Abhishek Jani **IT Department**

Roll No - 50

Experiment No.8
Explainable AI in healthcare for model interpretation
Date of Performance: 16/10/2023
Date of Submission:29/10/2023

Aim: To use Explainable AI in healthcare for model interpretation.
Objective: Given example statements or sentences with logo, use Explainable AI to display prediction probabilities using Lime TextExplainer
Theory:
Explainable AI (XAI) is a field of artificial intelligence that focuses on developing techniques and methods to make machine learning models more understandable and interpretable for humans. Interpretability is crucial for a variety of reasons, including trust, accountability, and making informed decisions based on AI predictions. Here are some key concepts and techniques in explainable AI for model interpretation:

Feature Importance: Understanding which features or input variables contribute the most to a model's predictions is a fundamental aspect of model interpretability. Common methods for feature importance include permutation feature importance, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations).

Model-Agnostic Interpretability: Many XAI techniques are model-agnostic, meaning they can be applied to various machine learning models. Techniques like LIME and SHAP are not tied to any specific model and can provide explanations for any black-box model.

Decision Trees and Rule-Based Models: Decision trees and rule-based models are inherently interpretable. They partition the data into branches based on feature values, allowing users to understand the decision-making process. However, they might not be as accurate as complex models for some tasks.

Local Interpretability: Local interpretability focuses on explaining the predictions of a model for a specific instance or data point. This can be useful for understanding why a model made a particular prediction in a given context.

Global Interpretability: Global interpretability aims to provide a holistic view of the model's behavior across the entire dataset. This can include insights into the relationships between features and the overall decision boundaries.

Visualization: Visualization techniques can help users understand model behavior by representing data and model predictions graphically. Tools like partial dependence plots and feature importance plots are commonly used for this purpose.

SHAP Values: SHAP (SHapley Additive exPlanations) is a widely used technique that provides a unified measure of feature importance and can explain the output of any machine learning model. It is based on game theory and provides a coherent and consistent way to allocate contributions of each feature to a prediction.

Counterfactual Explanations: Counterfactual explanations provide a different perspective by showing what changes in input features would lead to a different model prediction. This is particularly useful for understanding how to achieve a desired outcome or why a specific prediction was made.

Anchors: Anchors are simple, human-friendly rules that describe the behavior of a model for a given instance. They define a minimal set of conditions that, when met, guarantee a certain prediction.

Ethical Considerations: Explainable AI is crucial in ethical AI practices. It helps identify and mitigate biases in models and ensures that AI systems do not make discriminatory or harmful decisions.

Overall, explainable AI is an evolving field that offers a range of methods and techniques to enhance the transparency and interpretability of machine learning models. These techniques help make AI more trustworthy and accessible to users, including non-experts, regulators, and stakeholders who need to understand and trust AI decisions.

Code: -

Part 1:

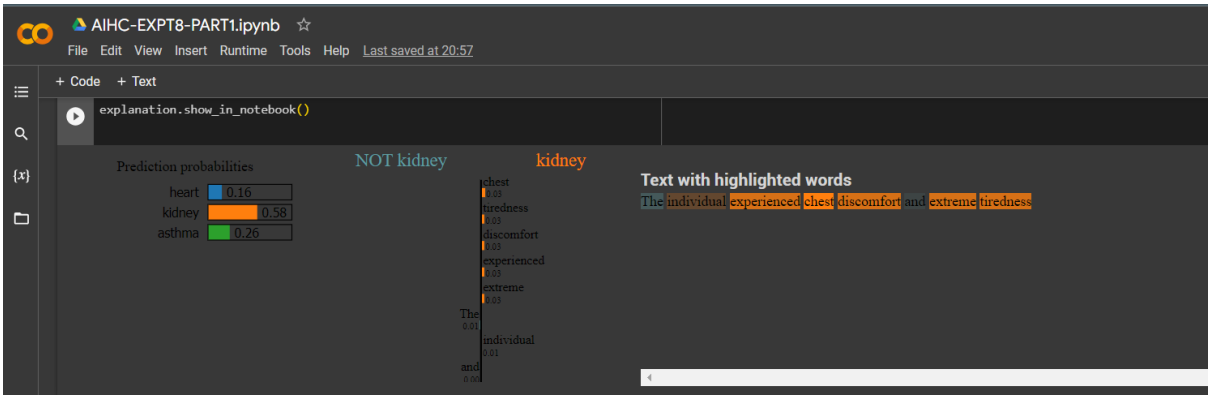
AIHC-EXPT8-PART1.ipynb

File Edit View Insert Runtime Tools Help Last saved at 20:57

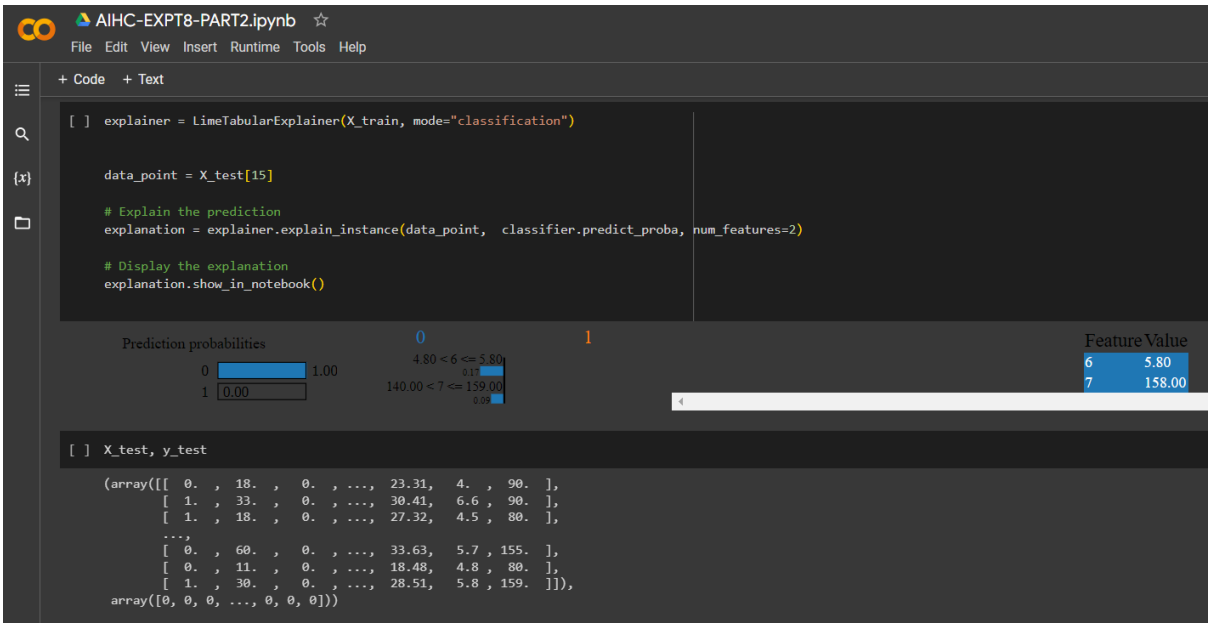
+ Code + Text

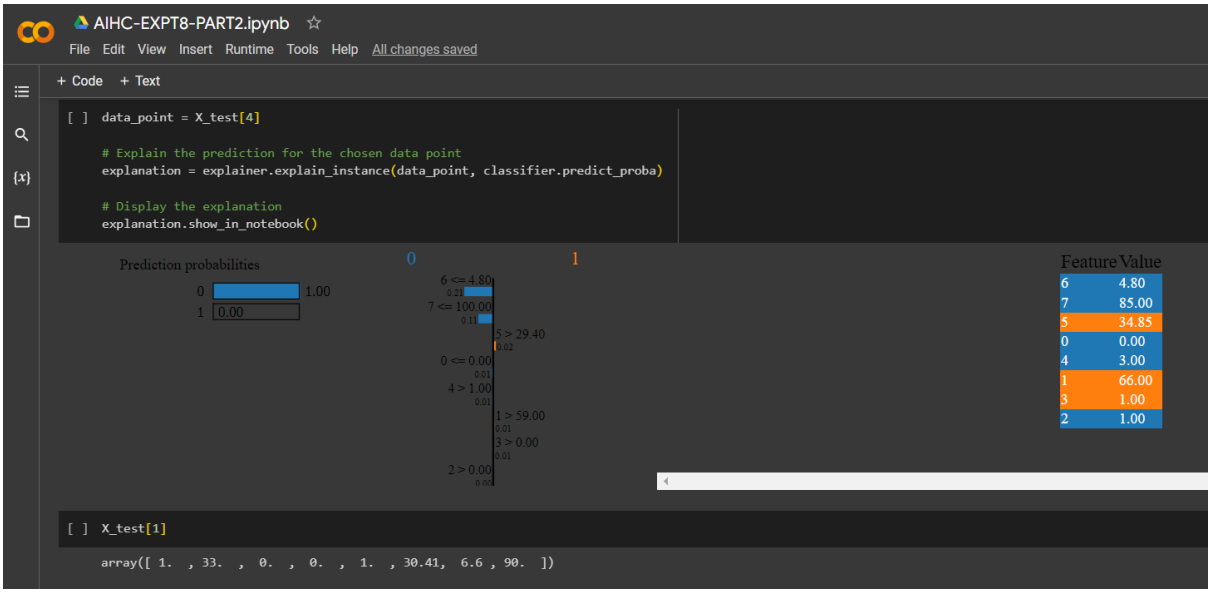
```
import lime
from lime.lime_text import LimeTextExplainer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
import numpy as np
import random

# Example sentences with labels
texts = [
    "The individual experienced chest discomfort and extreme tiredness",
    "The patient voiced concerns about breathlessness and a sense of lightheadedness",
    "There were signs of wheezing and persistent coughing in the individual",
    "Frequent changes in urination patterns and lower back pain were evident in the patient",
    "The patient reported chest constriction and feelings of weariness",
    "Swelling and noticeable difficulty in breathing were observed in the patient",
    "The patient presented with dizziness and complained of back pain",
    "The patient exhibited altered urination patterns and wheezing",
    "I had a sensation of chest pain along with frequent urination changes",
    "The patient's condition involved swelling and the description of tightness in the chest",


]

labels = [
    "heart", "heart", "asthma", "kidney",
    "heart", "kidney", "kidney", "asthma",
    "heart", "kidney"
]
```

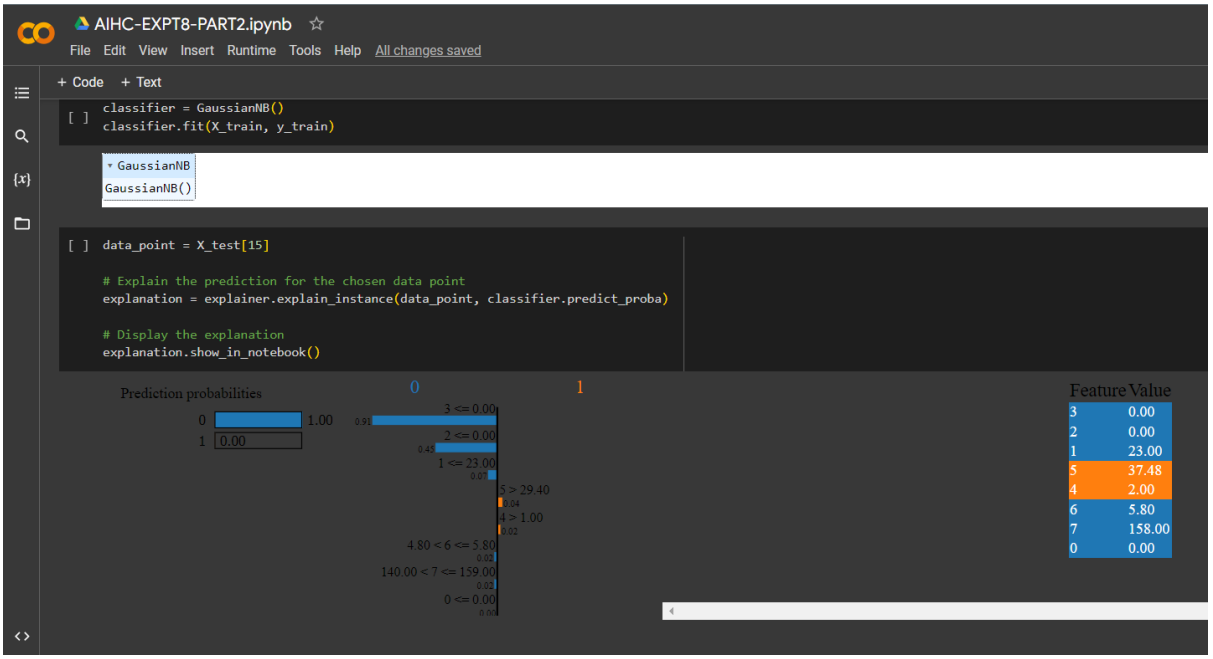
```
tfidf_vectorizer = TfidfVectorizer()
X = tfidf_vectorizer.fit_transform(texts)


naive_bayes_classifier = MultinomialNB()
naive_bayes_classifier.fit(X, labels)


def predict_proba(texts):
    X = tfidf_vectorizer.transform(texts)
    return naive_bayes_classifier.predict_proba(X)


explainer = LimeTextExplainer(class_names=["heart", "kidney", "asthma"])


prediction_to_explain = texts[0]
explanation = explainer.explain_instance(prediction_to_explain, predict_proba)


explanation.show_in_notebook()
```


AIHC-EXPT8-PART1.ipynb
File Edit View Insert Runtime Tools Help Last saved at 20:57
+ Code + Text
explanation.show_in_notebook()
Prediction probabilities
heart 0.16
kidney 0.58
asthma 0.26
NOT kidney
kidney
chest
tiredness
discomfort
experienced
extreme
The
individual
and
Text with highlighted words
The individual experienced chest discomfort and extreme tiredness

Part 2:


AIHC-EXPT8-PART2.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
[ ] explainer = LimeTabularExplainer(X_train, mode="classification")
data_point = X_test[15]
# Explain the prediction
explanation = explainer.explain_instance(data_point, classifier.predict_proba, num_features=2)
# Display the explanation
explanation.show_in_notebook()
Prediction probabilities
0 1.00
1 0.00
0
1
4.80 < 6 <= 5.80
140.00 < 7 <= 159.00
Feature Value
6 5.80
7 158.00
[ ] X_test, y_test
(array([[0. , 18. , 0. , ..., 23.31, 4. , 90.],
[1. , 33. , 0. , ..., 30.41, 6.6 , 90.],
[1. , 18. , 0. , ..., 27.32, 4.5 , 80.],
...,
[0. , 60. , 0. , ..., 33.63, 5.7 , 155.],
[0. , 11. , 0. , ..., 18.48, 4.8 , 80.],
[1. , 30. , 0. , ..., 28.51, 5.8 , 159.]]),
array([0, 0, 0, ..., 0, 0, 0]))





Google Collaboratory Link: -

AIHC-EXPT8-PART1.ipynb

AIHC-EXPT8-PART2.ipynb

Conclusion: -

Thus we have successfully performed Explainable AI in healthcare for model interpretation.