |
| CAT.MEDV | Is median value of owner-occupied homes in tract above $30,000 (CAT.MEDV = 1)  or not (CAT.MEDV = 0) |
| DIS | Weighted distances to five Boston employment centers |
| RAD | Index of accessibility to radial highways |

We have 2 variables as target variables used by Multiple Linear Regression and Decision Tree modeling. For the linear regression model, MEDV is the target variable and CAT.MEDV is the target variable for the decision tree model developed in this assignment. Students must not consider them as predictor variables. With respect to the Boston housing data set, answer the following questions using R software. You must provide answers, R codes, and the results of the analysis in the solution document.

1. Develop a multiple regression model to predict MEDV. Consider all other variables as predictors (except CAT.MEDV). Provide a summary of results (2 marks)
2. Which variables are statistically significant? (2 marks)
3. Evaluate the prediction performance of the model using R-square and RMSE parameters. Explain your findings (3marks)
4. Develop a decision tree model to predict CAT.MEDV. Consider all other variables as predictors (except MEDV). Provide a summary of results (2 marks)
5. Generate a training with 400 rows and a test set with 106 rows from the given dataset. Create a confusion matrix for the test set. Explain the results in terms of overall prediction error (3 marks)