LUNG CANCER PREDICTION

PROBLEM STATEMENT WITH DOMAIN

1)LUNG CANCER PREDICTION.

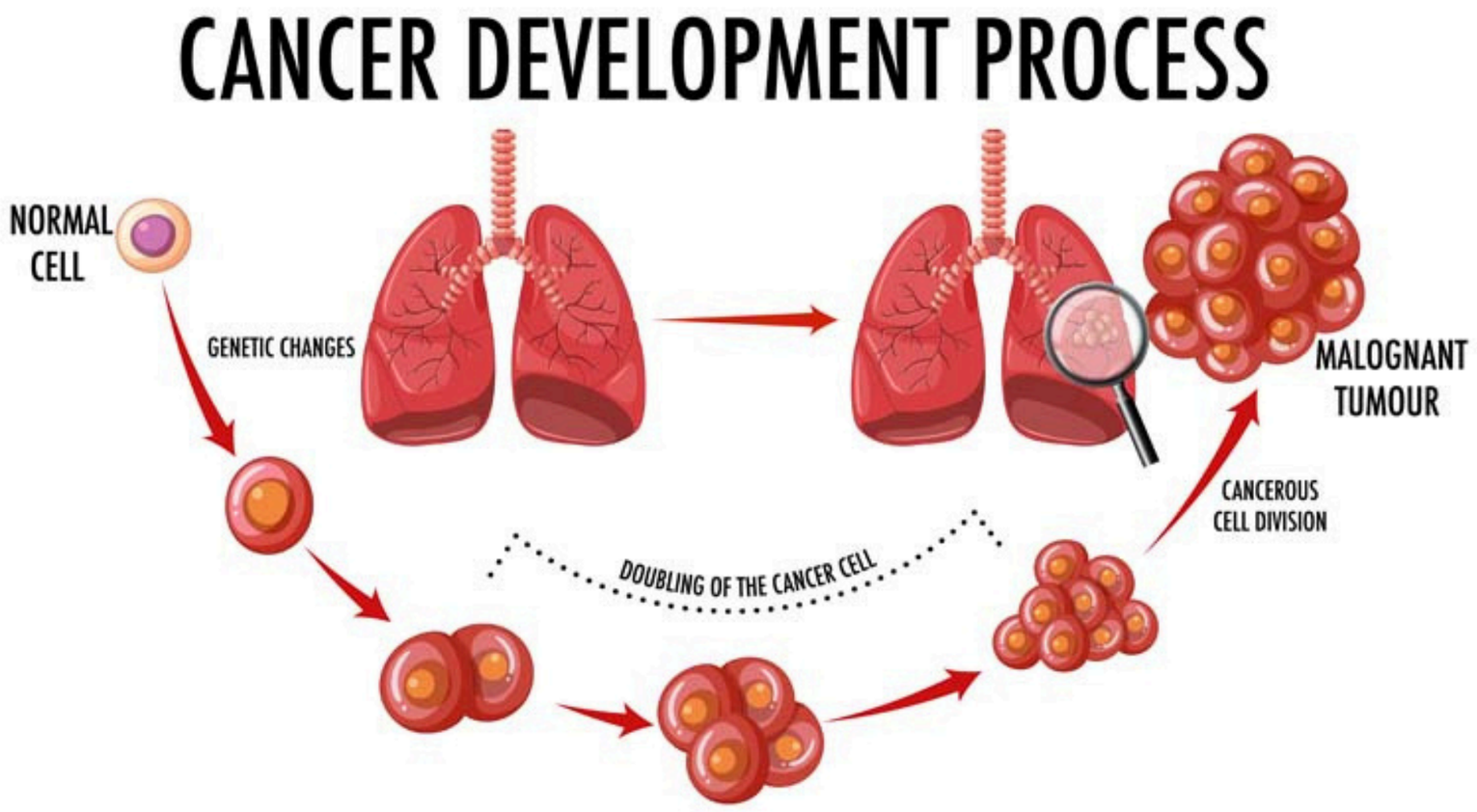
2)BIO MEDICAL DOMAIN.

3)ITS BASED ON SUPERVISED LEARNING - MACHINE LEARNING (NUMBERS).

4)IT IS CLASSIFICATION (CATEGORICAL VALUE OF OUTPUT).

SYMPTOMS

Common symptoms:



- Chest Pain
- Shortness of Breath
- Wheezing
- Coughing up Blood
- Chronic Fatigue

QUESTION :

DOES MACHINE LEARNING HAVE THE POTENTIAL TO CREATE NON INVASIVE INEXPENSIVE SCREENING TOOL TO PREDICT THOSE WHO WOULD NEED TO GO LDCT TESTING?

MACHINE LEARNING CAN CREATE NON-INVASIVE, AFFORDABLE SCREENING TOOLS BY ANALYZING VARIOUS PATIENT DATA TO IDENTIFY THOSE WHO NEED LDCT TESTING. IT IMPROVES RISK PREDICTION BY INTEGRATING DEMOGRAPHICS, MEDICAL HISTORY, AND LIFESTYLE FACTORS. THIS APPROACH HELPS FOCUS RESOURCES ON HIGH-RISK INDIVIDUALS, REDUCING UNNECESSARY TESTS AND ASSOCIATED COSTS. OVERALL, IT ENHANCES SCREENING ACCURACY AND EFFICIENCY.

RETAINED VARIABLE WITH JUSTIFICATION

RETAINING THESE VARIABLES ENHANCES THE PREDICTION MODEL BY INCORPORATING A RANGE OF RISK FACTORS AND SYMPTOMS. KEY ELEMENTS LIKE AGE, SMOKING_STATUS, AND COPD_DIAGNOSES DIRECTLY INFLUENCE LUNG CANCER RISK. ADDITIONAL FACTORS LIKE GENOMIC_SEX, DAILY_CIGARETTES, AND COUGHING PROVIDE CONTEXT AND REFINE RISK ASSESSMENT. COMBINING THESE VARIABLES IMPROVES ACCURACY IN IDENTIFYING INDIVIDUALS WHO MAY NEED FURTHER TESTING.

REMOVED VARIABLE WITH JUSTIFICATION

PATIENT_ID IS AN IDENTIFIER WITH NO PREDICTIVE VALUE FOR LUNG CANCER. ADOPTED_STATUS, NUMBER_OF_PREGNANCIES, AND MONTH_OF_BIRTH LACK DIRECT RELEVANCE TO LUNG CANCER RISK. YEAR_OF_BIRTH PROVIDES NO ADDITIONAL INSIGHT BEYOND AGE. CONSTANT_EXHAUSTION IS LESS SPECIFIC AND RELEVANT COMPARED TO OTHER SYMPTOMS AND RISK FACTORS IN PREDICTING LUNG CANCER.

AS A DATA SCIENTIST YOU ARE A LOGICIAN ,MATHEMATISIAN,TECHNICIAN,ANALYST NAND U NEED EPIDEMIOLOGISTS TO UNDERSTAND UR ANALYSES .EPIDEMIOLOGIST USUALLY BUSY INDIVIDUALS AND THEY DONT HAVE ALL THE TIME IN THE WORLD.ONE ESSENTIAL SKILL THAT U MUST ADHERE TO IS TO BE CONCISE AND STRAIGHT TO THE POINT . FOCUS ON THE ANSWERS ,NEEDED FOR EACH TASK AND PROVIDE JUST ENOUGH WORDS FOR THE ANSWERS ONLY .THERE IS NO NEED PROVIDE LENGTHY DESCRIPTIONS OF ALGORITHMS AND METHODS UNLESS U R ASKED TO DO

- RETAINED VARIABLES: AGE, SMOKING STATUS, COPD—CRITICAL FOR PREDICTING LUNG CANCER RISK.
- REMOVED VARIABLES: PATIENT ID, ADOPTED STATUS, PREGNANCIES—IRRELEVANT TO LUNG CANCER RISK.

Justification for retention or dropping the variable

S.No	Variable	Retain or Drop	Brief justification for retention or dropping
1	PATIENT-ID	Drop	Dropped due to data inaccuracies affecting the lung cancer prediction model's reliability.
2	GENOMIC SEX	Retain	Retained in the lung cancer prediction model as it significantly influences cancer risk and helps tailor treatment strategies.
3	AGE	Retain	It is a key factor that strongly correlates with cancer risk and progression.
4	BLOOD_TYPE	Retain	It may provide valuable insights into individual risk factors and disease progression.
5	NUMBER_OF_SIBLINGS	Retain	It may help identify genetic and familial risk factors associated with cancer.
6	YEAR_OF_BIRTH	Drop	Dropped in favor of using the available age, as age provides a more direct and accurate measure for the lung cancer prediction model.
7	ADOPTED_STATUS	Drop	Dropped from the lung cancer prediction model as it does not contribute significantly to the prediction accuracy compared to other clinical factors.
8	NUMBER_OF_PREGNANCIES	Drop	It does not have a significant impact on predicting lung cancer risk compared to other relevant factors.
9	MONTH_OF_BIRTH	Drop	Dropped - as age provides sufficient predictive value for the lung cancer model.
10	PARENT_ALIVE	Retain	It may provide insights into familial health patterns and genetic predispositions to cancer.
11	SMOKING_STATUS	Retain	smoking is a major risk factor that significantly influences cancer risk and progression.
12	DAILY_CIGARETTES	Retain	It quantifies smoking intensity, a critical factor in assessing lung cancer risk.
13	YELLOW_SKIN	Retain	It may indicate jaundice, which can be associated with advanced disease or liver involvement.
14	ANXIETY	Retain	It may reflect overall health status and could impact patient outcomes and disease progression.
15	PEER_PRESSURE	Retain	It can influence lifestyle behaviors, such as smoking, which are relevant to cancer risk.
16	COPD_DIAGNOSES	Retain	Chronic obstructive pulmonary disease is a significant risk factor that can indicate increased susceptibility to lung cancer.
17	Constant_Exhaustion	Drop	Its minimal impact on prediction accuracy compared to other clinical factors.
18	FATIGUE	Retain	It can be a symptom of underlying cancer and may provide insight into disease progression and patient health.
19	ALLERGY	Retain	It may offer insights into immune system status and potential interactions with cancer risk factors.
20	WHEEZING	Retain	It can be a symptom of respiratory issues and may indicate underlying lung pathology relevant to cancer risk.
21	ALCOHOL_CONSUMPTION	Retain	Alcohol use is a known risk factor that can influence cancer development and progression.
22	COUGHING	Retain	It is a common symptom of lung issues and may indicate the presence or progression of lung cancer.
23	SHORTNESS_OF_BREATH	Retain	It is a critical symptom that can signal lung abnormalities and potential cancer progression.
24	SWALLOWING_DIFFICULTY	Retain	It can indicate advanced disease or metastasis affecting the esophagus or surrounding structures.
25	CHEST_PAIN	Retain	It can be a significant symptom of lung cancer or related conditions, aiding in the assessment of disease presence and severity.
26	LUNG_CANCER	Retain	It is the primary condition being predicted, directly informing the accuracy and relevance of the model's risk assessments.

BEST MODEL WITH JUSTIFICATION THROUGH METRICS

Metrics	USE or DONOT USE	Justification in relation to the success criteria	Model Name	Test Score
ACCURACY	USE SVM to get best model	**SVM (0.82)** is the best choice due to its high accuracy and strong performance in predicting lung cancer. **KNN (0.79)** is also effective but may struggle with large datasets. **Naive Bayes (0.56)** has low accuracy and is less suitable. **ANN (0.20)** performs poorly and is currently not a viable option. Prioritize SVM for meeting your success criteria effectively.	ANN	0.20
			SVM	0.82
			KNN	0.79
			NB	0.56
RECALL	USE SVM to get best model	**KNN (0.94)** is the best choice with the highest recall, effectively identifying most positive cases. **SVM (0.92)** also performs well in detecting positives. **Naive Bayes (0.58)** has lower recall, missing many positive cases. **ANN (0.00)** fails to identify any positives and is currently unsuitable. Prioritize KNN for the highest recall and best performance in detecting lung cancer.	ANN	0
			SVM	0.92
			KNN	0.94
			NB	0.58
PRECISION	USE SVM to get best model	**Naive Bayes (0.88)** has the highest precision, effectively minimizing false positives, but has lower recall. **SVM (0.86)** also offers high precision and balances well with good recall. **KNN (0.82)** has slightly lower precision but compensates with high recall. **ANN (0.00)** is unsuitable due to its extremely low precision. Prioritize Naive Bayes or SVM for the best precision in predicting lung cancer.	ANN	0
			SVM	0.86
			KNN	0.82
			NB	0.88
F1 SCORE	USE SVM to get best model	**SVM (0.89)** has the highest F1 Score, showing the best balance between precision and recall. **KNN (0.88)** also performs well with a high F1 Score. **Naive Bayes (0.70)** has a lower F1 Score, indicating less effective performance. **ANN (0.00)** is unsuitable due to its extremely low F1 Score. Prioritize SVM or KNN for the best overall balance in lung cancer prediction.	ANN	0
			SVM	0.89
			KNN	0.88
			NB	0.70

MODELLING:CREATE PREDICTIVE CLASSIFICATION MODELS

Algorithm Name	Algorithm Type	Learnable Parameters	Some Possible Hyperparameters	Imported Python Package to use the algorithm
ANN	Machine Learning	19 Parameters	optimizer=adam, loss=binary_crossentropy, metrics=accuracy	From tensorflow.keras.models import Sequential, from tensorflow.keras.layers import Dense
SVM	Machine Learning	19 Parameters	kernel-linear, C-0.1, 1, 10, 100	from sklearn.svm import SVC
KNN	Machine Learning	19 Parameters		from sklearn.neighbors import KNeighborsClassifier
NB	Machine Learning	19 Parameters		from sklearn.naive_bayes import GaussianNB

VARIABLE- ISSUE WITH MITIGATION

Variable Name	Issue Description	Proposed Mitigation	Justification for used Mitigation
Fatigue	It shows not in dataframe during null values handling	To remove unwanted spaces	It having some space in letter before . so, I used stripping for removing whitespace from the column names.

OUTPUT

Lung Cancer Prediction

Please enter repective fields

GENOMIC SEX*

AGE*

BLOOD TYPE*

NUMBER OF SIBLINGS*

PARENT ALIVE*

SMOKING STATUS*

DAILY CIGARETTES*

YELLOW SKIN*

ANXIETY*

PEER PRESSURE*

COPD DIAGNOSES*

FATIGUE*

ALLERGY*

WHEEZING*

ALCOHOL CONSUMPTION*

COUGHING*

SHORTNESS OF BREATH*

SWALLOWING DIFFICULTY*

CHEST PAIN*

upload

OUTCOME POSITIVE

Lung Cancer Prediction

For the values

GENOMIC_SEX:1

AGE:55

BLOOD_TYPE:3

NUMBER_OF_SIBLINGS:1

PARENT_ALIVE:1

SMOKING_STATUS:2

DAILY_CIGARETTES:26

YELLOW_SKIN:1

ANXIETY:1

PEER_PRESSURE:2

COPD_DIAGNOSES:2

FATIGUE:2

ALLERGY:1

WHEEZING:1

ALCOHOL_CONSUMPTION:2

COUGHING:2

SHORTNESS_OF_BREATH:2

SWALLOWING_DIFFICULTY:1

CHEST_PAIN:2

Output Lung Cancer prediction:

You **will have** Lung Cancer in the future: **Positive**

LC Predicition

OUTCOME NEGATIVE

Lung Cancer Prediction

For the values

GENOMIC_SEX:1

AGE:55

BLOOD_TYPE:3

NUMBER_OF_SIBLINGS:1

PARENT_ALIVE:1

SMOKING_STATUS:2

DAILY_CIGARETTES:10

YELLOW_SKIN:1

ANXIETY:1

PEER_PRESSURE:1

COPD_DIAGNOSES:1

FATIGUE:2

ALLERGY:1

WHEEZING:2

ALCOHOL_CONSUMPTION:2

COUGHING:1

SHORTNESS_OF_BREATH:1

SWALLOWING_DIFFICULTY:1

CHEST_PAIN:1

Output Lung Cancer prediction:

You **will not** have Lung Cancer in the future:**Negative**

LC Predicition

Advantages

Early Detection: An accuracy of 82% helps in identifying potential lung cancer cases early, improving treatment outcomes and survival rates.

Support for Professionals: It assists doctors by providing a reliable tool to highlight high-risk patients, complementing their clinical expertise.

Cost Reduction: Accurate predictions can minimize unnecessary tests, potentially lowering overall healthcare costs.

WITH REGARDS
VERGEENA DEVI.C