Lab Assignment 07: Counterfactual Explanations for Classification Models

Name: K. Saiteja
Hall Ticket No: 2303A52325
Batch: 35

Objective

The aim of this lab is to understand and implement counterfactual explanations in machine learning models. Counterfactuals provide "what-if" scenarios that help interpret model decisions. In this assignment, the Lung Cancer Risk dataset was used to demonstrate how minimal feature changes can flip a prediction and improve trust in AI systems.

Dataset Description

The dataset used is the Lung Cancer Risk dataset. It contains 200 patient-related records with features such as age, gender, smoking habits, radon exposure, asbestos exposure, secondhand smoke exposure, COPD diagnosis, alcohol consumption, family history, and lung cancer diagnosis. Preprocessing steps: - Missing values in alcohol consumption were handled. - Categorical variables were encoded using Label Encoding. - Numerical features were scaled for model training. For faster execution, only the first 200 rows were used.

Methodology

1. Preprocessing: Missing values handled, categorical variables encoded, and features scaled. 2. Models: Logistic Regression and Random Forest were trained for binary classification. 3. Evaluation: Models were evaluated using Accuracy, Precision, Recall, and F1-score. 4. Counterfactuals: DiCE library was used to generate counterfactual examples for test instances predicted negative. 5. Analysis: Influential features were identified, and realism/actionability of counterfactuals was discussed.

Model Performance Results

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.64	0.67	0.89	0.76
Random Forest	0.75	0.74	0.94	0.83

Counterfactual Examples

Counterfactual examples were generated using the DiCE library. For an instance predicted as negative (no cancer), three counterfactuals were produced that flipped the prediction to positive (cancer). Key influential features in the counterfactual generation included Age and Radon Exposure. These counterfactuals show minimal but realistic changes in features that alter model outcomes.

Analysis and Reflection

Analysis indicated that features like age and radon exposure had the most influence in changing predictions. These are realistic and actionable since they relate to lifestyle and environmental conditions. When comparing Euclidean and Manhattan distance metrics, slight differences were observed in which features were prioritized for modification. Counterfactual explanations improve trust and transparency by showing patients and doctors how small changes influence predictions. Beyond healthcare, counterfactuals are valuable in finance (loan approvals), education (admissions), and hiring systems for fairness and explainability.