Exercise 1: Analyzing Second Hand Car Sales Data with Supervised and Unsupervised Learning Models.

a. In second hand car sales dataset for each car sold, it contains the information about Manufacturer (the name of the manufacturer that produced the car), Model (the name of the model of the car), Engine size (the size of the engine, in litres), Fuel type (the type of fuel that the engine uses), Year of manufacture (the year in which the car was made), Mileage (the total number of miles that the car has been driven), Price (the price that the car was sold for, in Pound Sterling (GBP)). Among all of these features variables Engine size, Year of manufacture, and Mileage are the numerical feature input. When it was run a linear regression model among all of these numerical feature variables, year of manufacture gives the better price prediction. It was calculated the R-squared value for the three linear regression model and got the following results-

```
The R-square Score in Linear Regression using

Engine size = 0.1579553281999313
Year of manufacture = 0.510984724303591
Mileage = 0.3983649608124491
```

Since, The R-squared value of the year of manufacture larger than the other two, it can be concluded that Year of Manufacture gives the better prediction compared to the engine size and mileage.

For each numerical feature variable, it was built a polynomial regression model (non-linear regression). Since the prediction made on a single feature variable, the same R-squared value was calculated like the linear regression models.

```
The R-square Score in Polynomial Regression using

Engine size = 0.1579553281999313
Year of manufacture = 0.510984724303591
Mileage = 0.3983649608124491
```

b. A linear regression model and a polynomial regression model was also built using of all numerical feature input (Engine size, Year of manufacture, and Mileage). Then the prediction accuracy increased horribly. The R-squared score of these models attached below:

```
R-squre Score in Linear Regression = 0.6749798831772409
R-square Score in Polynomial Regression = 0.8941503326884358
```

From the R-square value of the Polynomial regression it can be said that, it gives the better prediction accuracy compared to linear regression model. The inclusion of multiple input features improves the accuracy of the model's prediction compared to the single-input feature models that was explored in part (a).

c. In this phase, a Random Forest Regression model was built using both categorical and numerical input features variable. It gave the R-square value 0.9984705869992906. In (b), the R-square value in polynomial regression was 0.89415. So, it can be said that this surpassed the accuracy achieved by polynomial regression, showcasing the importance of considering all features.

```
R-square score in Random Forest Regression model: 0.9984705869992906
```

d. Architecture of the Artificial Neural Network (ANN) model-

Input Layer: The input layer has 128 nodes, and it uses the ReLU (Rectified Linear Unit) activation function.

Hidden Layer: There is one hidden layer with 64 nodes and a ReLU activation function. This layer helps the model learn complex patterns from the input data.

Output Layer: The output layer has 1 node, indicating that this is a regression task (predicting the car price, a continuous variable). The activation function is linear, suitable for regression tasks.

Hyperparameter Tuning:

Optimizer: Used the Adam optimizer, because it is a popular choice for gradient-based optimization algorithms.

Loss Function: The mean squared error (MSE) is used as the loss function.

Epochs: The model is trained for 50 epochs, meaning it goes through the entire training dataset 50 times during training.

Batch Size: During each epoch, the training dataset is divided into batches of size 32. The weights of the model are updated after processing each batch.

Validation Split: 10% of the training data is used as validation data to monitor the model's performance during training and prevent overfitting.

In the Artificial Neural Network (ANN) model, R-square value was calculated 0.9990076753145802 whereas the R-square value of random forest regression (supervised learning) was 0.9984705869992906. So, it can be concluded that in this case ANN gives the better accuracy to predict the price of a car.

e. Among linear regression model, polynomial regression model, random forest model and Artificial Neural Network (ANN) model, ANN makes the best prediction compared to others.

```
R-squre Score in Linear Regression = 0.6749798831772409
R-square Score in Polynomial Regression = 0.8941503326884358
R-square score in Random Forest Regression model: 0.9984705869992906
R-square score in ANN model: 0.9990076753145802
```

That's why ANN is the best model for predicting the price of a car.

f. Using all numerical input features, k-mean clustering was built for k equals 2 to 10. For all of these cluster, inertia, silhouette and Davies Bouldin score was measured. We have known that,

Inertia (within-cluster sum of squares): The lower, the better.

Silhouette score: A higher score indicates better-defined clusters.

Davies-Bouldin Index: A lower score indicates better clustering.

Analyzing these scores, it can be said that cluster 3 is the optimal cluster.

```
Clusters: 2, Inertia: 116530.87591309159, Silhouette: 0.39735675661460274, Davies-Bouldin: 0.9841527072766522
Clusters: 3, Inertia: 87492.31979501166, Silhouette: 0.4093777993502264, Davies-Bouldin: 0.957912509564022
Clusters: 4, Inertia: 69304.38347564946, Silhouette: 0.328995478659553, Davies-Bouldin: 0.9876761833564394
Clusters: 5, Inertia: 54699.89629693521, Silhouette: 0.343744398210847, Davies-Bouldin: 0.8953580078939852
Clusters: 6, Inertia: 47052.289475394784, Silhouette: 0.29801437627490474, Davies-Bouldin: 0.9524050784773186
Clusters: 7, Inertia: 42343.923009551996, Silhouette: 0.27303272152241786, Davies-Bouldin: 1.0097146049595984
Clusters: 8, Inertia: 38495.91528808992, Silhouette: 0.2800694520350489, Davies-Bouldin: 1.0132378712529535
Clusters: 9, Inertia: 35293.32738858744, Silhouette: 0.28069235267142595, Davies-Bouldin: 1.0184967931412017
Clusters: 10, Inertia: 32535.28853946303, Silhouette: 0.28223384257860057, Davies-Bouldin: 1.0120142329281643
```

K-means cluster was also built using only two numerical feature input (engine size and mileage) for cluster k = 2 to 10. The evaluation metrices was looks-

```
k-means clustering using 2 numerical feature input
Clusters: 2, Inertia: 98091.45262098621, Silhouette: 0.3669037012204407, Davies-Bouldin: 1.1078968604155395
Clusters: 3, Inertia: 67640.20684663932, Silhouette: 0.3837362495977886, Davies-Bouldin: 0.9464029615658083
Clusters: 4, Inertia: 53355.18029571917, Silhouette: 0.3939015989244124, Davies-Bouldin: 0.8875475042141868
Clusters: 5, Inertia: 40952.017640387996, Silhouette: 0.3551896769108442, Davies-Bouldin: 0.865004339535966
Clusters: 6, Inertia: 34662.87493969449, Silhouette: 0.3188912041933196, Davies-Bouldin: 0.8978029803587639
Clusters: 7, Inertia: 29914.267549593955, Silhouette: 0.3340851222070231, Davies-Bouldin: 0.8437406928459491
Clusters: 8, Inertia: 26899.462637665187, Silhouette: 0.3110482480099364, Davies-Bouldin: 0.8780938100030149
Clusters: 9, Inertia: 24156.625624845357, Silhouette: 0.323078080141373, Davies-Bouldin: 0.8524310418261667
Clusters: 10, Inertia: 21934.467389429134, Silhouette: 0.3224778871061413, Davies-Bouldin: 0.8701647875029245
```

In this case cluster 3 is the optimal cluster.

Comparing the two sets (for feature input 2 and feature input 3) of clustering results:

1. **First Set:** For Clusters 3, Inertia: 87492.32, Silhouette: 0.4094, Davies-Bouldin: 0.9579
2. **Second Set:** For cluster 3, Inertia: 67640.21, Silhouette: 0.3837, Davies-Bouldin: 0.9464

In this case: The second set has a lower inertia, which is generally favorable, the first set has a higher silhouette score, indicating better-defined clusters and the second set has a slightly lower Davies-Bouldin index, suggesting better clustering. Considering all three metrics, the second set with Clusters: 3, Inertia: 67640.21, Silhouette: 0.3837, and Davies-Bouldin: 0.9464 seems to be the more optimal choice as it exhibits a good balance between lower inertia and acceptable values for silhouette and Davies-Bouldin index.

g. Comparing between K-means clustering and Agglomerative Clustering (hierarchical clustering), Agglomerative Clustering is the best clustering. Because analyzing the silhouette and Davis Bouldin score of these cluster, it can be concluded that hierarchical clustering produces the best clustering.

```
Hierarchical Clustering:
Silhouette: 0.38276355502401527, Davies-Bouldin: 0.9219150632524684

K-means Clustering:
Silhouette: 0.4016670124291188, Davies-Bouldin: 0.9190927767103263
```