ASSIGNMENT 2

Description of the dataset:

The dataset I chose is Visualizing Soil. There are four features, of which three are numeric and, one is nominal, one target in this ML task. The features are northing, easing, resistivity and isns. The target is to keep track of soil properties given those four features. There is a total of 8641 instances.

Configurations of the methods:

There are no instances with the missing attributes.

- **Decision Tree Regressor:** For the parameter tuning, the min_samples_leaf values taken are 2,4,6,8,10. 10 folds have been used for cross-validation. Also, 20%, 40%, 60%, 80%, and 100% of training data were used for the learning curve.

- **KNearestNeighbors Regressor:** For the parameter tuning, the number of neighbors for each sample chosen was 14. 10 folds have been used for cross-validation. Also, 20%, 40%, 60%, 80%, and 100% of training data were used for the learning curve.

- **Linear Regressor:** 10 folds have been used for cross-validation. Also, 20%, 40%, 60%, 80%, and 100% of training data were used for the learning curve.

- **SVM Regressor:** 10 folds have been used for cross-validation. Also, 20%, 40%, 60%, 80%, and 100% of training data were used for the learning curve.

- **Bagged decision tree regressor:** 10 folds have been used for cross-validation. Also, 20%, 40%, 60%, 80%, and 100% of training data were used for the learning curve.

- **Dummy Regressor:** 10 folds have been used for cross-validation. Also, 20%, 40%, 60%, 80%, and 100% of training data were used for the learning curve.

Results:

Table of RMSE values:

Method Name	RMSE mean value
Decision Tree Regressor	2.171264828951403
KNearestNeighborsRegressor*	4.989914688914912
Linear Regression*	5.281030818910198
Support Vector Machine*	11.487668809471277
Bagged Decision Tree	**1.821223253678928**
Dummy Regressor*	12.506713759161693

In the above table, KNN (p-value = 0.000714149071145302), Linear Regression (p-value=0.004781747844760739) , Support Vector Machine (p-value = 3.5*10^-5), Dummy Regressor (p-value = 0.000301382108953) are statistically significantly different when compared to Bagging Decision Tree. The decision tree's p-value when compared is 0.087 hence it is not statistically significantly different.

Graph of all learning curves:

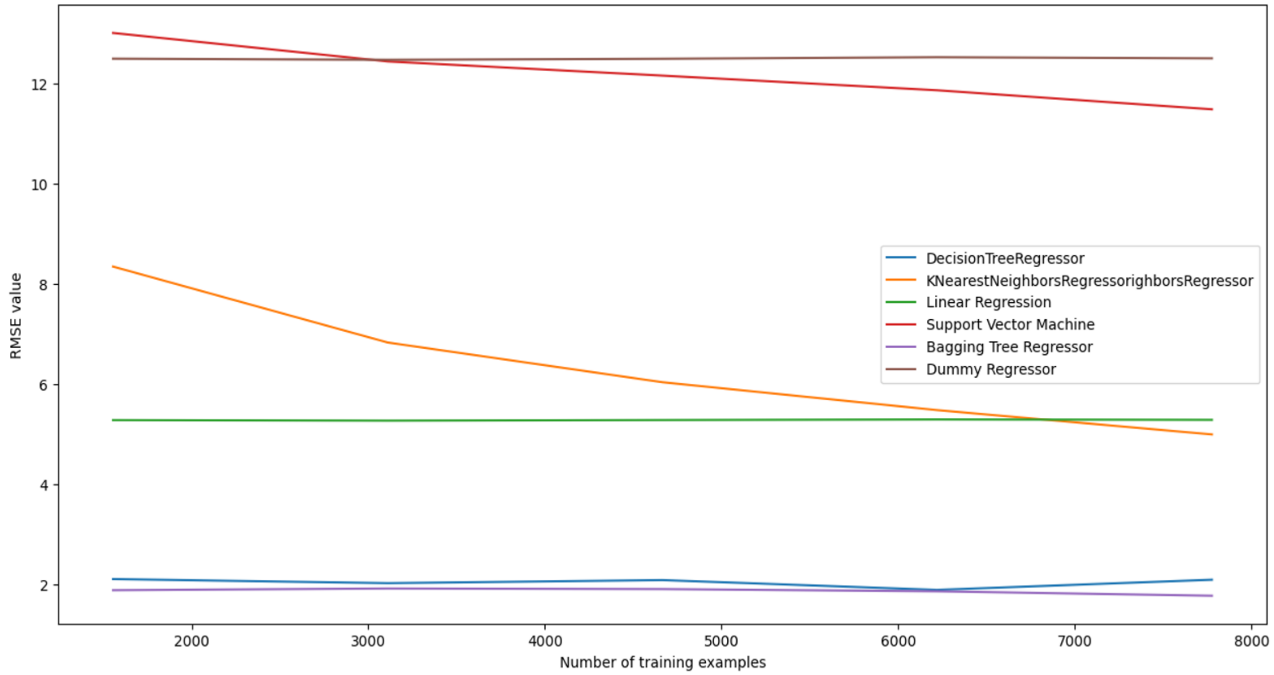

Table showing computational training and test times for the last point of learning curve:

Method Name	Computational Training time (in secs)	Computational Testing time (in secs)
Decision Tree Regressor	0.011427927017211913	0.0005125522613525391
KNearestNeighborsRegressor	0.006190180778503418	0.005892395973205566
Linear Regression	0.0021967172622680666	0.0010954618453979492
Support Vector Machine	2.0887311697006226	0.5108922243118286
Bagged Decision Tree	0.07193000316619873	0.002098274230957031
Dummy Regressor	0.0005994796752929688	0.0002997159957885742

Discussion:

From the learning curves, it can be observed that, for the initial 20% of training data, the RMSE value was high for all the models. As the training examples are increased, the RMSE value slowly decreases. For the 100% data, the Bagged Decision tree performed better than the rest of the models. The decision tree performed almost similar to the Bagged Decision tree. Whereas the dummy regressor performed least when compared to other models.

The same can be observed from the rmse values table. From the RMSE mean values, the RMSE mean value for the Bagged Decision tree is better since a lower RMSE value indicates a better fit, whose value is 1.82. The decision tree regressor also did perform closely with the RMSE mean value of 2.17. The Dummy regressor has the highest RMSE value of 12.5, and it seemed to be the least fitting for my dataset.

The computational training time for the dummy regressor is low even though it didn't perform well. The computational training time for the SVM regressor is high. The computational testing time is also low for Dummy Regressor. In comparison, it is high for the SVM regressor.