**DECISION TREE**

**ABSTRACT**

The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy* and *Information Gain* to construct a decision tree.

**In this session we are going to use the pima Indians dataset which classifies whether a person has diabetes or not using a binary classifier.**

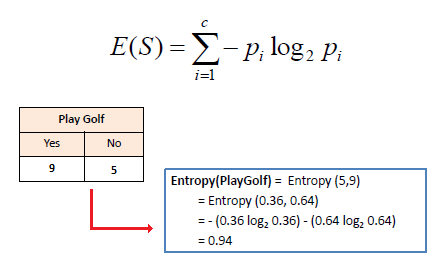
**I NTRODUCTION**

A **decision tree** is a [decision suppor](https://en.wikipedia.org/wiki/Decision_support_system)t tool that uses a [tree-lik](https://en.wikipedia.org/wiki/Tree_(graph_theory))e model of decisions and their possible consequences, including [chanc](https://en.wikipedia.org/wiki/Probability)e event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). It is one way to display an [algorith](https://en.wikipedia.org/wiki/Algorithm)m that only contains conditional control statements.

Decision trees are commonly used in [operations research](https://en.wikipedia.org/wiki/Operations_research), specifically in [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), to help identify a strategy most likely to reach a [goal](https://en.wikipedia.org/wiki/Goal), but also a popular tool in [machine learning](https://en.wikipedia.org/wiki/Decision_tree_learning).

**METHODOLOGY**

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.



**DATASET**

Pregnancies

Number of times pregnant

Glucose

Plasma glucose concentration a 2 hours in an oral glucose tolerance test

BloodPressure

Diastolic blood pressure (mm Hg)

SkinThickness

Triceps skin fold thickness (mm)

Insulin

2-Hour serum insulin (mu U/ml)

BMI

Body mass index (weight in kg/(height in m)^2)

DiabetesPedigreeFunction

Diabetes pedigree function

Age

Age (years)

Outcome

Class variable (0 or 1) 268 of 768 are 1, the others are 0

**ALGORITHM**

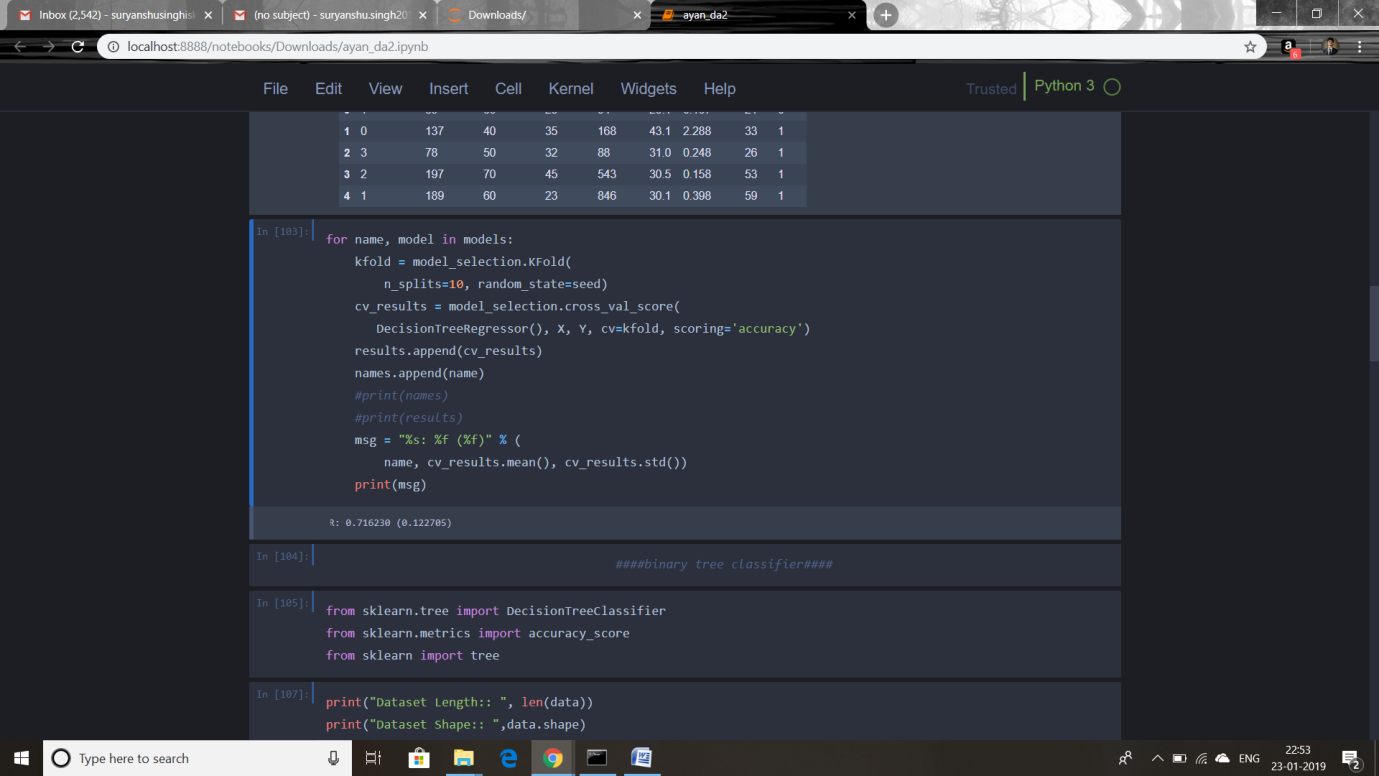
**Step 1**:Find the entropy for the target variable

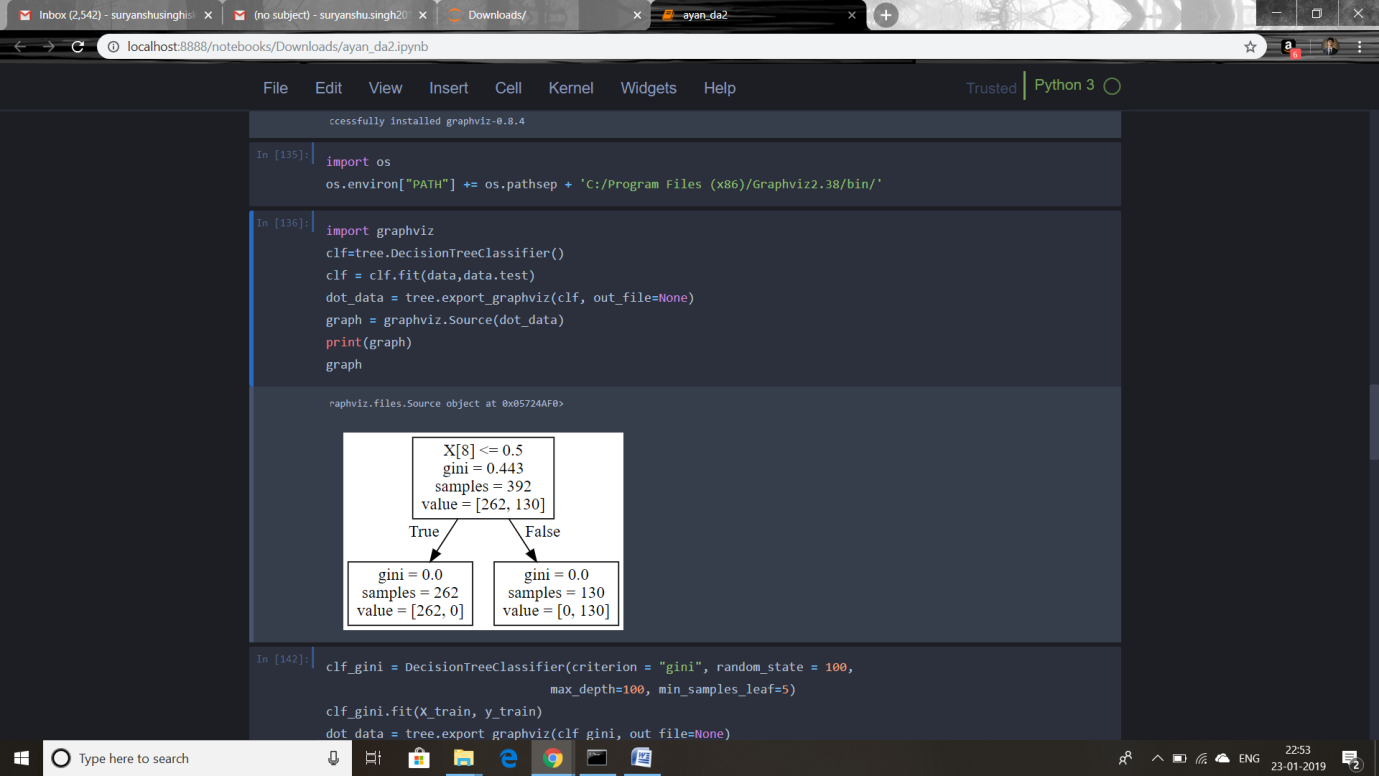
**Step 2**:Find entropy for each attribute and subtract it with that of the target variable to find information gain

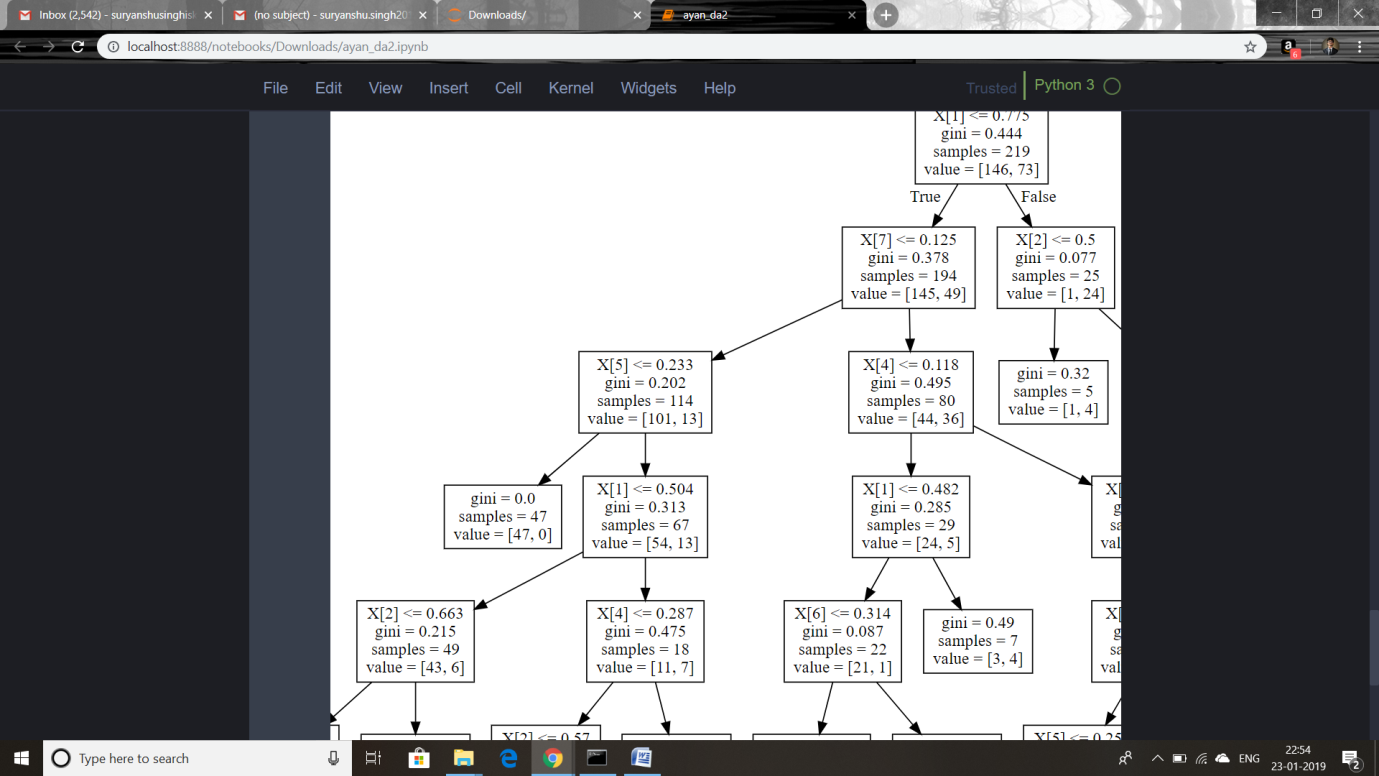
**Step 3**:Take the attribute with highest gain as the root and expand it

**Step 4**:If all the attributes in a single node correspond to only one class stop else perform steps(1-3) for each node.

**RESULTS**

****

****

****

**CONCLUSION**

We fit the data in the model and created a decision tree for classifying new data into the binary class(0 or 1)types present in the dataset .

**REFERENCES**

[**https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/**](https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/)

**https://www.kaggle.com/uciml/pima-indians-diabetes-database**