CH5650 Molecular Data Science and Informatics
EndSem Report

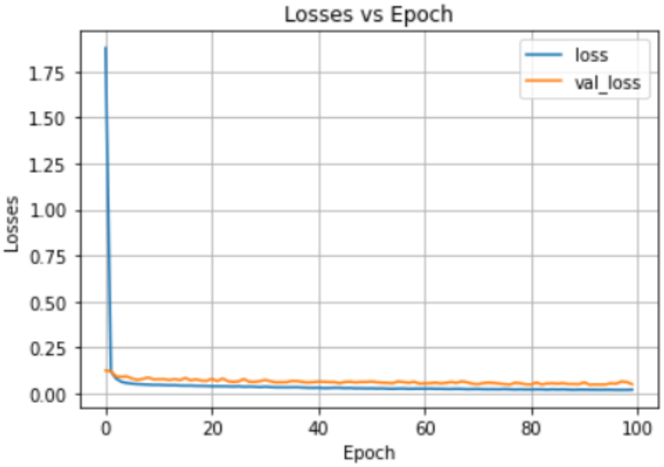
Shrivarshan K, MM20B058

Question 1:

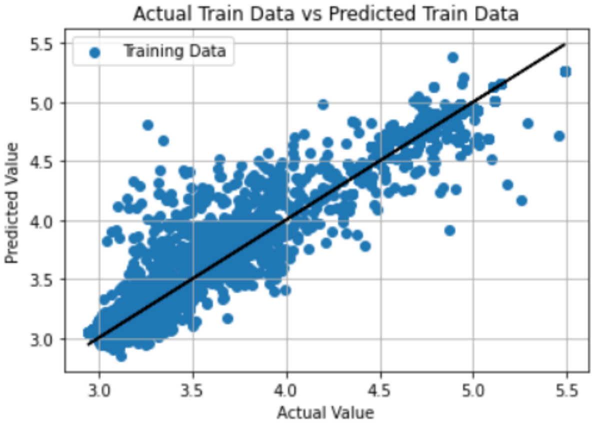
A deep neural network model is built with one hidden layer, having 10 neurons. The data is split into 80-20 for training and testing. R2_score is used for testing the model performance. The model gives the following results on training and testing data.

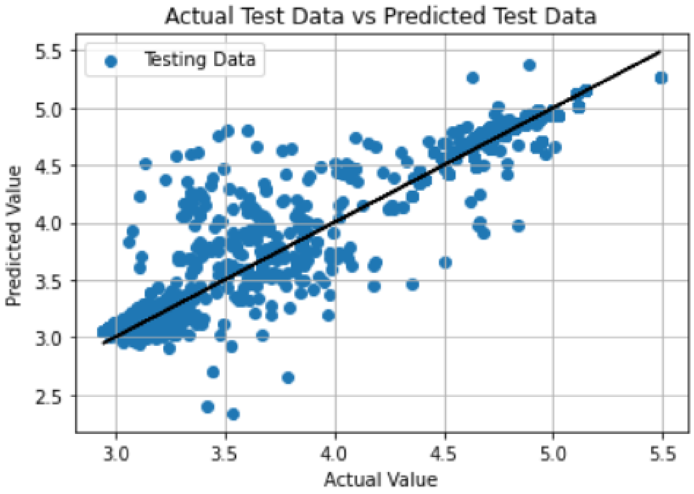
R2_score on test data = 0.9345428496837637
R2_score on train data = 0.9718136873882236

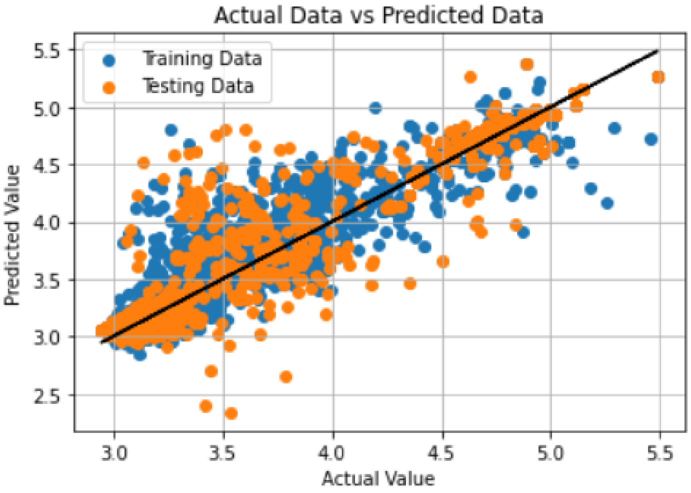
Loss v/s epoch plot:



Parity plots:

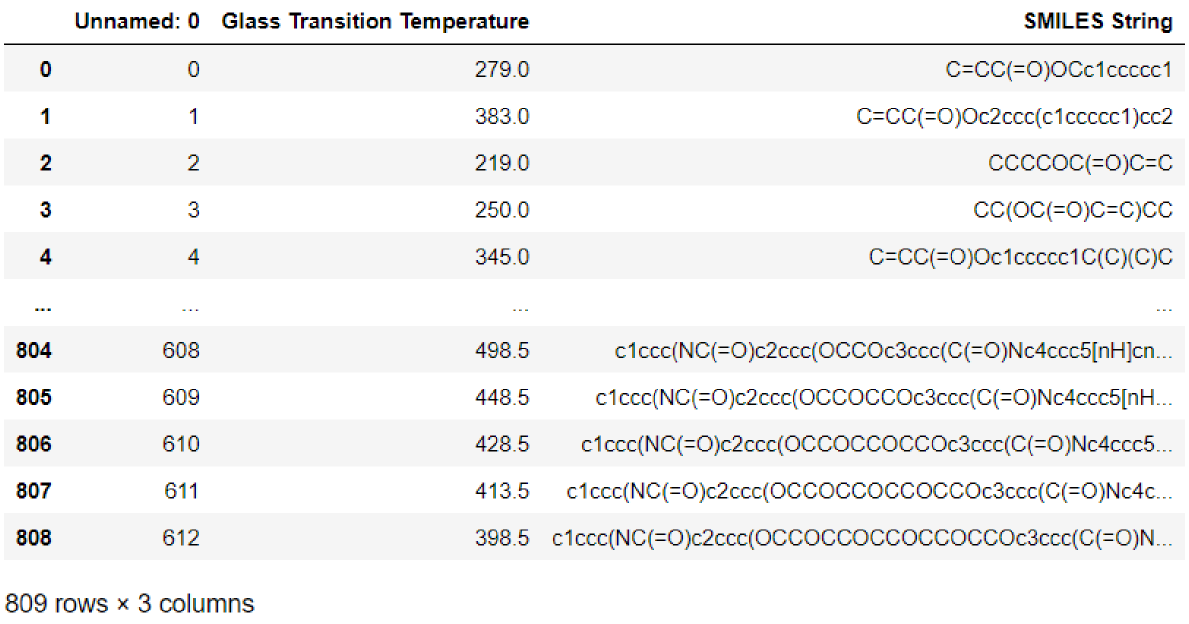






Question 2:

The dataset consists of SMILES string and the Tg (Glass Transition Temperature). We use RDKit functions to calculate the descriptor values for each molecule. This function gives us 209 descriptors, but we need we only the physicochemical descriptors. So we filter this and finally obtain 124 descriptors.



	Unnamed: 0	Glass Transition Temperature	SMILES String
0	0	279.0	C=CC(=O)OCc1ccccc1
1	1	383.0	C=CC(=O)Oc2ccc(c1ccccc1)cc2
2	2	219.0	CCCCOC(=O)C=C
3	3	250.0	CC(OC(=O)C=C)CC
4	4	345.0	C=CC(=O)Oc1ccccc1C(C)(C)C
...	...	...	...
804	608	498.5	c1ccc(NC(=O)c2ccc(OCCOc3ccc(C(=O)Nc4ccc5[nH]cn...
805	609	448.5	c1ccc(NC(=O)c2ccc(OCCOCCOc3ccc(C(=O)Nc4ccc5[nH...
806	610	428.5	c1ccc(NC(=O)c2ccc(OCCOCCOCCOc3ccc(C(=O)Nc4ccc5...
807	611	413.5	c1ccc(NC(=O)c2ccc(OCCOCCOCCOCCOc3ccc(C(=O)Nc4c...
808	612	398.5	c1ccc(NC(=O)c2ccc(OCCOCCOCCOCCOCCOc3ccc(C(=O)N...

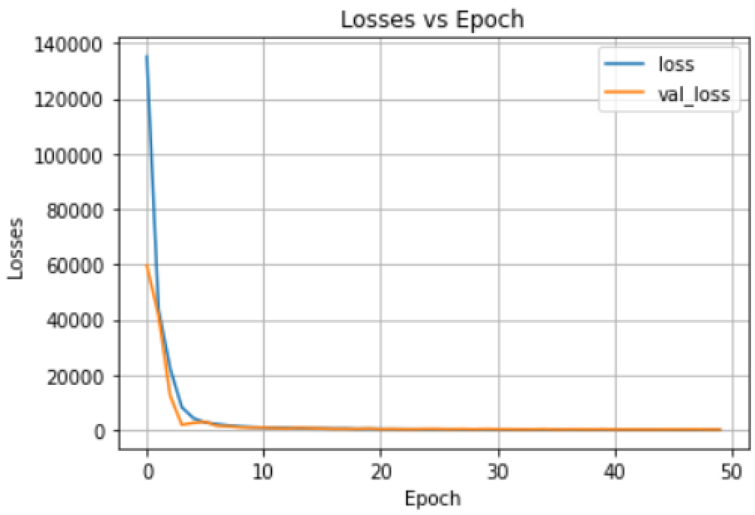
809 rows × 3 columns

Model:

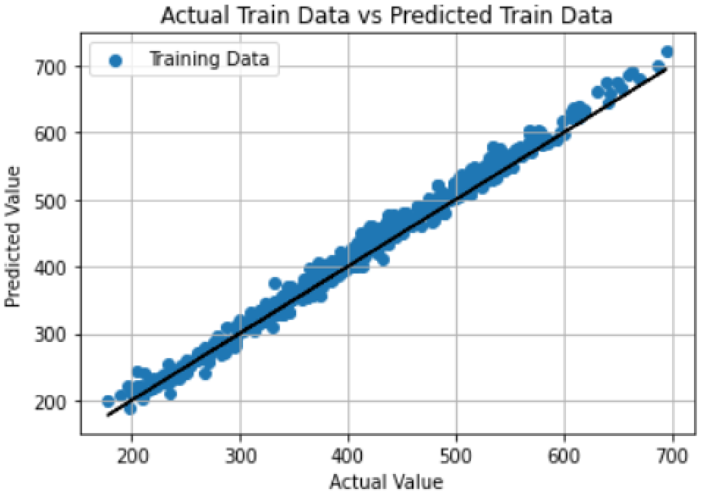
The deep neural network with 3 hidden layers, each containing 50, 20, 10 neurons is built. The data is split into 80-20 for training and testing. R2_score is used for testing the model performance. The model gives the following results on training and testing data.

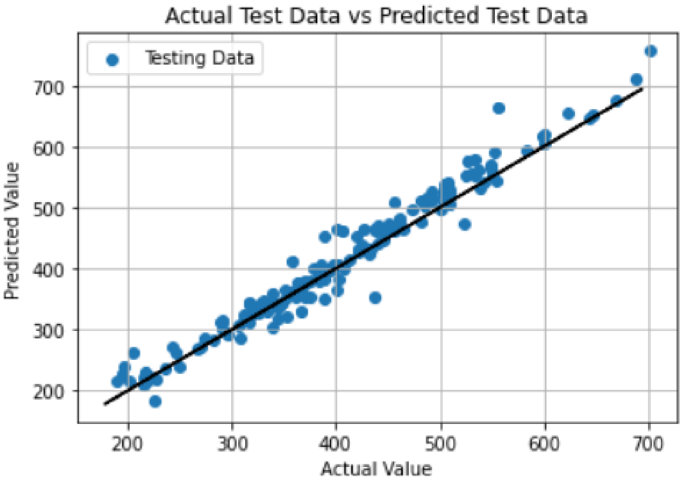
R2_score on train data = 0.9817171552244884
R2_score on test data = 0.9560919110515878

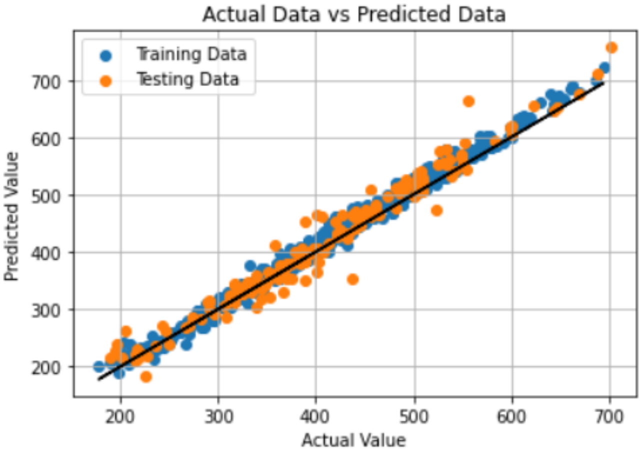
Loss v/s Epoch:



Parity plots:







Question 3:

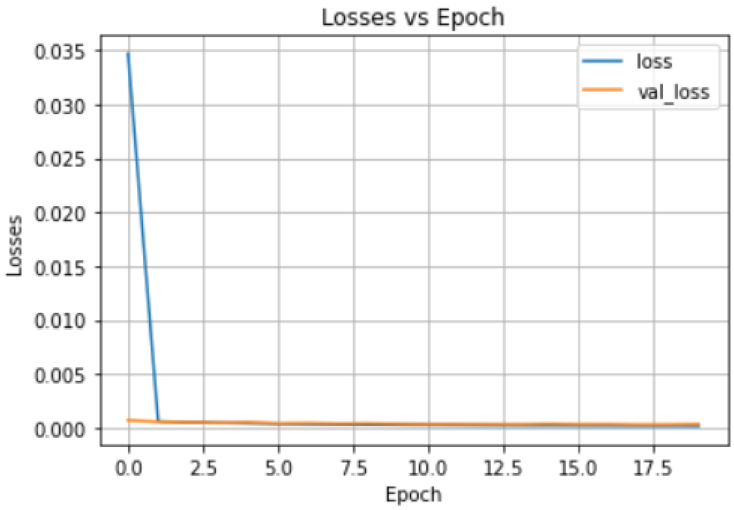
The set consists of the SMILES structure of each molecule along with the Formation Energy or HOMO-LUMO Band gap, the Enthalpy, and the Specific Heat. We use RDKit to find the structural keys. For the HOMO-LUMO band, Morgan FP was used. Morgan FP contains 1024 length binary feature vector.

HOMO LUMO Bond:

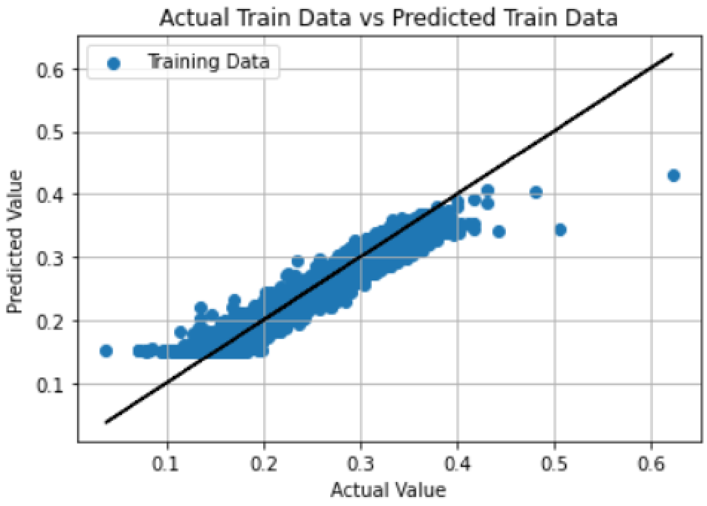
The deep neural network with 2 hidden layers, each containing 100, 50 neurons is built. A dropout layer of 0.2 is added. The data is split into 80-20 for training and testing. R2_score is used for testing the model performance. The model gives the following results on training and testing data.

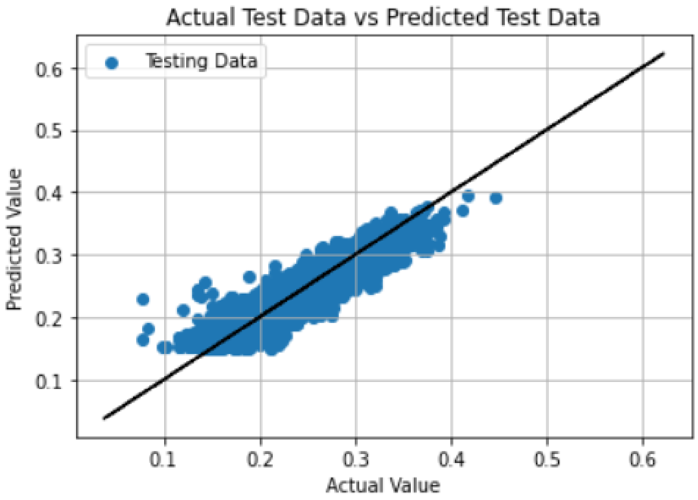
R2_score on train data = 0.9431987524046798
R2_score on test data = 0.8583064186129798

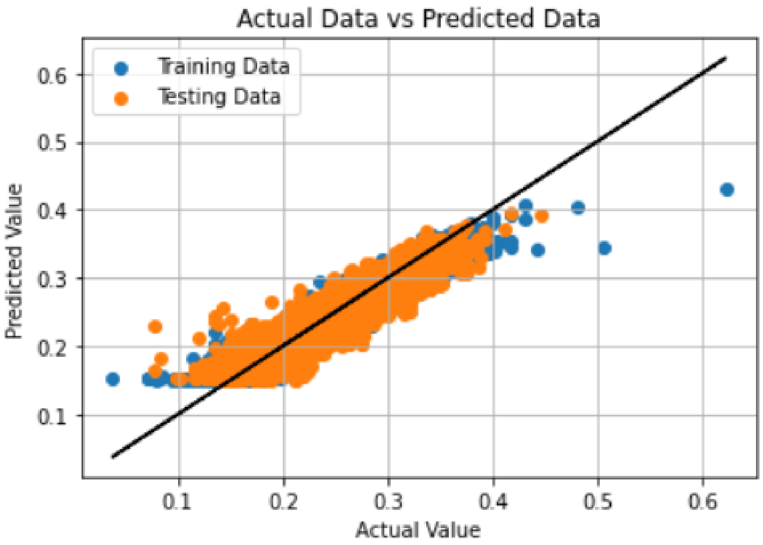
Loss v/s Epoch:



Parity plots:





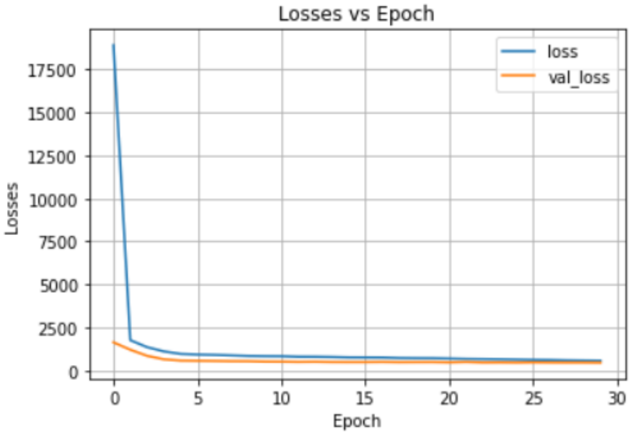


<u>Enthalpy:</u>

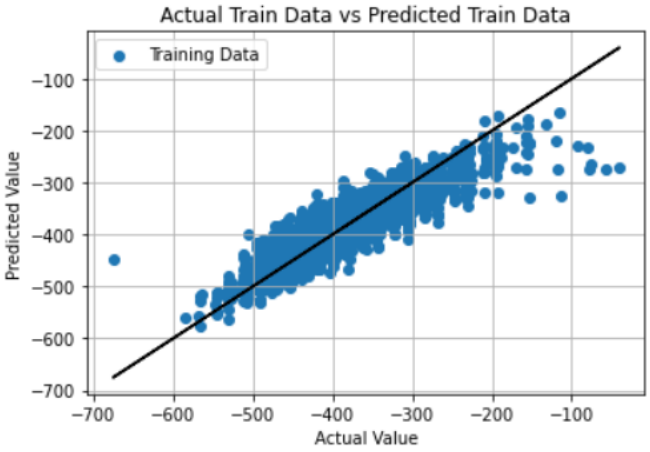
The deep neural network with 2 hidden layers, each containing 50, 30 neurons is built. A dropout layer of 0.1 is added. The data is split into 80-20 for training and testing. R2_score is used for testing the model performance. The model gives the following results on training and testing data.

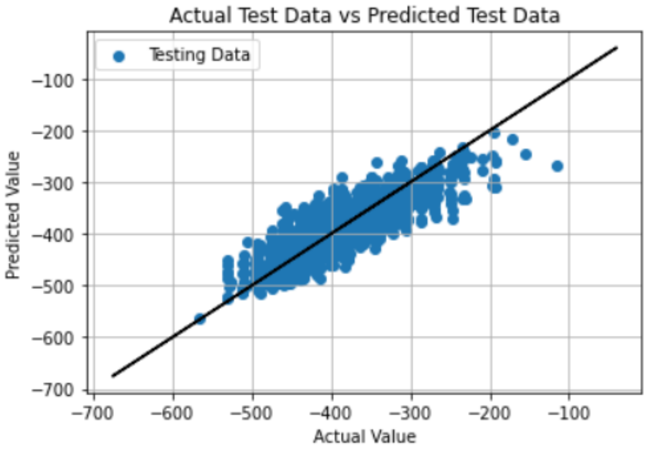
R2_score on train data = 0.8105393108787556
R2_score on test data = 0.7154025793114425

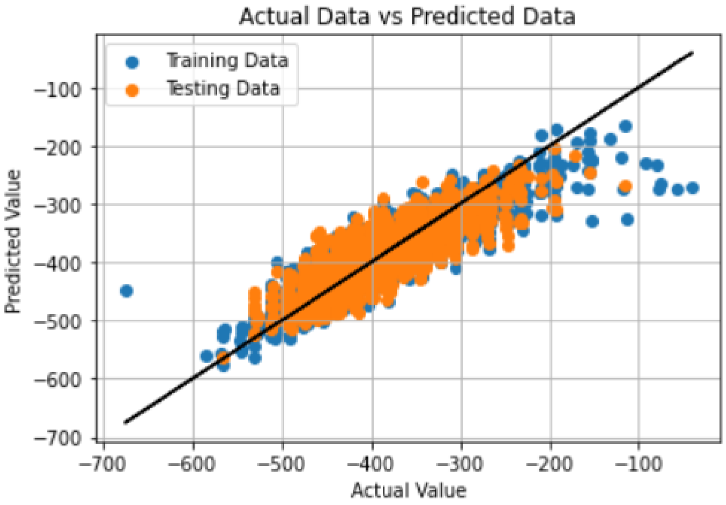
Loss v/s Epoch:



Parity plots:





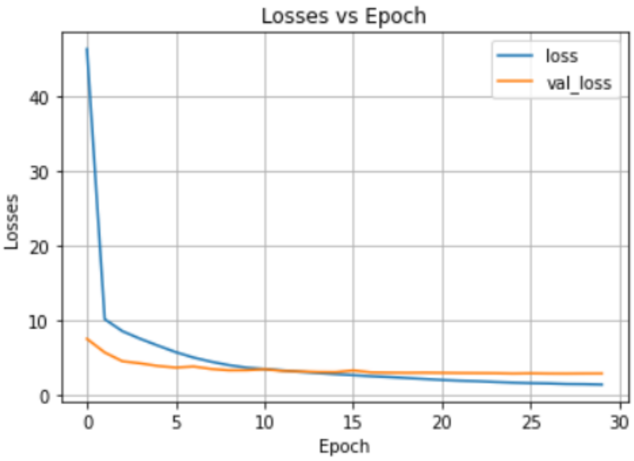


Specific Heat:

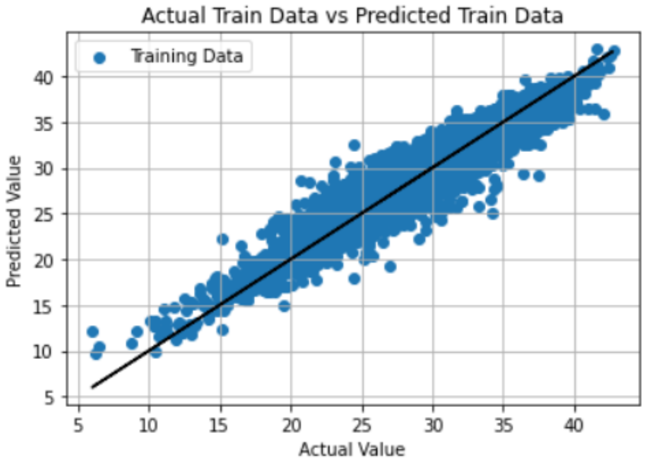
The deep neural network with 2 hidden layers, each containing 100, 42 neurons is built. A dropout layer of 0.3 is added. The data is split into 80-20 for training and testing. R2_score is used for testing the model performance. The model gives the following results on training and testing data.

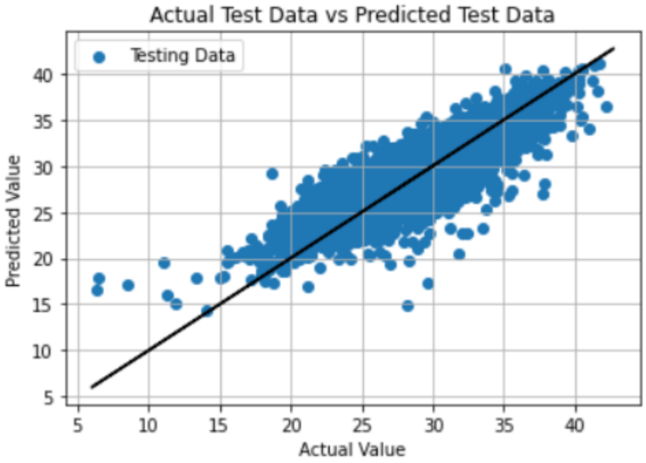
R2_score on train data = 0.9119541468312503
R2_score on test data = 0.6557644656113509

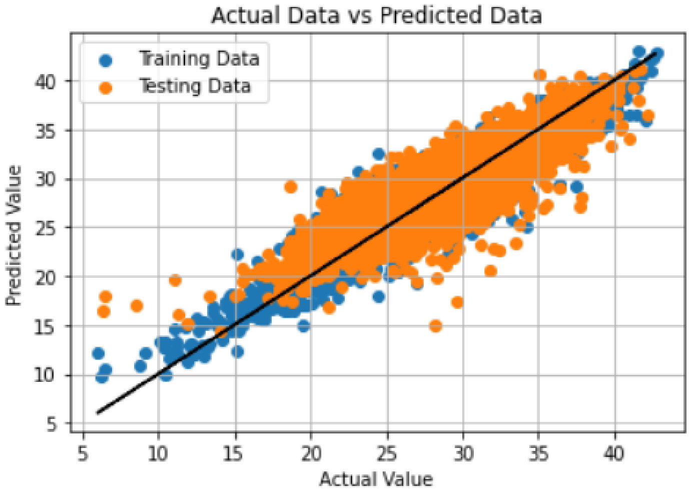
Loss v/s Epoch:



Parity plots:


Actual Train Data vs Predicted Train Data
Training Data
Predicted Value
Actual Value


Actual Test Data vs Predicted Test Data
Testing Data
Predicted Value
Actual Value


Actual Data vs Predicted Data
Training Data
Testing Data
Predicted Value
Actual Value