ARTIFICIAL INTELLINGENCE IN THE ANALYSIS OF LIFE EXPECTANCY

Analysis of the life expectancy in different countries using Supervised and Unsupervised machine learning models.

ABSTRACT

The longevity of life in a country is dependent on a lot of factors affecting the country. These factors which includes the rate of adult mortality, infants' deaths, the economic status of the country in terms of the country's Gross Domestic Product (GDP), and the development status of the country i.e., Developing or Developed, etc. This scientific paper provides a comprehensive analysis using supervised hyper parametric regression models and unsupervised hyper parametric clustering models to determine the correlations between the life expectancy and development status, adult mortality, infant's death, GDP and expectancy class for various countries for the year 2015. The analysis indicated that while adult mortality gave a good correlation and prediction to the life expectancy of a country, combining all features together still had a slight edge on the correlation with the life expectancy of the country. The analysis also indicated that K-Mean clustering model produced the best fit for life expectancy.

KEYWORDS:

Supervised learning model, Unsupervised learning model, Life expectancy, GDP, Adult Mortality, Infant deaths, Simple regression model, Polynomial regression model, K-Means clustering algorithm, K-Means clustering, Agglomerative Hierarchical clustering, DBSCAN model.

1. INTRODUCTION

Machine learning is when an agent in this case, a computer, learns and improves performance after making deductions and observations (Richings, et al., 2023). There are 3 types of machine learning which are supervised learning, unsupervised learning and reinforcement learning. In this scientific paper, we shall use supervised and unsupervised learning to determine the life expectancy of various countries for a year. Supervised learning is further divided into sections, which are "regression" and "classification" (Richings, et al., 2023). Regression's output is continuous while classification output is a finite set of classes (Richings, et al., 2023).

Unsupervised learning is a machine learning approach to analysing and clustering unlabelled datasets.(Richings, et al., 2023)

2. METHODOLOGY

The approach to this analysis is in two parts, which are the supervised learning model (regression models) and the unsupervised learning models (clustering models). Jupyter notebook is the medium used for analysis. The processes involve importing all the necessary libraries (csv, pandas, seaborn, matplotlib, NumPy, sklearn) and using pandas to load dataset. Exploratory data analysis was performed on the life expectancy dataset presented; this is to ensure that the dataset is clean before carrying out analysis on it.

2.1. Supervised learning models to predict life expectancy.

2.1.1. Simple linear regression model

Simple linear regression model: (i) GDP, (ii) Adult Mortality, (iii) Infant deaths was executed on jupyter notebook where the independent variable (x) takes the input feature columns of the data frame, and the dependent variable takes the Life expectancy columns. Since the input variable is an independent variable, it will be reshaped. The reshaped data set is then split into training set and testing set using the sklearn library. The application of standardization is performed on the independent variable (x) to ensure that the Mean of the data set is zero and the Standard deviation is one. After training the model, model evaluation is performed and the performance metrics which are mean absolute error (mae), mean squared error, root mean squared error (rmse) and coefficient of determination (cod) are evaluated. (Richings, et al., 2023)

2.1.2. Polynomial regression model

This is performed by importing the polynomial features from the sklearn library after the input feature column of "GDP" has been split into train and test data, scaled and standardized, and performed its correlation to life expectancy. This scientific paper explored the different value of polynomial degree from 2 to 5. Model evaluation for performance metrics is carried out. (Richings, et al., 2023)

2.1.3. Multiple input regression model

While the above sections involved the single input feature on the regression model, this work went further by exploring the options of multiple input feature of "GDP", "Adult Mortality", "infant deaths" and its correlation to life expectancy. Since it is a multiple input feature, reshaping is not required. The combined input is split into train and test data set, scaled, standardized and performed its correlation to life expectancy. Then the model evaluations for performance metrics is obtained. K-Nearest Neighbours model was used to predict life expectancy class from the above input features. (Richings, et al., 2023)

2.2. Unsupervised learning models to predict life expectancy.

2.2.1 K-Mean Cluster Algorithm

The K-Mean cluster algorithm was used to identify clusters of countries based on their life expectancy vs GDP. Elbow method was used to determine the optimal value of k. K-Mean was also performed on life expectancy vs Adult Mortality. The two clusters' models were compared to find the best clustering for life expectancy. (Richings, et al., 2023)

2.2.2 Agglomerative Hierarchical Clustering & DBSCAN models

Agglomerative hierarchical clustering and DBSCAN model was performed on life expectancy dataset using life expectancy vs GDP and comparative correlation was done to determine which of the three models produces the best clustering for this data. (Richings, et al., 2023)

3. RESULTS & DISCUSSION

Figure 1 shows the loaded data frame of life expectancy data after all the necessary libraries had been imported into the jupyter notebook. Exploratory data analysis showing the visual representation of the life expectancy data and ensuring that the dataset is clean and fit for analysis.

Out[4]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	GDP	Life expectancy class
0	Afghanistan	2015	Developing	65.0	263.0	62	584.259210	Low
1	Albania	2015	Developing	77.8	74.0	0	3954.227830	High
2	Algeria	2015	Developing	75.6	19.0	21	4132.762920	Medium
3	Angola	2015	Developing	52.4	335.0	66	3695.793748	Low
4	Antigua and Barbuda	2015	Developing	76.4	13.0	0	13566.954100	High
...	...	...	...	...	...	...	...	...
178	Venezuela (Bolivarian Republic of)	2015	Developing	74.1	157.0	9	4110.000000	Medium
179	Viet Nam	2015	Developing	76.0	127.0	28	2600.000000	High
180	Yemen	2015	Developing	65.7	224.0	37	1490.000000	Low
181	Zambia	2015	Developing	61.8	33.0	27	1313.889646	Low
182	Zimbabwe	2015	Developing	67.0	336.0	22	118.693830	Low

183 rows × 8 columns

Figure 1. Loaded data set to pandas' data frame.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183 entries, 0 to 182
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Country                183 non-null    object
 1   Year                   183 non-null    int64
 2   Status                 183 non-null    object
 3   Life expectancy        183 non-null    float64
 4   Adult Mortality        183 non-null    float64
 5   infant deaths          183 non-null    int64
 6   GDP                    183 non-null    float64
 7   Life expectancy class  183 non-null    object
dtypes: float64(3), int64(2), object(3)
memory usage: 11.6+ KB
```

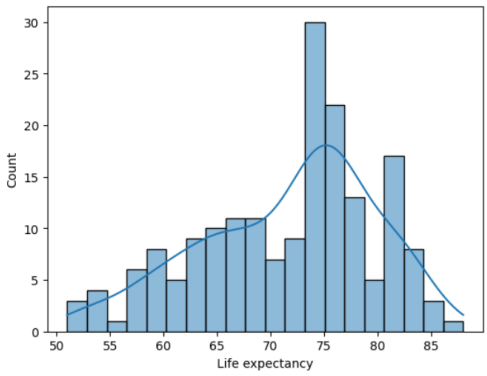
Figure 2a Datatype of feature column

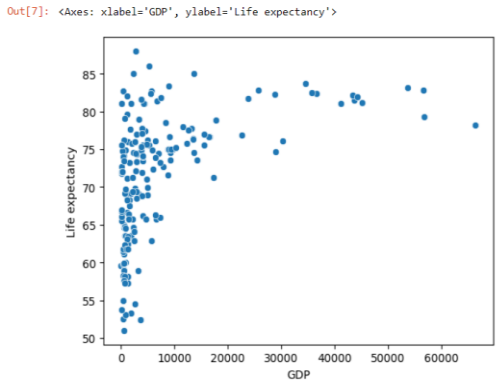
```
In [3]: #Check for missing values in the dataframe.
        life_dfr.isna().sum()

Out[3]: Country                0
        Year                   0
        Status                 0
        Life expectancy        0
        Adult Mortality        0
        infant deaths          0
        GDP                    0
        Life expectancy class  0
        dtype: int64
```

Figure 2b Empty cells in column

Figure 3(a) Histogram showing mean of life expectancy (b) Scatterplot of life expectancy vs GDP (c) Scatterplot of life expectancy vs adult mortality (d) Scatterplot of life expectancy vs infant deaths (e) Heat map of life expectancy. The heatmap of all the input features of life expectancy, adult mortality, infant deaths, and GDP and the most significant number of -0.78 indicates a negative correlation and the best correlation between life expectancy and adult mortality.





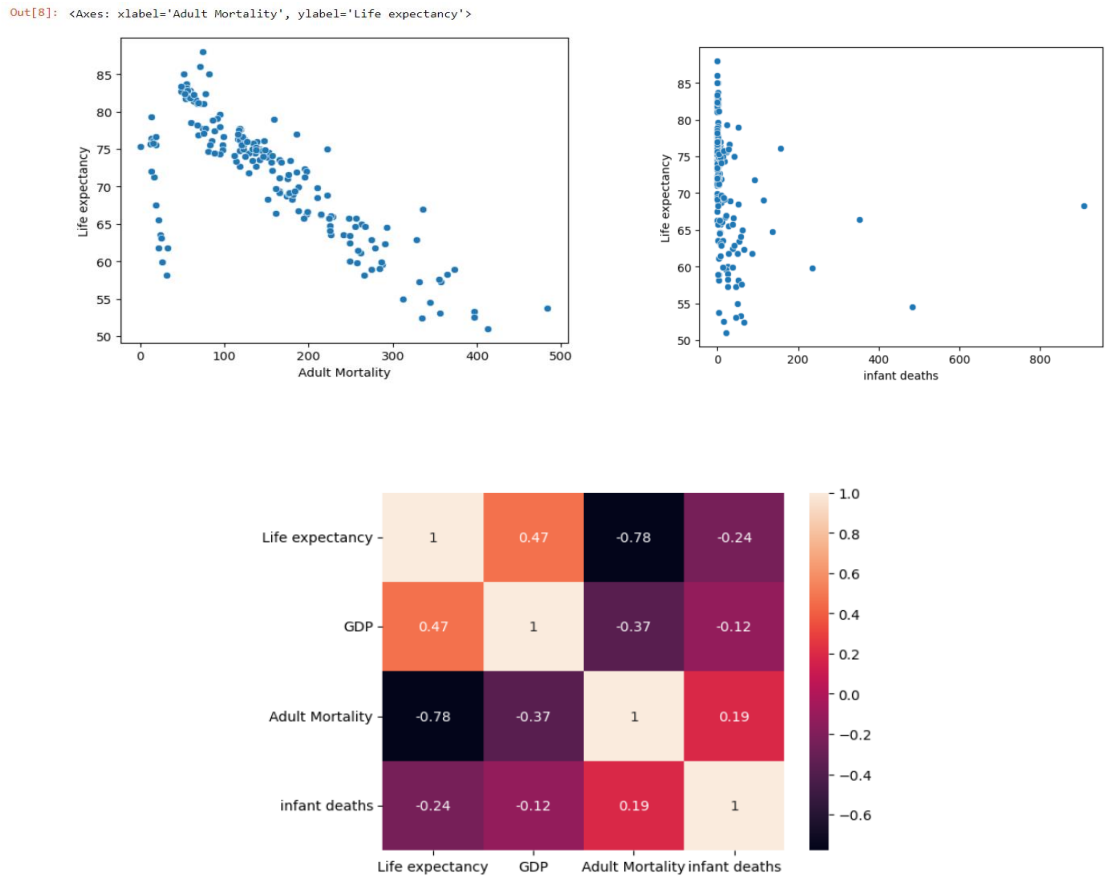


Figure 3(a) Histogram showing mean of life expectancy (b) Scatterplot of life expectancy vs GDP (c) Scatterplot of life expectancy vs adult mortality (d) Scatterplot of life expectancy vs infant deaths (e) Heat map of life expectancy

3.1.1 Simple linear regression model

Table 1. Performance metrics of simple linear regression models on life expectancy vs GDP, Adult mortality & Infant deaths

Performance metrics	Life expectancy vs GDP	Life expectancy vs Adult Mortality	Life expectancy vs Infant deaths
Mean absolute error (mae)	6.19	3.61	6.80
Mean squared error (mse)	58.34	27.83	67.94
Root mean squared error(rmse)	7.64	5.28	8.24
Coefficient of determination (R2)	0.15	0.60	0.01

Table 1 shows that the adult mortality has minimal errors and more significant R2 hence it best fits the simple linear regression model.

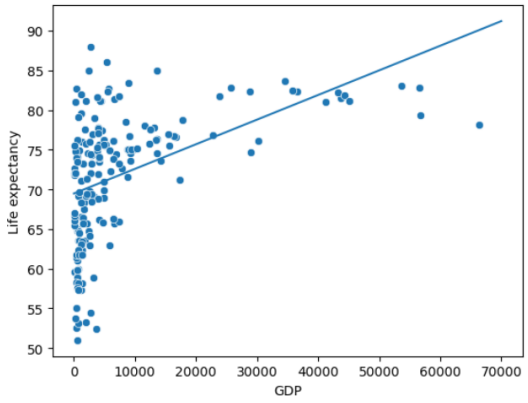


Figure 4a. Life expectancy vs GDP

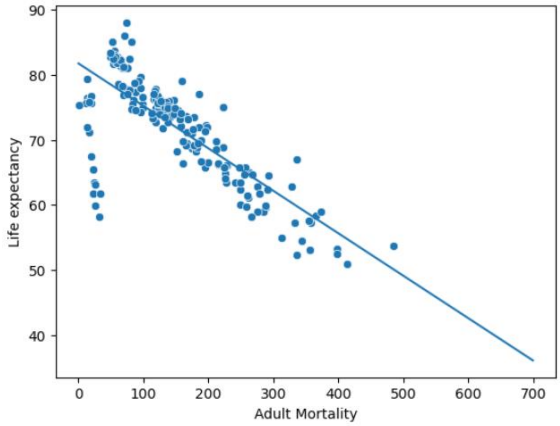


Figure 5b. Life expectancy vs Adult mortality

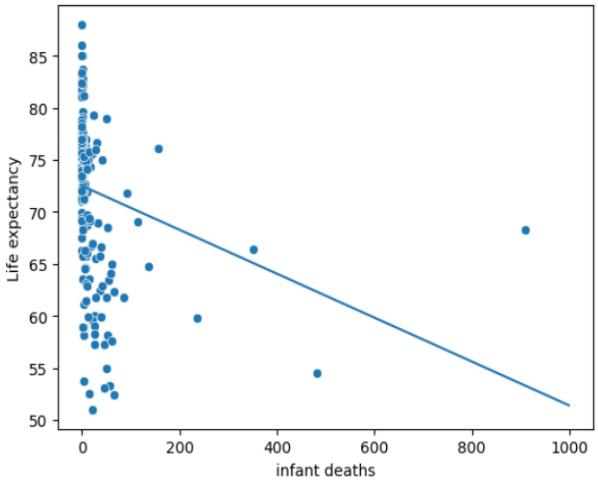


Figure 6c. Life expectancy vs infant deaths

Figure 5 Life expectancy vs adult mortality showed a better fit.

3.1.2 Polynomial regression model

Table 2. Perfomance metrics of a polynomial regression model of Life expectancy vs GDP

Performance metrics	Degree 2	Degree 3	Degree 4	Degree 5
Mean absolute error (mae)	5.51	5.46	5.46	5.30
Mean squared error (mse)	50.90	49.01	47.96	46.50
Root mean squared error(rmse)	7.13	7.00	6.93	6.82
Coefficient of determination (R2)	0.26	0.29	0.30	0.32

Table 2 shows that the highest degree of polynomial gave the least error and the highest R2 hence it is the best fit for the polynomial regression model.

3.1.3. Multiple input regression model

Table 3. Multiple input regression model

Performance Metrics	Multiple input features
Mean absolute error (mae)	3.57
Mean squared error (mse)	26.21
Root mean squared error(rmse)	5.12
Coefficient of determination (R2)	0.62

The multiple input regression model indicates lesser errors and a slightly higher R2 score.

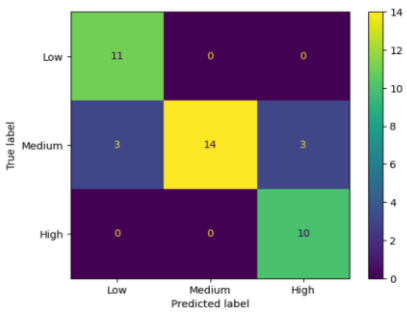


Figure 7. Confusion matrix plot of life expectancy class

3.2.1 K-Mean Cluster Algorithm

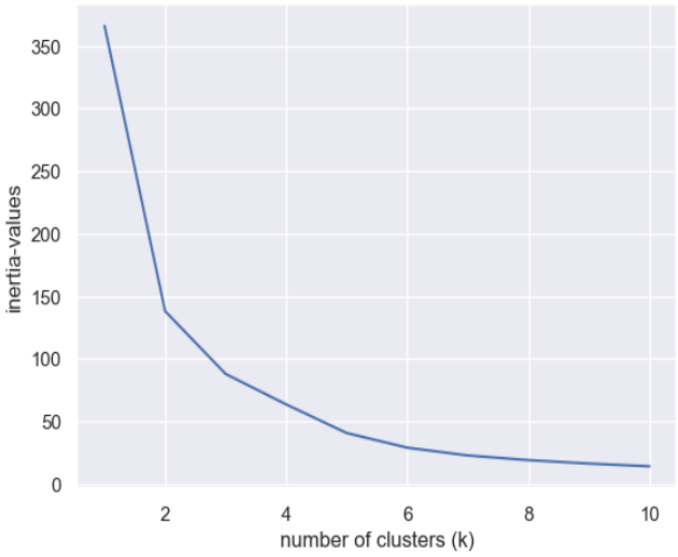


Figure 8a. optimal k value for life expectancy vs GDP

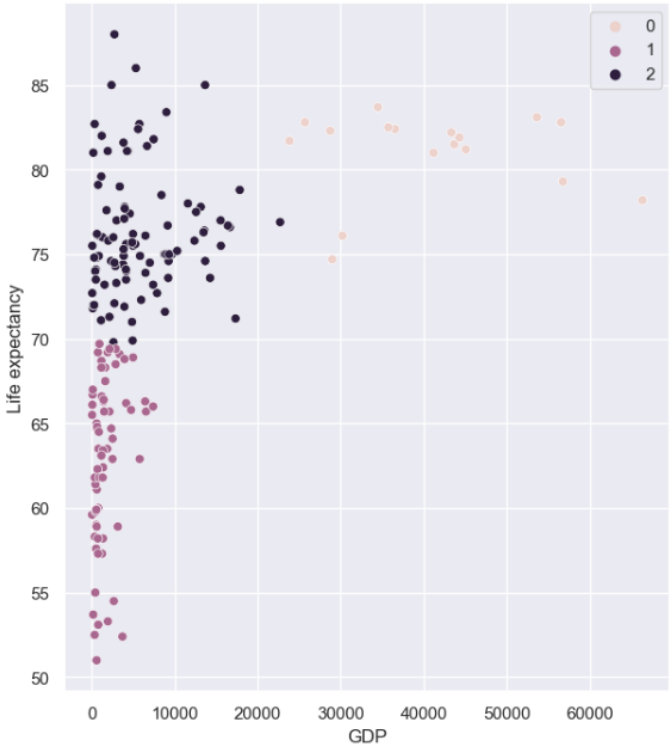


Figure 8b. K-Mean cluster for life expectancy vs GDP

Figure 8a, c indicated that the optimal value of k is 3, where the plot started flattening out, and this is known as the "Elbow method".

Table 4 shows that K-Mean cluster for life expectancy vs GDP produced a better clustering

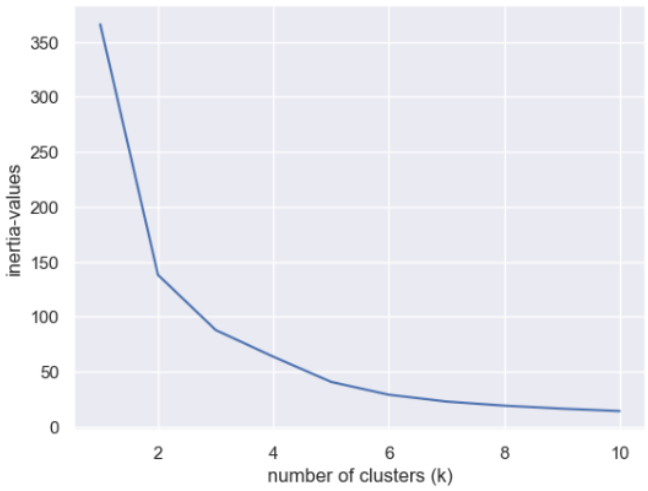


Figure 8c. Optimal k value for life expectancy vs adult mortality

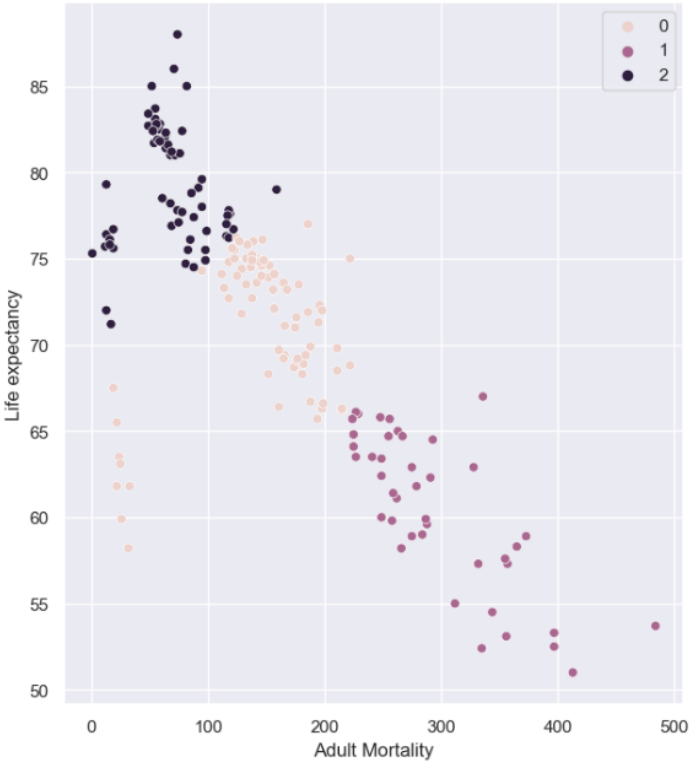


Figure 8d. K-Mean cluster for life expectancy vs adult mortality

Table 4. K-Mean cluster; performance metrics for life expectancy vs GDP & life expectancy vs adult mortality

Performance metrics	Life expectancy vs GDP	Life expectancy vs Adult mortality
David_Bouldin Score	0.58	0.79
Silhouette Score	0.55	0.44
Calinski –Harabasz Score	317.51	284.34

3.2.2 Agglomerative Hierarchical Clustering & DBSCAN models

Figure 9. Agglomerative cluster for life expectancy vs GDP

Figure 10. DBSCAN cluster for life expectancy vs GDP

Table 5. Agglomerative cluster & DBSCAN cluster; performance metrics for life expectancy vs GDP

Performance metrics	Agglomerative clustering for Life expectancy vs GDP	DBSCAN clustering for Life expectancy vs GDP
David_Bouldin Score	0.56	1.15
Silhouette Score	0.44	0.30
Calinski –Harabasz Score	129.50	47.83

Table 5 shows that Agglomerative clustering indicated a better fit than DBSCAN clustering, but DBSCAN showed its robustness in capturing outliers. Comparing Table 4 & Table 5 they indicate that K-Mean produced the best clustering of life expectancy vs GDP because of the compactness of the clustering which buttresses the aim of getting more compact intra distance and further inter distance of cluster data set.

4. CONCLUSION

The life expectancy for a country is dependent on a number of factors like GDP, adult mortality, infant deaths socioeconomic status, the development state of the country. The most correlated factor out of above mentioned is adult mortality, in which an increase in adult mortality results to a decrease in life expectancy. K-Mean clustering model showed the best fit for the life expectancy class.

5. REFERENCES

Richings, D. A., Davis, D. O., Fagbola, D. T. & Rose, S., 2023. University of Hull; Understanding Artificial Intelligence 771763_A23_T1; Week 6: Introduction to Supervised Learning. [Online]
Available at: https://canvas.hull.ac.uk/courses/67474/files/4950856?module_item_id=1014202
[Accessed 15 November 2023].

Richings, D. A., Davis, D. O., Fagbola, D. T. & Rose, S., 2023. *University of Hull; Understanding Artificial Intelligence 771763_A23_T1; Workshop2 - Supervised Learning.* [Online]
Available at: https://canvas.hull.ac.uk/courses/67474/files/4978635?module_item_id=1019803
[Accessed 15 November 2023].

Richings, D. A., Davis, D. O., Fagbola, D. T. & Rose, S., 2023. *University of Hull; Understanding Artificial Intelligence 771763_A23_T1; Week7: Unspervised Learning.* [Online]
Available at: https://canvas.hull.ac.uk/courses/67474/files/4964048?module_item_id=1016821
[Accessed 15 November 2023].

Richings, D. A., Davis, D. O., Fagbola, D. T. & Rose, S., 2023. *University of Hull; Understanding Artificial Intelligence 771763_A23_T1; Workshop 3 - Unspervised Learning.* [Online]
Available at: https://canvas.hull.ac.uk/courses/67474/files/4978677?module_item_id=1019809
[Accessed 15 November 2023].