Machine learning assignment report

First I started to read the excel sheet which is my data set and then I will use pandas library to put the values in this dataset in a frame (data frame) and then I will take a copy of the data and start to work on it and then I will not use all the features Which is in the dataset due to complexity and not to lead to over fitting so we will decrease the number of the features so i will use correlation function to get the relation between the features then I will have a matrix which will have numbers that represent the correlation between the features and each other. So I started to for loop on rows and columns and see the relation if the relation between a feature in row and another feature in a column is greater than 0.5 so we will remove this feature from the excel sheet and if it's not written in the correlation matrix we will have to add it and if its written before we will just remove it from the excel sheet and then we will call our function which is the correlation function and then from my training data x i will remove the column of the price because this column is what i want to predict so I don't need it and then i will remove also from the training data the first column and the 2nd column which is the date and the id because the date has numbers and letters which turns into string not like the other features so it may cause problems and the id is not important and we will not use it so we will remove them and then all of this we was working with data frame which uses pandas library so we will convert it to np and the data will be float and the y will be the predicted price of the house . I also inserted column of ones because of theta zero is not multiplied by x so this column is to make theta zero as if its multiplied by x but this x is ones so it will not change but till now I still has many features so I want to decrease them so I will see what features that we will benefit from going through polynomials (theta0+theta1X1+theta2x2+ theta3x3+theta4x4+....+theta1x1*2+theta2x2*2+...) so if I have features which its value is 0,1 I will not benefit or change it when it goes through polynomial so I will make an array which will take the index of the features that will go through polynomial and the I will see different plots of each feature using scatter plots and see the graph of each feature in the array and I will see which graph will make use of the polynomial and the graph that its shape doesn't change when a polynomial is applied to it so i will remove this feature also from the array of polynomials so there will be 5 features only which remain that will be used when we use a polynomial with a degrees so now the features will not be doubled it will be less than the double and this is better in terms of complexity

The second thing we will make a function to calculate the polynomial i will give it the x trained and give it the features and the degree I want to make so it will add a column with degree I want so if i have 2nd degree (x1+x2) *2 i will take x1*2 and x2*2 but I will not take the factor 2x1x2 to remove the complexity and make it easier and then i will start to make linear regression first I will compute the cost of the function j(theta0,theta1)=1/2m summation (htheta(xi)-yi)*2 and we apply it in the code and then i will calculate the gradient descent to minimize the cost function as much as i can so we will have x which is the array of the input and y the expected value and the theta which is array of theta values and alpha which is the learning rate and the number of iterations that the gradient descent will go through this

iterations and we will see the thetas that will give me the best cost and then will make the features normalize through standard scaler to avoid overflow and to make the values close to each other's and then we will go through the first technique k fold and we will choose the number of splits and then I will normalize them and i will try them on different degrees so the number of the splits was 10 so i will train 9 of them and one for test then in the second iteration I will another one for test and the other 9 will be trained so i will use all the date in train and test and then I will make gradient descent to have the best thetas and i will multiply this thetas with the XTest and then I will take the mean squared error of the y predicted and y test and will see which one has the least error so we will try with each degree and see each error and the degree which will give me the least the error will be considered the best degree and the results shows that degree two was the best degree with least errors and then will make a stratified k fold which is the same as k fold the only difference that we will divide them according to their labels and when we predict we predict with same percentage they are classified into and we also found that degree 2 was the best degree . The last technique was regularization we will compute the cost function with the regularization term and then the gradient descent with the regularization term and the. We will choose the regularization factor so to choose it we first will create array of lamdas and then we split the data to 60% train and 20% test and 20% cross validation and then we will normalize the data and then choose the alpha and the number of iterations so we will start the 1st iteration we will take each lamda with each degree to see the best degree so we will have gradient descent and we Will give it the x train and y train and theta, alpha and number of iteration with theta0 and with theta1 we will give it the same things but we will add lamda and then we will compute the cost of each and see the least cost cased by which degree and which lamda and then we will make sure by the testing values as we will test them with the degree and the lamda that gave us the least cost and see if it gives us a small cost or not.

Finally the results shows that when take lamda that gives us the least cost, we try it in the test and it gives us a high cost and this lamda was large and this was under fitting and the cost we have here in the regularization is high than kfold and the stratified k-fold and the cost of the k-fold was less than the cost of the stratified k-fold