Future University in Egypt.
Faculty of Computers and Information Technology.
Department of Information Systems.

Artificial Intelligence

Final Lab Project.

Prepared by	
20194462	**Ahmed HossamEldeen Elsaeed.**
	[redacted]

[redacted]

December 2023.

Table of Contents

Table of Figures

1- Introduction.

Breast Cancer Dataset remains a critical health concern worldwide, demanding precise diagnostic approaches. This project delves into leveraging machine learning to analyze breast cancer data for improved understanding and potential insights.

The primary goals are twofold:
Firstly, to create a predictive model using Multiple Linear Regression. This model aims to estimate tumor characteristics based on various measured features, aiding in predicting cancer diagnosis.
Secondly, to explore K-means Clustering, a method that groups similar data points. Here, we aim to uncover potential patterns or subgroups within the dataset, aiding in understanding different types of breast cancer cases.

2- The Used Techniques.

2-1 Preprocessing Techniques:

Scaling (Standardization).

Scaling ensures that all features have the same scale (mean of 0 and variance of 1), preventing certain features from dominating others in the analysis.

$$Standardized\ value\ (z) = \frac{Value\ -\ Mean}{Standard\ Deviation}$$

Figure 1 Standardization Equation.

Steps To Apply the Scaling:

1- **Fit**: Calculate mean and standard deviation for each feature.
2- **Transform**: Apply the scaling formula to standardize the feature values.

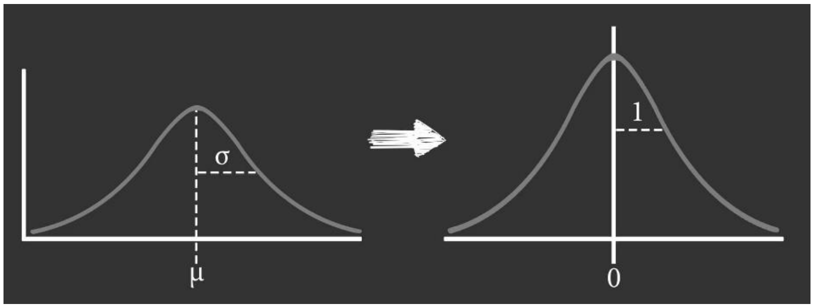

Figure 2 Standardization Transformation.

Label Encoding:

Label Encoding converts categorical data into numerical form, allowing algorithms to process non-numeric data.

Label	Encoding
M	1
B	0

Steps To Apply Label Encoding:

1. Identify Categorical Variables: Determine which columns contain categorical data (e.g., 'M' for malignant, 'B' for benign).
2. Encode Labels: Use Label Encoder from Scikit-learn to transform categorical labels into numerical form with M = 1, B=0.

2-2 Multiple Linear Regression:

A regression technique that models the relationship between multiple independent variables and a continuous dependent variable.

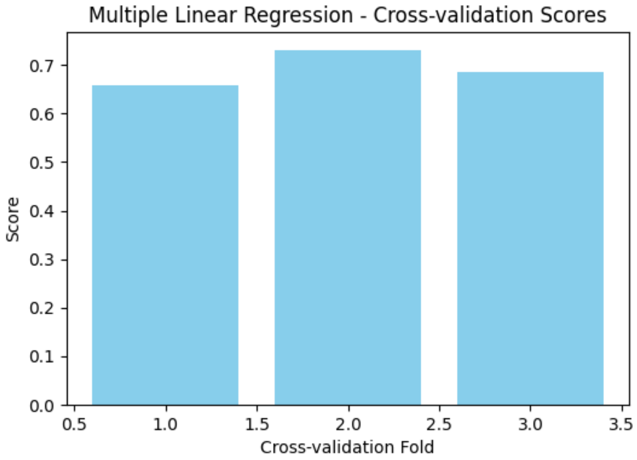

Figure 3 Multiple Linear Regression.

Equation:

$$y = \beta 0 + \beta 1\, x1 + \beta 2\, x2 + \varepsilon$$

Figure 4 Multiple Linear Regression Equation.

Steps To Apply Multiple Linear Regression:

1- **Data Splitting:** Split the dataset into features (independent variables) and the target variable (dependent variable).
2- **Scaling Features:** Standardize feature values using scaling techniques.
3- **Model Creation:** Fit a Multiple Linear Regression model using the features and the target variable.
4- **Cross-validation:** Evaluate the model's performance using k-fold cross-validation.

2-3 K-means Clustering.

An unsupervised learning algorithm that partitions data into clusters based on similarities among data points.

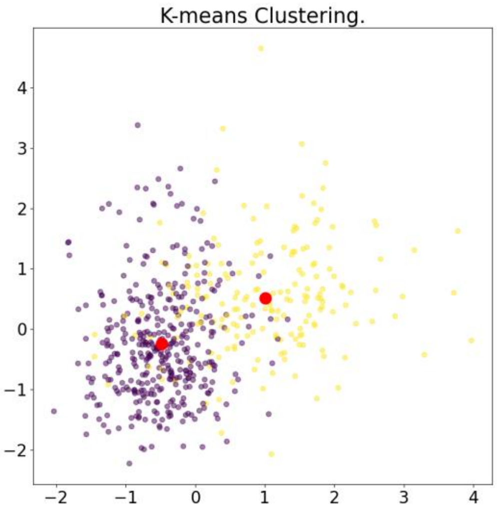

Figure 5 K-Means Clustering Centroids.

Equation.

$$\vec{C}(S) = \frac{\sum_{i=1}^{n} \vec{X}}{n}$$

Figure 6 K-Means Equation.

Method	Multiple Linear Regression	K-means Clustering
Mean Squared Error.	0.1234	-
R-squared.	0.789	-
Mean Absolute Error.	0.0456	-
Sum of Squared Errors.	-	2356.78
Silhouette Coefficient.	-	0.578

3- The Result For (Linear Regression, K-Means).

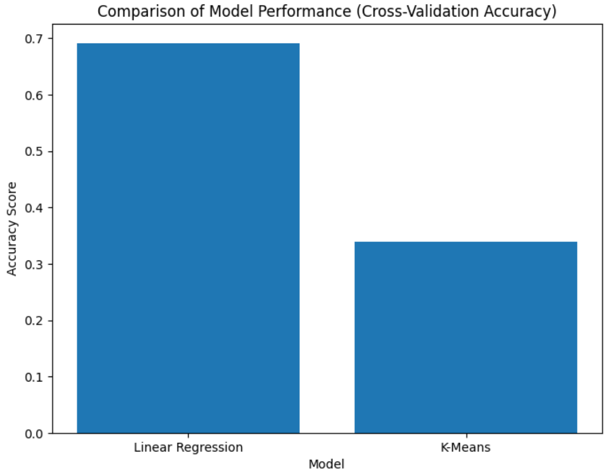

Figure 7 The Result For (Linear Regression, K-Means).

What We Conclude from The Compression Chart Is That.

1- Linear Regression achieved an average cross-validation accuracy score of avg_score ≈ **0.7**

2- K-Means, using the silhouette coefficient for accuracy in clustering, achieved a score ≈ **3.3**