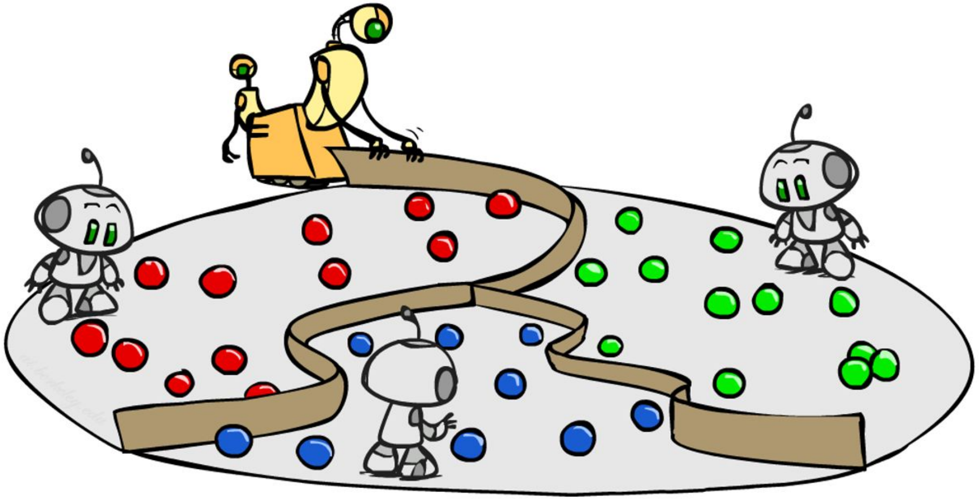
CS-ELEC1A: Advanced Intelligent Systems

Lab Exercise #2: Train Your Own Decision Tree



Problem Context

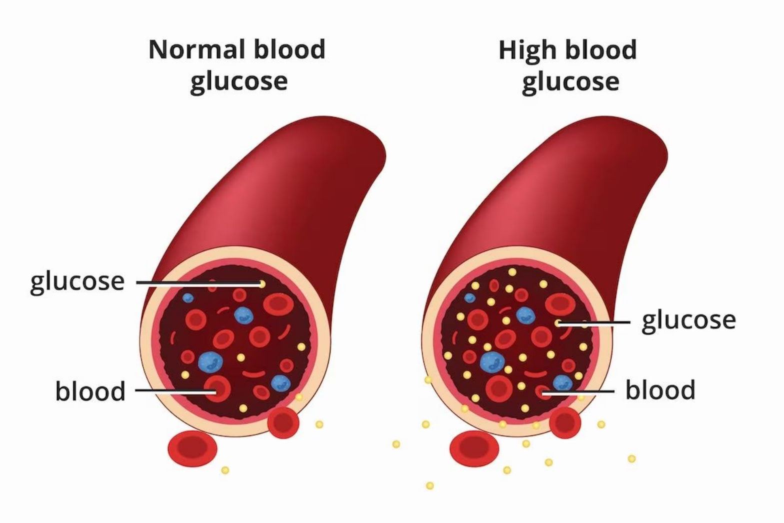
Context: Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. In the past 3 decades the prevalence of type 2 diabetes has risen dramatically in countries of all income levels. Type 1 diabetes, once known as juvenile diabetes or insulin-dependent diabetes, is a chronic condition in which the pancreas produces little or no insulin by itself. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival. There is a globally agreed target to halt the rise in diabetes and obesity by 2025.

Problem Context

World Health Organization

Suppose you are working as a Machine Learning Scientist at a Non-Profit Organization. Your company is overwhelmed by the number of patients that want to have a diagnosis. Given your expertise in the field of Machine Learning, your goal is to create a model that identifies if a person is likely to have diabetes based on the patient’s Number of Pregancies, Glucose Levels, Blood Pressure, Skin Thickness, Insulin, Body Mass Index, Diabetes Pedigree Function, and Age.

Provided Dataset



Pregnancies: Number of Pregnancies

Glucose: Glucose Levels in Blood

Blood Pressure: Blood Pressure measurement

Skin Thickness: Thickness of the skin

Insulin: Insulin Levels in Blood

BMI: Body Mass Index

Diabetes Pedigree Function: Diabetes likelihood depending on the subject's age and his/her diabetic family history

Age: How old the patient is

Step #2: Downloading the Dataset

Retrieve the dataset from https://drive.google.com/file/d/1XVv0BT50CM9avOhGiMr3T6e3pxo3d9Y4/view?usp=sharing

Step #3: Import Statements

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score

import seaborn as sns
import pandas as pd
import numpy as np
```

Step #4: Loading and Checking the Dataset

```
df = pd.read_csv("diabetes.csv")

df.head(10)

df.info()

df['Outcome'].value_counts()
```

Step #5: Simple Exploratory Data Analysis

```
sns.distplot(df['BloodPressure'])

sns.distplot(df['SkinThickness'])

sns.distplot(df['Age'])
```

Step #6: Train-Test Split

```
from sklearn.model_selection import train_test_split

X = df[['BloodPressure', 'SkinThickness', 'Age']]
y = df['Outcome']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Step #7: Training the Decision Tree

```
clf = DecisionTreeClassifier(random_state=0, criterion='entropy', max_depth=2,
                                       min_samples_split=2,
                                       min_samples_leaf=1)
clf.fit(X_train, y_train)
```

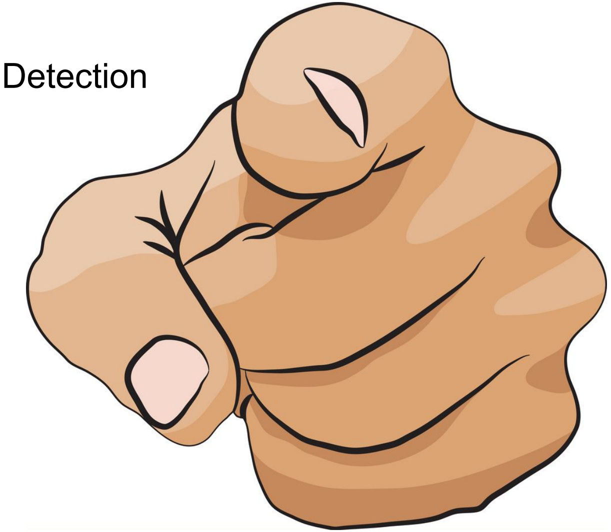
Step #8: Evaluating Performance

```
y_preds = clf.predict(X_test)

print("Accuracy %.4f" % accuracy_score(y_test, y_preds))
print("Precision %.4f" % precision_score(y_test, y_preds))
print("Recall %.4f" % recall_score(y_test, y_preds))
print("F1 %.4f" % f1_score(y_test, y_preds))

plot_tree(clf,
        feature_names = X.columns,
        class_names = ['No Diabetes', 'Diabetes'],
        filled = True)
```

What You Need to Do

Objective:

Improve the Accuracy, Precision, Recall, and **F1** metric for Diabetes Detection



Possible Things To Experiment On:

- Other **preprocessing methods**
- Conduct **feature engineering** (add, create, delete features)
- Make **changes to hyperparameters**
- And many more!

For the Write-Up:

- Recommended to have:
    - Introduction: discussion of premise and data exploration
    - Methodology: details of overall methodology
    - Experiments: explanation of various trials and experiments
    - Results and Analysis: discussion of why the results came to be with some additional analysis
    - Conclusions & Recommendations: highlight of write-up, thoughts, improvements