A brief description of the dataset (what is the task, what are the features and the target)

N/B; An API to the UCI Machine Learning Repository is required to fetch the dataset by running; !pip3 install -U ucimlrepo. Source - https://archive.ics.uci.edu/dataset/1/abalone

Dataset is Abalone (ID: 183). Number of Instances: 4177, Number of Features: 8 (7 numerical and one nominal). Features include;

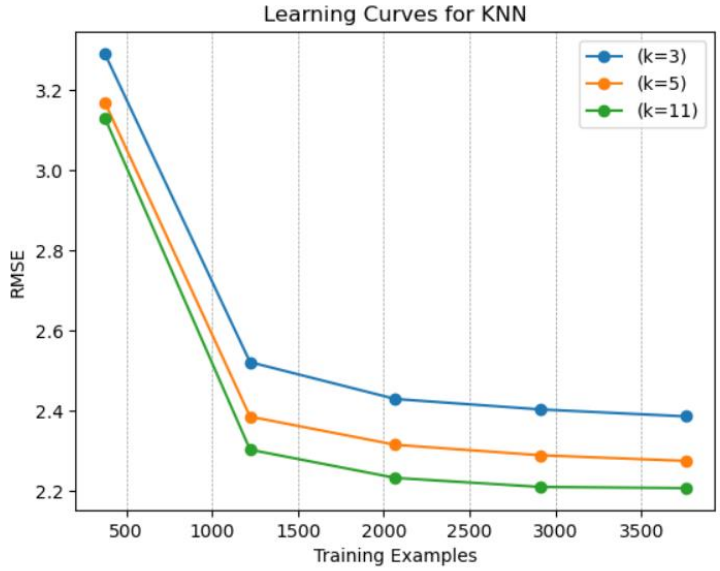
```
In [22]: # variable information
         abalone.variables
```

Out[22]:

	name	role	type	demographic	description	units	missing_values
0	Sex	Feature	Categorical	None	M, F, and I (infant)	None	no
1	Length	Feature	Continuous	None	Longest shell measurement	mm	no
2	Diameter	Feature	Continuous	None	perpendicular to length	mm	no
3	Height	Feature	Continuous	None	with meat in shell	mm	no
4	Whole_weight	Feature	Continuous	None	whole abalone	grams	no
5	Shucked_weight	Feature	Continuous	None	weight of meat	grams	no
6	Viscera_weight	Feature	Continuous	None	gut weight (after bleeding)	grams	no
7	Shell_weight	Feature	Continuous	None	after being dried	grams	no
8	Rings	Target	Integer	None	+1.5 gives the age in years	None	no

The dataset is about predicting the age (the target, computed by adding 1.5 to Rings column) of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. Other measurements, which are easier to obtain, are used to predict the age.

Results for Task 1 in the form of graph and table.

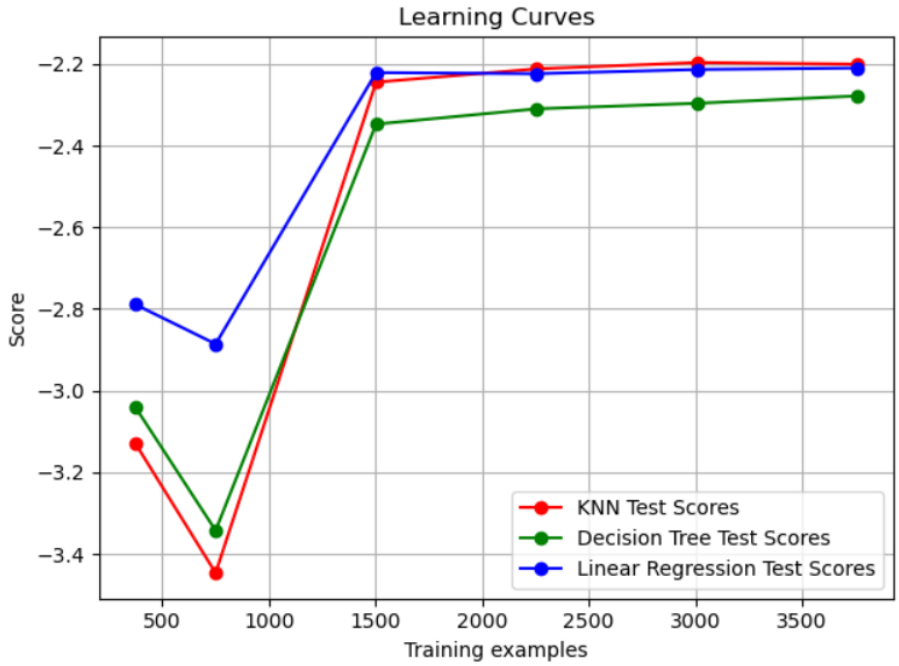



```
Task 1: Comparison of k-nearest neighbor regressor (k values)
            RMSE
(k=3)   2.380214
(k=5)   2.267836
(k=11)  2.200232
```

Discussion on the results of Task 1.

The learning curves plot the relationship between the number of training examples and the RMSE on the test set for different values of k (3, 5, and 11). We can see as the number of training examples increases the RMSE decreases for all three k values and the RMSE decreases as the value of k increases. The k value of 11 seems to have the lowest RMSE across different training sizes indicating its potential suitability for the regression task compared to the other k values. The curve eventually plateaus at around 2100 training examples for all the k values.

Results for Task 2 in the form of graph and table.




```
Best parameters for KNN: {'n_neighbors': 11}
Best parameters for Decision Tree: {'min_samples_leaf': 20}

RMSE for best parameters:
                      RMSE
KNN               2.153483
Decision Tree     2.265476
Linear Regression 2.209699
```

Discussion on the results of Task 2.

From the plot we observe plateaus in the learning curves of all models at later stages but we can see KNN has a steep initial decrease in rmse score compared to the other two which shows that it learns faster with few training examples. The plateaus later converge for KNN and linear regression meaning the models eventually achieve similar levels of performance as they are provided with more training data. It also shows that with sufficient data, the models tend to capture the underlying patterns in the data effectively, regardless of their differences.