LIFESTYLE CORRELATED CANCER RISK FACTOR PREDICTION SUPPORTED BY UNSUPERVISED MACHINE LEARNING MODEL USING K MEANS CLUSTERING ALGORITHM

Overview:

Apart from the elevated technologies and comforting lifestyle, prevalence of patients with cancer disease still exists. A healthcare dataset is introduced which includes the factors based on lifestyle and genetics that indicates the risk of developing a cancer. Finding the correlation between each feature, a K Means Clustering model is trained without any guidance such that it tends to form a set of clusters that predicts the severity of the cancer. The risk level of cancer is scaled between low, medium, and high. The foreseen outcome could provide a vision to advance in healthcare support that could radically diminish the cancer numbers to healthy values.

Problem Statement:

People are no longer concerned about the possibility of being exposed to cancer owing to the fear of health examination and the cost of the healthcare services consulted. This lowers awareness in the centre of the population, unwittingly increasing illness prevalence. The availability of such freely accessible platforms is critical so that they does not interfere with the tasks at hand. In our everyday routine, a single minute of self-analysis would have a higher influence on increasing life expectancy and preventing cancer. This demonstration utilizes the machine learning model built with the K means clustering algorithm to segment the individuals into lifestyle based cancer risk categories, providing an accessible data driven alternative to timely prevention.

Main Objective(s):

- To categories each person into three levels of cancer risk categories based on the lifestyle and genetics features indicated.
- To train the unsupervised model with K means clustering that outlines the required clusters (low, medium, and high) and predicts accurate category for the given latest data.

DATASET DESCRIPTION:

A. **Source of data:** Kaggle: https://www.kaggle.com/datasets/tarekmasryo/cancer-risk-factors-dataset

B. **Structure of data:**

   a. **Rows:** 2000 records

   b. **Columns:** 21 features

C. **Data columns:**

COLUMN NAME	COLUMN DESCRIPTION
Patient_ID	Unique ID of the patient in the hospital records.
Cancer_Type	Diagnosed cancer type.
Age	Age of the patient.
Gender	Gender of the patient (Male / Female).
Smoking	Smoking score based on their smoking activity. 0 – Min, 10 – Max.
Alcohol_Use	Alcohol intake score based on their periodic consumption. 0 – Min, 10 – Max.
Obesity	Obesity score scaled after testing with height and weight value. 0 – Min, 10 – Max.
Family_History	Affirmation case of any family member encountered cancer issue. 0 – No, 1- Yes.
Diet_Red_Meat	Total meat consumed in a single meal as a score. 0 – Min, 10 – Max.
Diet_Salted_Processed	Overall salt used in a single meal in a score. 0 – Min, 10 – Max.
Fruit_Veg_Intake	Overall fruits and vegetables consumed in a single meal in a range. 0 – Min, 10 – Max.
Physical_Activity	Rating of patient’s physical activity. 0 – Min, 10 – Max.
Air_Pollution	Region and area based air pollution rating. 0 – Min, 10 – Max.
Occupational_Hazards	Hazard exposure severity rating of patient’s job. 0 – Min, 10 – Max.
BRCA_Mutation	Tumor suppressing genes overall score of the patient. 0 – Min, 10 – Max.
H_Pylori_Infection	Helicobacter pylori bacterium infection severity of the patient.

	0 – Min, 10 – Max.
Calcium_Intake	Cumulative score of calcium content in a meal. 0 – Min, 10 – Max.
Overall_Risk_Score	Normalized score correlated with features of Smoking, Alcohol_Use, Obesity, Diet_Salted_Processed, Air_Pollution, Occupational_Hazards.
BMI	Calculated Body Mass Index value from height and weight of the patient.
Risk_Level	Categories derived from the value of Overall_Risk_Score feature. Categories: Low (0 to 0.32), Medium (0.33 to 0.659) , High (0.66 to 1).

D. Key Variables:

a. **X – Independent variables:** Age, Gender, Smoking, Alcohol_Use, Obesity, Family_History, Diet_Red_Meat, Diet_Salted_Processed, Fruit_Veg_Intake, Physical_Activity, Air_Pollution, Occupational_Hazards, BRCA_Mutation, H_Pylori_Infection, Calcium_Intake, BMI

b. **Y – Dependent variables:** Cancer_Type, Overall_Risk_Score, Risk_Level

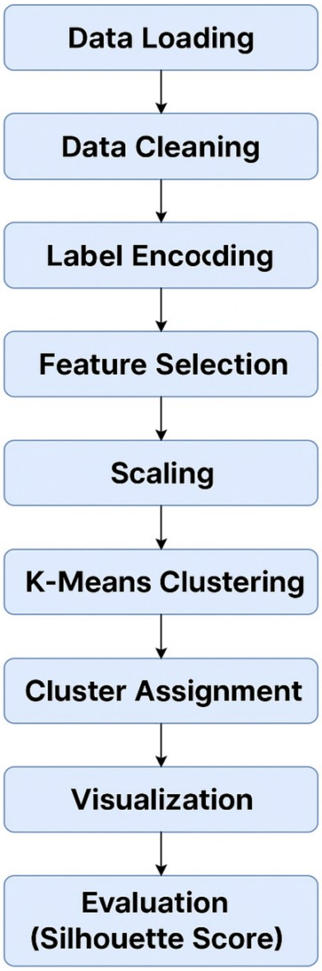
Y DATA PROFILE:

cancer-risk-factors-data-profile-report.htr

Features	count	mean	std	Min	25%	50%	75%	max
Cancer_Type	2000	1.774	1.3458	0.0000	1.0000	2.0000	3.0000	4.0000
Age	2000	63.248	10.4629	25.0000	56.0000	64.0000	70.0000	90.0000
Gender	2000	0.489	0.5000	0.0000	0.0000	0.0000	1.0000	1.0000
Smoking	2000	5.157	3.3253	0.0000	2.0000	5.0000	8.0000	10.0000
Alcohol_Use	2000	5.035	3.2610	0.0000	2.0000	5.0000	8.0000	10.0000
Obesity	2000	5.968	3.0614	0.0000	4.0000	6.0000	9.0000	10.0000
Family_History	2000	0.195	0.3959	0.0000	0.0000	0.0000	0.0000	1.0000
Diet_Red_Meat	2000	5.190	3.1545	0.0000	3.0000	5.0000	8.0000	10.0000
Diet_Salted_Processed	2000	4.564	3.0883	0.0000	2.0000	4.0000	7.0000	10.0000
Fruit_Veg_Intake	2000	4.928	3.0453	0.0000	3.0000	5.0000	8.0000	10.0000

Physical_Activity	2000	4.015	2.9785	0.0000	1.0000	4.0000	6.0000	10.0000
Air_Pollution	2000	5.323	3.2075	0.0000	3.0000	5.0000	8.0000	10.0000
Occupational_Hazards	2000	4.979	3.2129	0.0000	2.0000	5.0000	8.0000	10.0000
BRCA_Mutation	2000	0.033	0.1774	0.0000	0.0000	0.0000	0.0000	1.0000
H_Pylori_Infection	2000	0.197	0.3975	0.0000	0.0000	0.0000	0.0000	1.0000
Calcium_Intake	2000	3.941	3.0489	0.0000	1.0000	4.0000	6.0000	10.0000
Overall_Risk_Score	2000	0.454	0.1231	0.0293	0.3670	0.4554	0.5398	0.8522
BMI	2000	26.183	3.9475	15.0000	23.5000	26.2000	28.7000	41.4000
Physical_Activity_Level	2000	4.939	3.1660	0.0000	2.0000	5.0000	8.0000	10.0000
Risk_Level	2000	1.736	0.5445	0.0000	2.0000	2.0000	2.0000	2.0000
Clusters	2000	0.987	0.7891	0.0000	0.0000	1.0000	2.0000	2.0000

METHODOLOGY:



TRAINING – TESTING SPLIT:

Since the model is an unsupervised machine learning K means clustering algorithm, the dataset X independent variables are utilized thoroughly for training.

MODEL SELECTION:

Model: K MEANS CLUSTERING (Unsupervised Model)

Description:

- Without any guidance or supervision, the model learns the data and recognizes its own pattern to predict the proposed new data
- It is efficient for grouping data into selected cluster

MODEL PARAMETERS:

Parameters Reference:

PARAMETERS	DESCRIPTION
n_clusters	Number of groups created for segmentation
random_state	Model learns data through random combinations ensuring reproducibility.

Parameters values used:

- n_clusters : "3" (Due to three levels of risk factors)
- random_state: "42". Used to shuffle the dataset for learning various combinations

MODEL TRAINING:

- Total Dataset used = 2000 records
- Scaler used: StandardScaler. Used to normalize multiple features
- 3 clusters are created by mentioning n_clusters = 3
- random state 42 is implemented to help model learn different possibilities

MODEL EVALUATION:

Sample data values obtained:

X_SCALED PARAMETERS						CLUSTERS
AGE	SMOKING	ALCOHOL_USE	DIET_RED_MEAT	DIET_SALTED_PROCESSED	OVERALL_RISK_SCORE	
0.4543	0.5544	-0.9309	-0.0601	-0.5064	-0.4531	1.0000
1.0279	0.8552	1.2162	-1.6456	-0.5064	-0.2450	2.0000
-0.7885	0.5544	1.5229	-0.6943	-0.5064	1.2242	2.0000
-0.2149	0.2536	-0.9309	0.2570	-0.8303	-1.1053	1.0000
0.3587	1.4568	0.6027	0.2570	-0.5064	0.5682	2.0000
1.3147	1.4568	0.9095	0.2570	-1.4780	0.3594	2.0000
-0.4061	1.4568	1.5229	1.2083	-0.1825	1.6897	0.0000
1.0279	0.8552	0.2960	-0.6943	-0.5064	0.2025	2.0000
0.7411	1.1560	-1.5444	1.5254	-0.1825	0.3509	0.0000
-0.7885	0.5544	-1.2377	-1.6456	-0.1825	-0.4032	2.0000

EVALUATION METRICS:

Metrics used:

- Silhouette Score: Utilized while developing a machine learning algorithm (specifically for K Means Clustering), to measure the performance of clusters formed.
- It forms the clusters by using the formula:

$$s = \frac{b - a}{\max(a, b)}$$

Where:

$a = average\ distance\ from\ the\ concern\ point\ to\ all\ other\ points\ in\ the\ same\ cluster$

$b\ = average\ distance\ from\ the\ concern\ point\ to\ all\ points\ with\ nearest\ neighboring\ cluster$

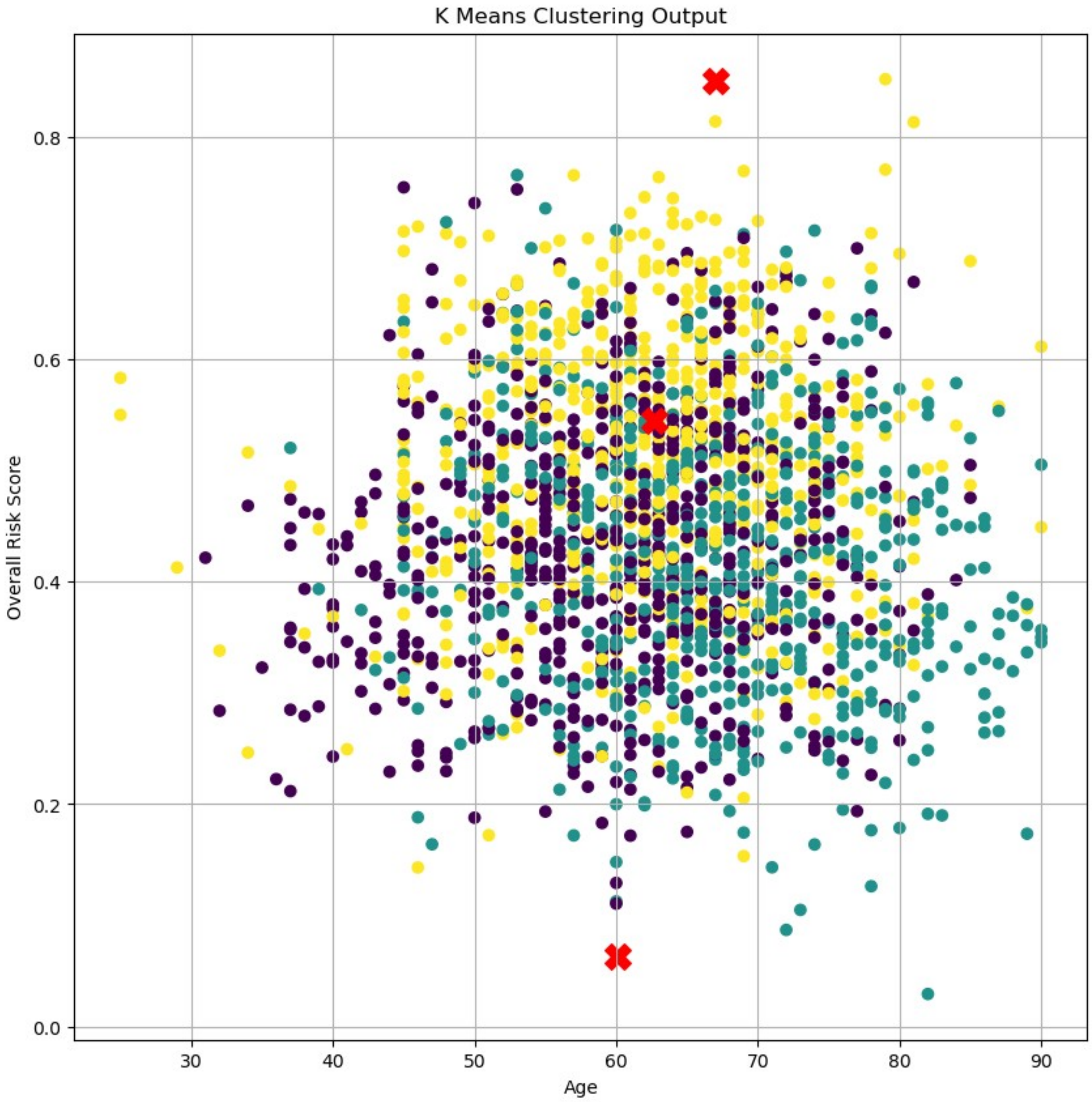
- It should be observed that the silhouette score should be ranged between 0 and 1. Scores between 0 to 0.20 are considered to be wrong clustering, between 0.21 to 0.60 are weak clustering, and between 0.61 to 1 ensures the model predicts the correct cluster.

Evaluated Value from the model:

- Silhouette score = 0.165

- This states that the model predicts wrong clustering. Hence, the model has to be boosted to increase the silhouette score.

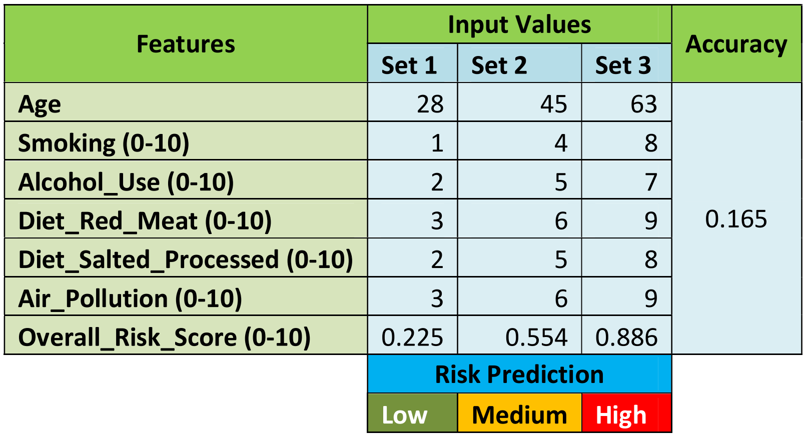
PLOTS:



PLOT INTERPRETATION:

- From the above visible scattering plot, comparison of feature between age and overall risk score is being analyzed. Here, clusters overlap with each other and doesn’t produce meaningful outcome.
- This clears that the model has to be trained with precise values of features and increase the records of the data.
- The model represents weak clusters. Hence, model couldn’t predict required outcome.

NEW DATA PREDICTION:



Features	Input Values			Accuracy
	Set 1	Set 2	Set 3	
Age	28	45	63	0.165
Smoking (0-10)	1	4	8	
Alcohol_Use (0-10)	2	5	7	
Diet_Red_Meat (0-10)	3	6	9	
Diet_Salted_Processed (0-10)	2	5	8	
Air_Pollution (0-10)	3	6	9	
Overall_Risk_Score (0-10)	0.225	0.554	0.886	
	Risk Prediction			
	Low	Medium	High	

CONCLUSION:

The K means clustering model has successfully predicted the unknown data to the approximate risk levels. However, the model exhibits low accuracy value (0.165), the risk levels are weakly clustered. There are no clear defined groups formed by the specified features. Hence, the model has to be enhanced using optimization as well as boosting techniques such that the model could strongly cluster to required factors.