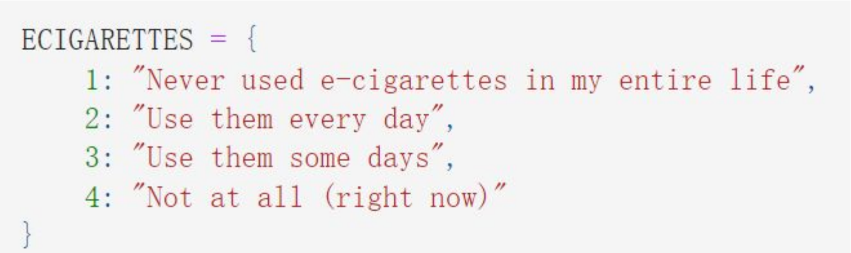
Predicting the Risk of Heart Disease Using Key Features: A Comparative Analysis of Machine Learning Models

Question: Can we reduce racial bias in ML predictions using different models and preprocessing methods?
Why: Heart disease is still a leading cause of death across various races in the United States that almost half of the American population contains at least one major risk factor. With modern computational tools and techniques, there is a need to analyze and understand these **risk factors** better and predict heart disease earlier.
Dataset: The dataset, titled “Indicators of Heart Disease (2022 UPDATE),” is a source from CDC’s annual Behaviour Risk Factor Surveillance System survey, presenting over 400000 adult responses related to health status from 2022. **The target variable “HadHeartAttack” will be treated as a binary response “yes” or “no”.**



```
ECIGARETTES = {
    1: "Never used e-cigarettes in my entire life",
    2: "Use them every day",
    3: "Use them some days",
    4: "Not at all (right now)"
}
```

Models to be used: **Knn, decision tree, neuron network**
Methods:
- preprocessing
- feature selection and model regularization
- hyperparameter tuning:

For KNN, we will test with both Euclidean Distance and Manhattan Distance.

For Decision Tree, we will go through the basic components such as min leaf sample, max tree depth, and so on

For neuron network, we use SGD, and determining the number of layers and neurons on each layers