Report

Task 1:

All regressors are included in the folder and implemented in the main.py.

Task 2:

- [housing dataset]

All features are used to train and test models.

- [California Renewable Production 2010-2018 dataset] – [CRP_new dataset]
- See main.py before main.

Data pre-processing:

Features selected: BIOMASS + BIOGAS

Target: mean (BIOMASS + BIOGAS), shift (-1), then fillna(0).

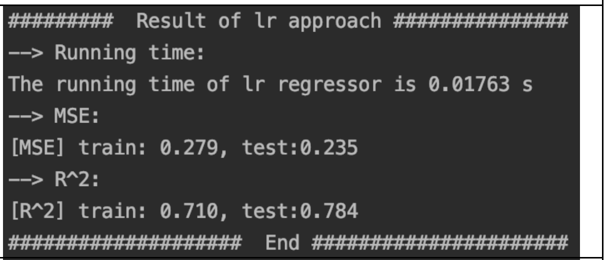
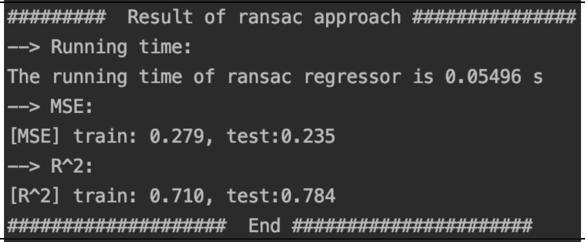
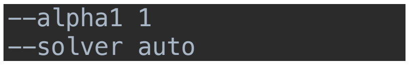
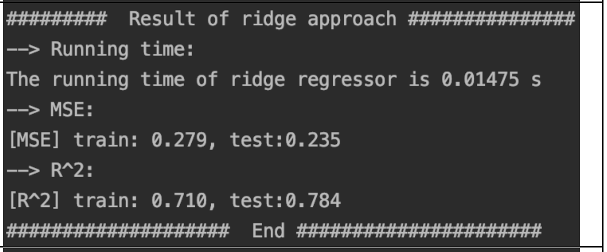
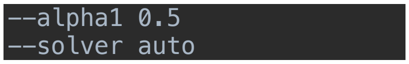
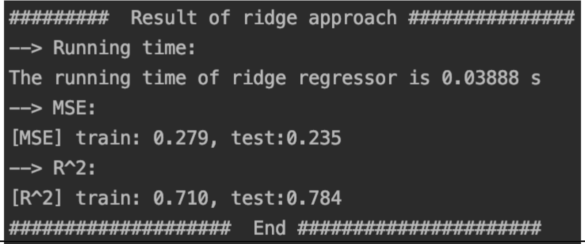
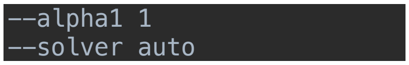
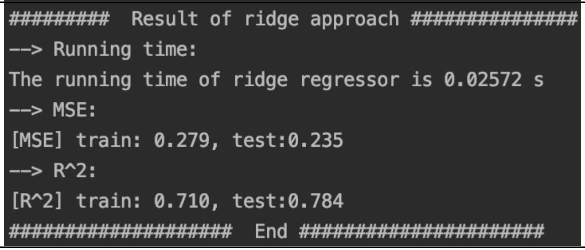
Purpose:

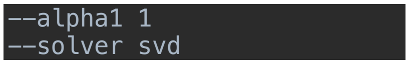
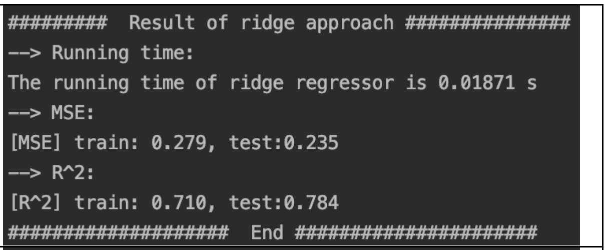
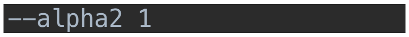
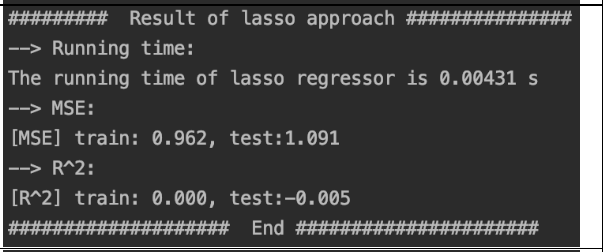
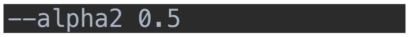
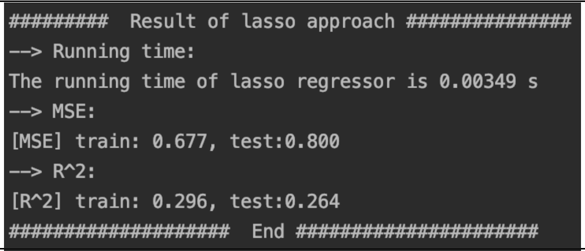
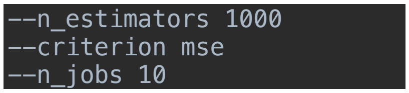
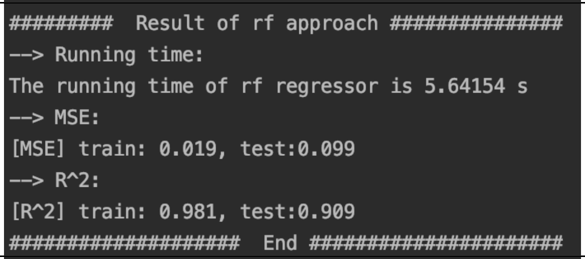
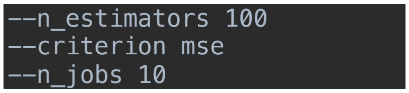
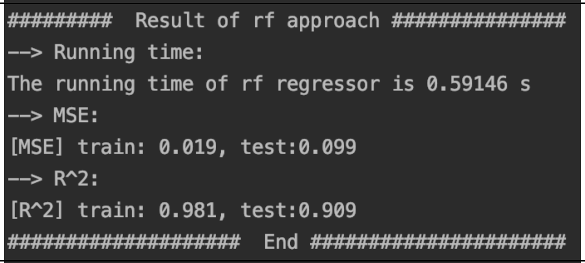
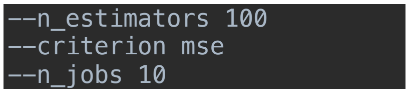
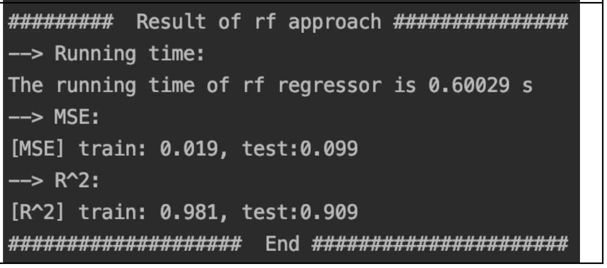
Predict the mean energy produced by BIO industry in the future.

- Each regressor has been tested by these two datasets.

Task 3

Table 1 Performance of different regressors [housing dataset]

Regressors	Parameters	Run Time and Model Evaluation Parameters
lr	default	 ######### Result of lr approach ############### --> Running time: The running time of lr regressor is 0.01763 s --> MSE: [MSE] train: 0.279, test:0.235 --> R^2: [R^2] train: 0.710, test:0.784 #################### End ######################
ransac	default	 ######### Result of ransac approach ############### --> Running time: The running time of ransac regressor is 0.05496 s --> MSE: [MSE] train: 0.279, test:0.235 --> R^2: [R^2] train: 0.710, test:0.784 #################### End ######################
ridge	 --alpha1 1 --solver auto	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.01475 s --> MSE: [MSE] train: 0.279, test:0.235 --> R^2: [R^2] train: 0.710, test:0.784 #################### End ######################
	 --alpha1 0.5 --solver auto	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.03888 s --> MSE: [MSE] train: 0.279, test:0.235 --> R^2: [R^2] train: 0.710, test:0.784 #################### End ######################
	 --alpha1 1 --solver auto	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.02572 s --> MSE: [MSE] train: 0.279, test:0.235 --> R^2: [R^2] train: 0.710, test:0.784 #################### End ######################

	 --alpha1 1 --solver svd	 ######### Result of ridge approach ############## --> Running time: The running time of ridge regressor is 0.01871 s --> MSE: [MSE] train: 0.279, test:0.235 --> R^2: [R^2] train: 0.710, test:0.784 ################### End #####################
lasso	 --alpha2 1	 ######### Result of lasso approach ############## --> Running time: The running time of lasso regressor is 0.00431 s --> MSE: [MSE] train: 0.962, test:1.091 --> R^2: [R^2] train: 0.000, test:-0.005 ################### End #####################
	 --alpha2 0.5	 ######### Result of lasso approach ############## --> Running time: The running time of lasso regressor is 0.00349 s --> MSE: [MSE] train: 0.677, test:0.800 --> R^2: [R^2] train: 0.296, test:0.264 ################### End #####################
rf	 --n_estimators 1000 --criterion mse --n_jobs 10	 ######### Result of rf approach ############## --> Running time: The running time of rf regressor is 5.64154 s --> MSE: [MSE] train: 0.019, test:0.099 --> R^2: [R^2] train: 0.981, test:0.909 ################### End #####################
	 --n_estimators 100 --criterion mse --n_jobs 10	 ######### Result of rf approach ############## --> Running time: The running time of rf regressor is 0.59146 s --> MSE: [MSE] train: 0.019, test:0.099 --> R^2: [R^2] train: 0.981, test:0.909 ################### End #####################
	 --n_estimators 100 --criterion mse --n_jobs 10	 ######### Result of rf approach ############## --> Running time: The running time of rf regressor is 0.60029 s --> MSE: [MSE] train: 0.019, test:0.099 --> R^2: [R^2] train: 0.981, test:0.909 ################### End #####################

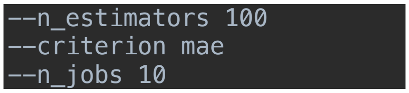
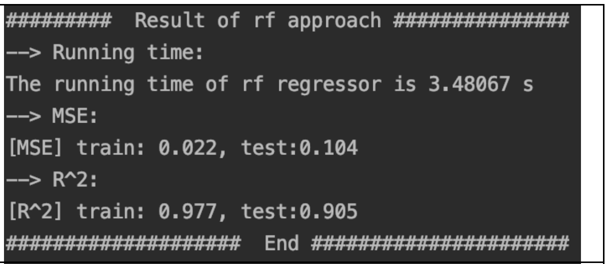
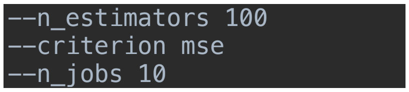
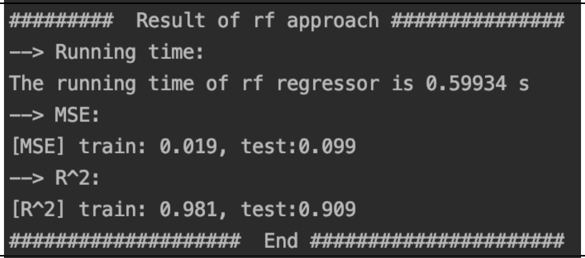
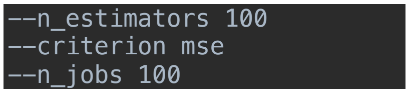
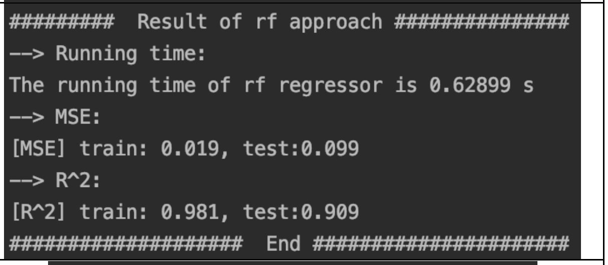
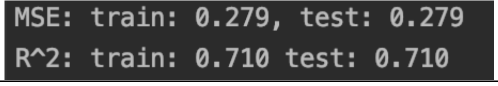
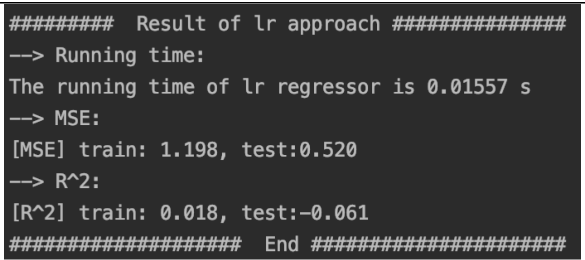
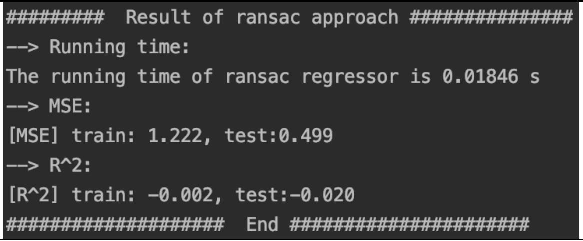
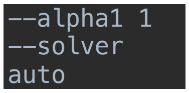
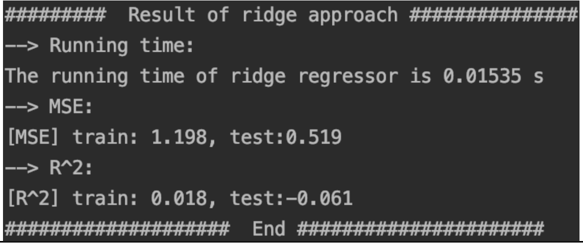
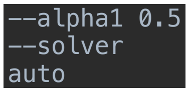
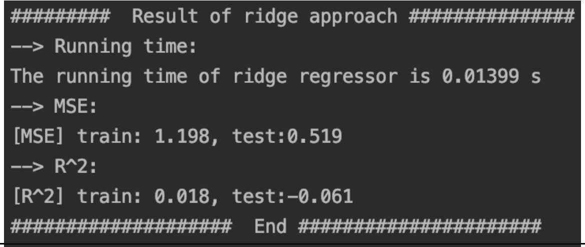
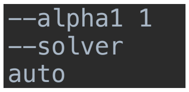
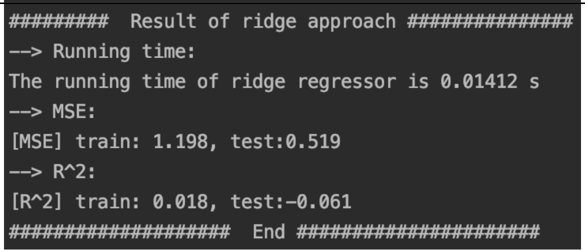
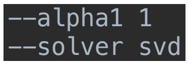
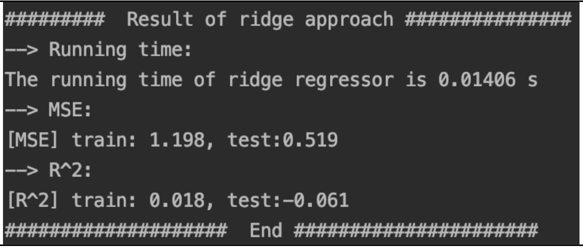
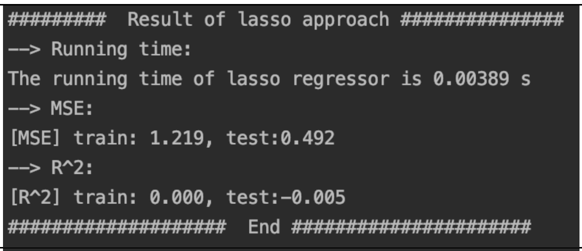
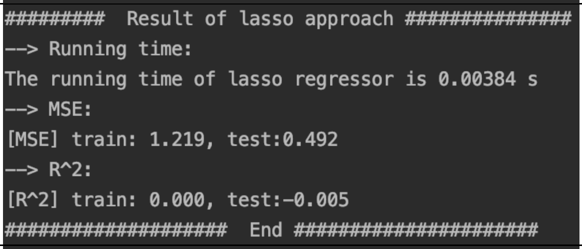
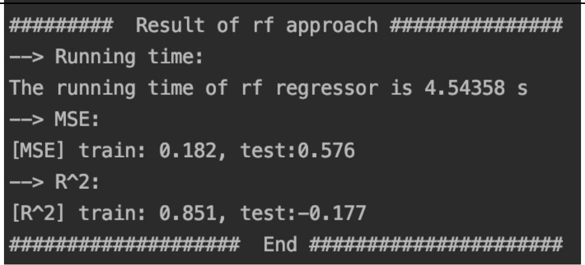
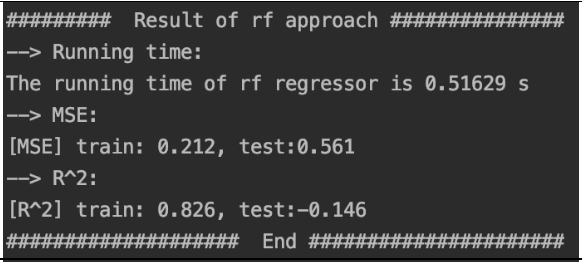
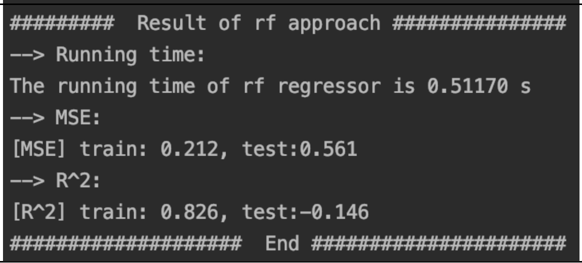
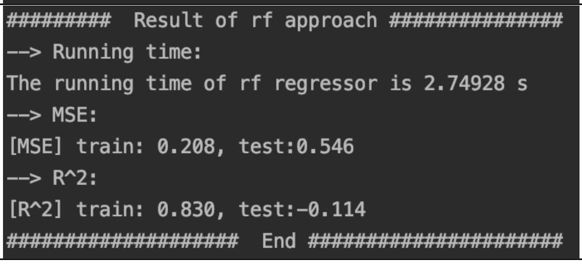
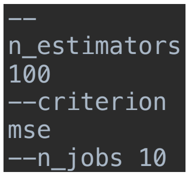
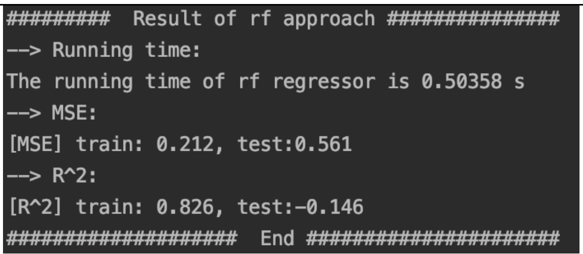
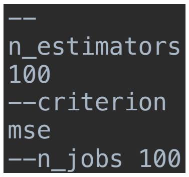
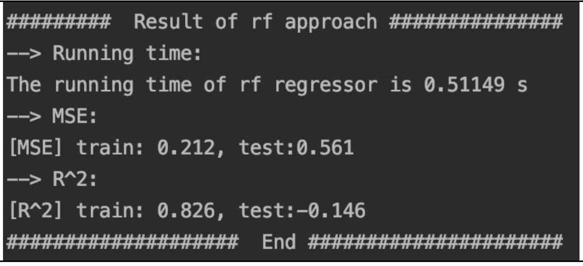
	--n_estimators 100 --criterion mae --n_jobs 10	######### Result of rf approach ############### --> Running time: The running time of rf regressor is 3.48067 s --> MSE: [MSE] train: 0.022, test:0.104 --> R^2: [R^2] train: 0.977, test:0.905 #################### End ######################
	--n_estimators 100 --criterion mse --n_jobs 10	######### Result of rf approach ############### --> Running time: The running time of rf regressor is 0.59934 s --> MSE: [MSE] train: 0.019, test:0.099 --> R^2: [R^2] train: 0.981, test:0.909 #################### End ######################
	--n_estimators 100 --criterion mse --n_jobs 100	######### Result of rf approach ############### --> Running time: The running time of rf regressor is 0.62899 s --> MSE: [MSE] train: 0.019, test:0.099 --> R^2: [R^2] train: 0.981, test:0.909 #################### End ######################
normal	default	MSE: train: 0.279, test: 0.279 R^2: train: 0.710 test: 0.710

Table 2 Performance of different regressors [CRP_new dataset]

Regressors	Parameters	Run Time and Model Evaluation Parameters
lr	default	 ######### Result of lr approach ############### --> Running time: The running time of lr regressor is 0.01557 s --> MSE: [MSE] train: 1.198, test:0.520 --> R^2: [R^2] train: 0.018, test:-0.061 #################### End ######################
ransac	default	 ######### Result of ransac approach ############### --> Running time: The running time of ransac regressor is 0.01846 s --> MSE: [MSE] train: 1.222, test:0.499 --> R^2: [R^2] train: -0.002, test:-0.020 #################### End ######################
ridge	 --alpha1 1 --solver auto	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.01535 s --> MSE: [MSE] train: 1.198, test:0.519 --> R^2: [R^2] train: 0.018, test:-0.061 #################### End ######################
	 --alpha1 0.5 --solver auto	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.01399 s --> MSE: [MSE] train: 1.198, test:0.519 --> R^2: [R^2] train: 0.018, test:-0.061 #################### End ######################
	 --alpha1 1 --solver auto	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.01412 s --> MSE: [MSE] train: 1.198, test:0.519 --> R^2: [R^2] train: 0.018, test:-0.061 #################### End ######################
	 --alpha1 1 --solver svd	 ######### Result of ridge approach ############### --> Running time: The running time of ridge regressor is 0.01406 s --> MSE: [MSE] train: 1.198, test:0.519 --> R^2: [R^2] train: 0.018, test:-0.061 #################### End ######################

lasso	--alpha2 1		 ######### Result of lasso approach ############### --> Running time: The running time of lasso regressor is 0.00389 s --> MSE: [MSE] train: 1.219, test:0.492 --> R^2: [R^2] train: 0.000, test:-0.005 ################### End #####################	
	--alpha2 0.5		 ######### Result of lasso approach ############### --> Running time: The running time of lasso regressor is 0.00384 s --> MSE: [MSE] train: 1.219, test:0.492 --> R^2: [R^2] train: 0.000, test:-0.005 ################### End #####################	
rf	--n_estimators 1000 --criterion mse --n_jobs 10		 ######### Result of rf approach ############### --> Running time: The running time of rf regressor is 4.54358 s --> MSE: [MSE] train: 0.182, test:0.576 --> R^2: [R^2] train: 0.851, test:-0.177 ################### End #####################	
	--n_estimators 100 --criterion mse --n_jobs 10		 ######### Result of rf approach ############### --> Running time: The running time of rf regressor is 0.51629 s --> MSE: [MSE] train: 0.212, test:0.561 --> R^2: [R^2] train: 0.826, test:-0.146 ################### End #####################	
	--n_estimators 100 --criterion mse --n_jobs 10		 ######### Result of rf approach ############### --> Running time: The running time of rf regressor is 0.51170 s --> MSE: [MSE] train: 0.212, test:0.561 --> R^2: [R^2] train: 0.826, test:-0.146 ################### End #####################	
	--n_estimators 100 --criterion mae --n_jobs 10		 ######### Result of rf approach ############### --> Running time: The running time of rf regressor is 2.74928 s --> MSE: [MSE] train: 0.208, test:0.546 --> R^2: [R^2] train: 0.830, test:-0.114 ################### End #####################	

	 --n_estimators 100 --criterion mse --n_jobs 10		 ######### Result of rf approach ############### --> Running time: The running time of rf regressor is 0.50358 s --> MSE: [MSE] train: 0.212, test:0.561 --> R^2: [R^2] train: 0.826, test:-0.146 #################### End ######################	
	 --n_estimators 100 --criterion mse --n_jobs 100		 ######### Result of rf approach ############### --> Running time: The running time of rf regressor is 0.51149 s --> MSE: [MSE] train: 0.212, test:0.561 --> R^2: [R^2] train: 0.826, test:-0.146 #################### End ######################	

In Tables 1 and 2, all regressions are tested by different parameters. The performances are evaluated with parameters, including training MSE+R2, testing MSE+R2.

For housing dataset:

- For lr regressors, training and testing MSE+R2 in the same level. [no overfitting]
- For ransac regressors, same result as lr regressors. [no overfitting]
- For ridge regressors, same result as lr regressors. Changing of alpha and solver do not influence the prediction results [no overfitting]
- For lasso regressors, the value of alpha influence the result significantly. The lower value of alpha makes a better prediction. But the overall MSE and R2 is very close between train and test datasets. [no overfitting]
- For rf (nonlinear) regressors, 6 sets of parameters are considered. A relatively high R2 and very low MSE are obtained. It is concluded that nonlinear regressor (RandomForest) perform much better than linear regressors for housing dataset. [no overfitting]
- For normal equation regressor, the same result of MSE and R2 will be obtained for training and testing datasets since this regressor is an analytical method. Thus, no overfitting will take place.

For CRP_new dataset:

- For lr/ransac/ridge/lasso regressors, a strong overfitting is observed since the MSE of training is higher than testing. In addition, the R2 is almost zero since the target is randomly made by us. There is no linear relationship between features and target.
- For rf nonlinear regressor, the overall MSE decrease a lot, thereby leading to a very high R2 for training dataset but a low R2 for testing, which mean a strong overfitting.

Task 4

A readme.txt file is attached.