**Lab 3**

**Due 2nd Feb**

**Worth 5%**

In this lab you build a Decision Tree classifier with Scikit-learn.

We will use the Pima Indians Diabetes Database which is already available on Blackboard.

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The dataset consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**Setup**

You will need to install the following Python modules:

Pandas

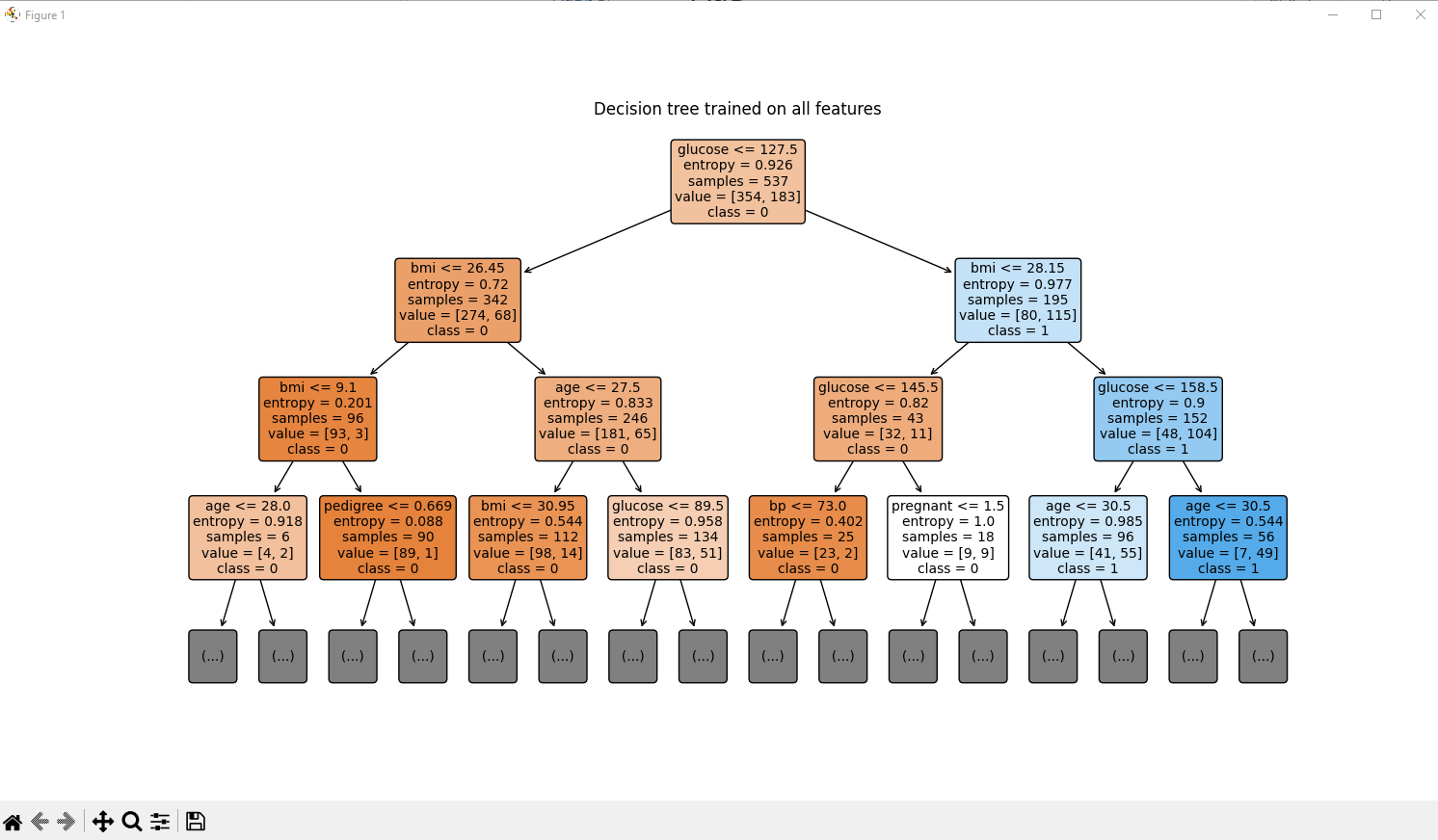
Scikit-learn

Graphviz (if you are using a jupyter notebook. You will also need to install the Graphviz application on your computer. The Graphviz python module is used to call the Graphviz application on the computer).

Download and get ID3.py to run properly.

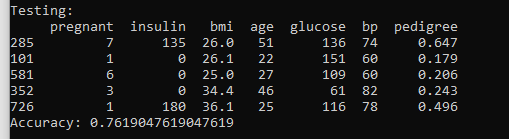
You should get a window to pop up with something like the figure below. This is a visual representation of the classification model created by the DecisionTreeClassifier in sklearn.

Note it doesn’t show the whole tree. Modify the script to show the whole tree.



**To do:**

1. Now that you have a functioning ID3.py, you can see how the model “clf” is created. You can also see how well it has been trained and is able to predict the outcome for the test data. Modify the program so that it is possible for a user to check the data from a new patient against the model and give a prediction about whether they have diabetes or not.



1. On Blackboard there is another dataset called Thyroid\_Diff.csv. You can read about the dataset here: <https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence/data>. Create a decision tree classifier as above but with this new dataset. Open the datafile and examine it. Compare it to the diabetes dataset. What do you notice?

**Hint:** Our AI algorithms only work with numbers! You might want to use the “.replace” function or something similar.