TITLE : CANCER PATIENTS DATASET USING MACHINE LEARNING ALGORITHMS

ABSTRACT : Artificial intelligence (AI) is definitely reshaping the landscape and horizons of oncology, opening new important chance for improving the management of cancer patients[1]. Cancer is considered as one of the deadly diseases in the world. Lung, Prostate, and Breast Cancer are some of the Cancer types that are donate most to the Mortality Rateour proposed research work includes Data Collection which is further analyzed modelled using Machine Learning Techniques. AI-based devices that have already obtained the official approval by the federal drug administration (FDA), here we show that cancer diagnostics is the oncology-related area in which AI is already entered with the largest impact into clinical practice[4]. Furthermore, breast, lung and prostate cancers represent the specific cancer types that now are experiencing more advantages from AI-based devices. The future perspectives of AI in oncology are discussed: the creation of multidisciplinary platforms, the grasp of the importance of all neoplasms, including rare tumours and the continuous support for undertake its growth represent in this time the most important challenges for finalising the AI uprising in oncology[5].

1.INTRODUCTION : Now a days growing exponential number of patients around the globe has cancer patients as the maximum in number. Cancer has come out to be a major threat to human life. Artificial intelligence is exactly reshaping our lives and it is time to understand its evolution and reaching to model future development strategies. This is true also for oncology and related fields, where AI is now opening new important opportunities for improving the management of cancer patients[11].

AI represents an emerging and quickly evolving model that regards different scientific fields, also those allocate to the management of cancer patients[13].The applications of AI are scale upand include new approaches for cancer detection, screening, diagnosis and classification, the characterisation of cancer genomics, the analysis of tumour microenvironment, the evaluation of biomarkers with forcasting and predictive purposes and of strategies for followup and drug discovery. this study has used many classification algorithms. Hence in this paper has tried to detect early cancer in humans and help them to reduce the serious impact on human life,this concept will also help to save lives and time too.

About the data :

The cancer patient dataset has 1k entries. Each entry
contains the following information about an individual:

The age individual integer number greater than 0 , gender , airpolltution,swallowing difficulty,dry cough, snoring, wheezing,shortness of breath,frequent cold,weight loss,balanced diet,chronic lung disease,occupational hazard,dust allergeyand alcohol use all this field present in integer numeric format.level is present in high,low and medium.

Preprocessing : The collected data may conatin missing values that may lead to in consistency.To gain better result data need to be preprocessing and so it will better the effectiveness of the algorithms.

Data collection : Collect data on cancer patient, including their quality characteristics such as age,gender, weight loss,air pollution ,balanced diet and level etc .

Feature selection : Identify the most important features that can help predict the cancer patients. This can be done using statistical analysis or explainable AI using machine learning techniques.

Model training : Select an appropriate XAI using machine learning algorithm like Logistic Regression and train the model on the data. The model is trained to predict the cancer pateint of samples based on their characteristics.

Train model on training data set : Now we should train the model on the training dataset and make for soothsaying the test dataset.We can divide our train dataset into two tract train and testimony.We can train the model on this part and using that make predict for the testimony part.In this way we can validate our soothsaying as we've the true soothsaying for the true soothsayings for the testimony part(which we don't have for the test dataset).

Correlation attributes : Grounded on the correlation among attributes it was observed more likely to their cancer patient. The attribute that are individual and significant can includeage,gender,air pollution,wheezing,dry cough,alcohol used,genetic risk which is since by insight it's considered as important. The correlation among attributes can associated using box in AI platform.

Model evaluation : Evaluate the performance of the model on separate test dataset to determine its accuracy.

2.Data exploration and analysis : In this data set we done some exploratory data analysis .we have plotted some boxplot ,countplot,regression chart .

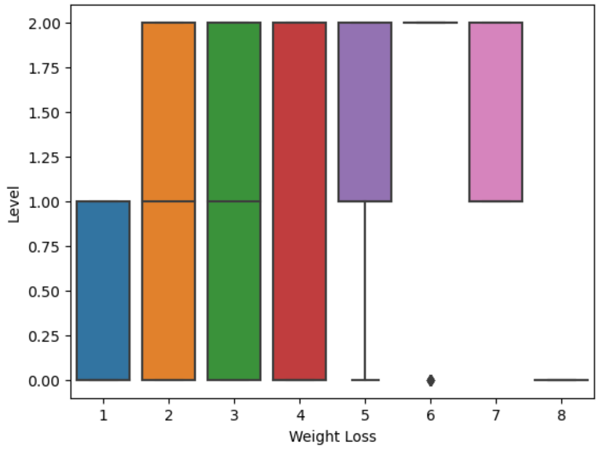


Fig 2.1 Boxplot Weight Loss Vs Level

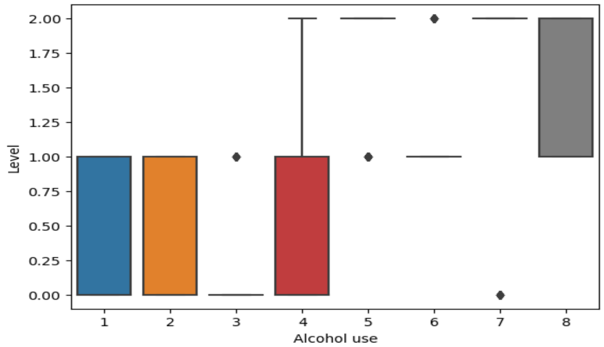


Fig 2.2 Boxplot Alcohol use Vs Level

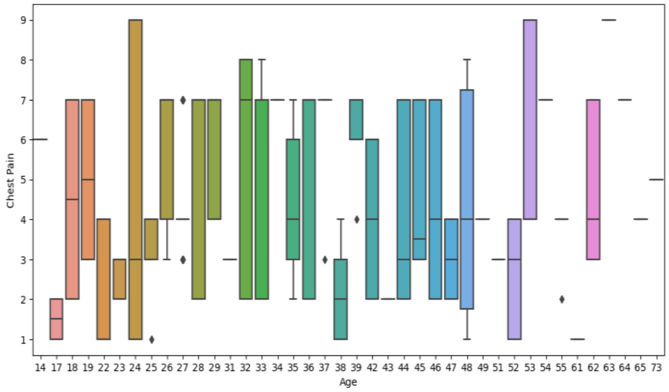


Fig 2.3 Boxplot Age Vs Chest Pain

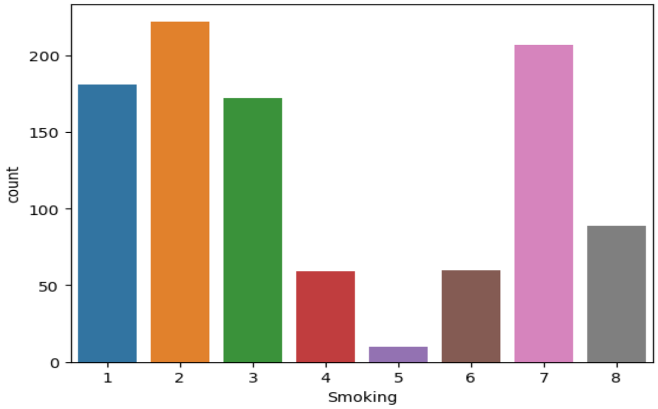


Fig 2.4 Countplot Smoking Vs Count

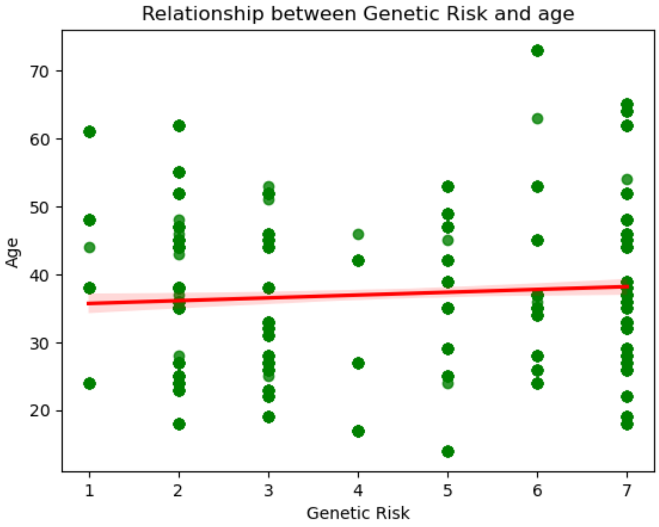


Fig 2.5 RegressionChart Genetic Risk Vs Age

3 . Experiments : We used cancer patient dataset in this we have categorical variable for ex- high , medium and low then we then we cannot directly provide it to machine it won’t be able to understand so explainable AI using ML algorithm involves LIME technique.in this lots of mathematical calculation so in this case we will use map function because we have features in target variable cancer patient level is present in high,low and medium. In our given dataset there is null values so we need to drop that missing values .After this stage, complete the data cleaning and manipulation process on the dataset.

3.1 Evaluation Metrics : Accuracy score is not usually the best form of evaluation in machine learning. Evaluation metrics where we build the model and apply that model to our given dataset that means how well a model is able to make prediction or classifications based on a set of input data.

3.2 procedure

3.2.1 any further processing?

Further data processing such as to use standardscaler is to improve the performance of the XAI using machine learning models. Many machine learning algorithms such as regression are sensitive to the scale of input feature. when some features have a larger scale than other so it can lead to biased results and reduced performance. By StanderScaler, we ensure that all feature have similar scale and thus avoid this issue. StandardScaler is typically used as a preprocessing step before training a machine learning model[1]. It involves fitting the StandardScaler on the training data and then transforming both the training and test data using the same scaler[2].

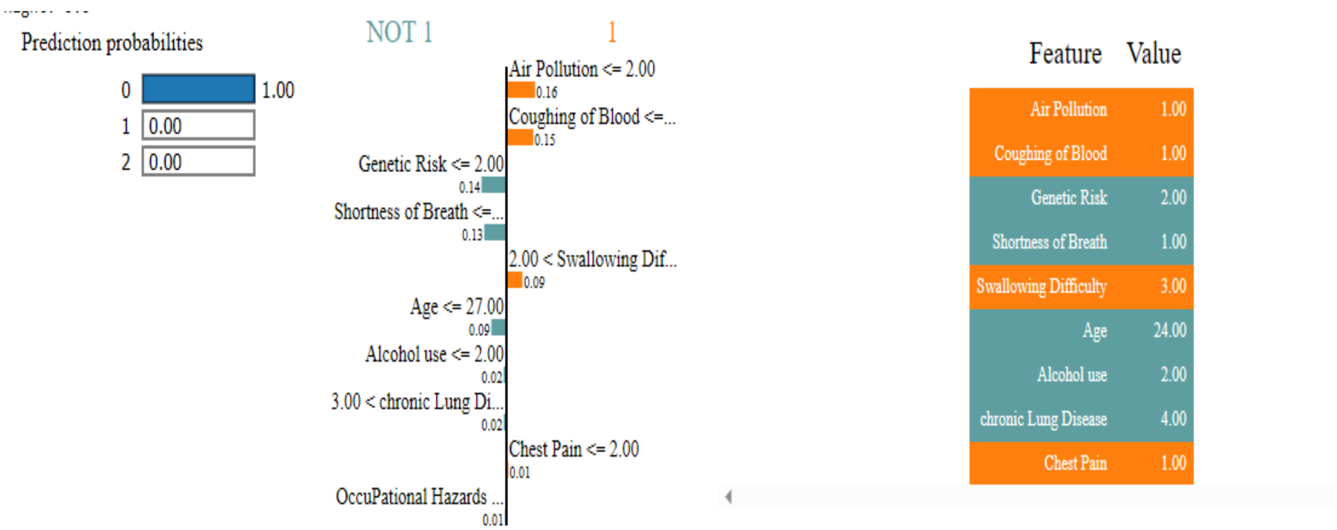
This ensures that the scaling is consistent across the training and test data. The main use of StandardScaler is to standardize the features of a dataset to improve the performance of machine learning models, especially those that are sensitive to the scale of the input features[3].

3.2.2 Models used:

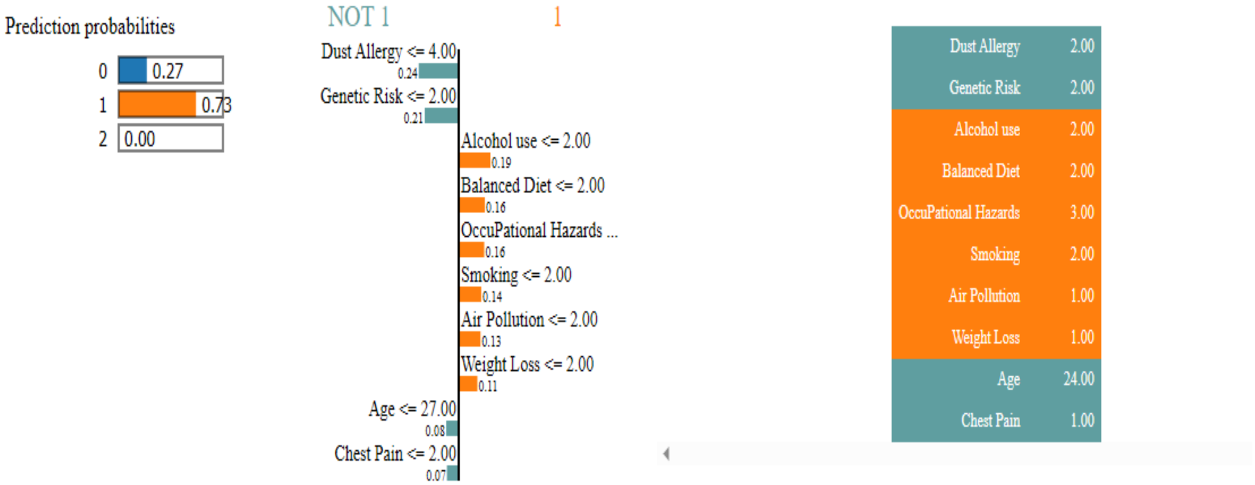
There are several models used in machine learning classification, but for this use case we have label data i.e. target column contains high,medium and low. Some of the popular one is Logistic Regression in this used explainable AI using LIME technique and check model gives best suitable accuracy and model we can deploy further.

4.Result : Before we applied LIME techniques I have seen the dataset this dataset is in the form of classes that why we apply the logistic regression algorithm in logistic regression we apply the technique called LIME, LIME was used for the local interpretability. It focuses on training local surrogate models to explain individual predictions and so the decider can understand why the model predicted a certain class for a particular instance.

LIME Local Explanation



Feature	Value
Air Pollution	1.00
Coughing of Blood	1.00
Genetic Risk	2.00
Shortness of Breath	1.00
Swallowing Difficulty	3.00
Age	24.00
Alcohol use	2.00
chronic Lung Disease	4.00
Chest Pain	1.00



Dust Allergy	2.00
Genetic Risk	2.00
Alcohol use	2.00
Balanced Diet	2.00
OccuPational Hazards	3.00
Smoking	2.00
Air Pollution	1.00
Weight Loss	1.00
Age	24.00
Chest Pain	1.00

5. DISCUSSION :

5.1 Justify four algorithms : Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to “self-learn” from training data and improve over time, without being explicitly programmed[1]. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. ML algorithms are mainly divided into three categories firs one in supervised machine learning algorithm and second one is unsupervised and third one is Reinforcement machine learning algorithm[5]. supervised machine learning algorithms are a type of machine learning algorithm that learns to make predictions based on labeled training data. In supervised learning, the algorithm is trained using a set of input output pairs, where the input is the data and the output is the corresponding label or target value that we want the algorithm to predict[7]. The goal of the algorithm is to learn a mapping between the inputs and outputs, so that it can make accurate predictions on new, unseen data. There are many different types of supervised learning algorithms, including Regression algorithms are used to predict a continuous value, such as a price or a temperature[13]. Classification: Classification algorithms are used to predict a categorical value, such as whether an email is spam or not[15].

5.2 Logistic regression : Logistic regression is a type of supervised learning. It is used to calculate or predict the probability of a binary (yes/no,high/low/medium etc) event occurring[1].

Logistic regression is used to solve classification problems, and the most common use case is binary logistic regression, where the outcome is binary (yes or no, high/low/medium etc). In the real world, you can see logistic regression applied across multiple areas and fields.

- In health care, logistic regression can be used to predict if a tumor is likely to be benign or malignant[1].
- In the financial industry, logistic regression can be used to predict if a transaction is specious or not[2].
- In marketing, logistic regression can be used to predict if a targeted audience will respond or not[3].

6.CONCLUSION : From a proper analysis of analysis this system can be used perfect for detection of cancer who are suffer. LIME interpretability framework proved to be highly useful in understanding the model’s behavior and increasing dependableness which can help domain experts understand or discover hidden path. interpretability techniques whether global or local since it will be interesting to enhance the dependableness of models as well as the understanding of how those techniques work.

References :

1.https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data

2.https://www.datacamp.com/blog/what-is-machine-learning

3.https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8#:~:text=Logistic%20Regression%20is%20a%20Machine,%2C%20failure%2C%20etc.

4.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8946688/

5. Gupta, P.: Cross-Validation in Machine Learning - Towards Data Science (2017)

6.Reed, R., MarksII, R.J.: Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, p. 38 (1999)

7. https://pubmed.ncbi.nlm.nih.gov/34837074/

8.https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/

9.https://monkeylearn.com/machine-learning/#:~:text=Machine%20learning%20(ML)%20is%20a,to%20make%20their%20own%20predictions.

10.https://www.nature.com/articles/s41416-021-01633-1