Introduction:

Lately, Machine Learning algorithms to make previsions or automatic classification increased their performance and complexity.

Some families of algorithms, like decision trees, always give a perceptible and interpretable path how the algorithm takes the multiple decisions to arrive at the final result.

In opposition, some families like SVM's, Deep neural networks, or Random Forests can get very good results, but the interpretability and explanation of the decisions are totally out of easy comprehension to humans.

Those algorithms act like a Black Box, we can get good results but almost nothing about the interpretability or explanation of the process.

That problem is very relevant when we have a good results algorithm, but we need to understand the decision and check if the algorithm is fair in the internal decisions.

Many approaches exist to analyze these "Black Box" algorithms, one of them being the LIME (Local interpretable model-agnostic explanations) algorithm.

Task:

Attached, you can find a training data set (train_lblencoder_df.csv) about health insurances and an already trained BlackBox model (LinearRegressor_Lblencoder_model.pkl) to give a Linear Regression about that data.

In the data set we have the following columns:

age, sex, bmi, children, smoker, region, charges

Where the first 6 columns are the features, and the last, "charges", is the target/goal of the Linear Regression that we want to predict.

The Goal is to explain with LIME the predictions of the given model for the 2 samples (toExplain.txt).

Hints:

Python

Load the model (LinearRegressor_Lblencoder_model.pkl) with the lib joblib

Python Lime implementation: https://github.com/marcotcr/lime