Project : Youtube Adview Prediction

5. Normalise your data and split the data into training, validation and test set in the appropriate ratio.

We need to split the data into training and test data. We use training data to learn patterns in the data and then check if it generalises well on unseen data. The split percentage can be varied and is generally on the amount of data we have. For a relatively small dataset like this, the split percentage should be high (80:20 rather than 99:1 with respect to train:test).

Normalisation is done to ensure all the features all weighted appropriately in the training stage. Just because some features have high scale should not translate to having higher influence on the model. Normalisation can be done using Standard Scalar or MinMax Scalar among others. In this particular problem, MinMax Scalar has been used which basically transforms each variable in the range of 0 to 1.

- [ ] Split dataset in train and test as well as into inputs and outputs
- [ ] Normalise the dataset using scalars

```
# Split Data
Y_train  = pd.DataFrame(data = data_train.iloc[:, 1].values, columns = ['target'])
data_train=data_train.drop(["adview"],axis=1)
data_train=data_train.drop(["vidid"],axis=1)
data_train.head()

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data_train, Y_train, test_size=0.2, random_state=42)

X_train.shape

# Normalise Data
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.fit_transform(X_test)

X_train.mean()
```