Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Кафедра Штучного інтелекту

Звіт з лабораторної роботи №3

з курсу Машинне навчання

на тему “ Навчання і класифікація за допомогою дерев рішень ”

Виконав:

студент гр. ІТШІ-17-1

Гура А.О.

Прийняла:

Кулішова Н.Є.

Харків 2019

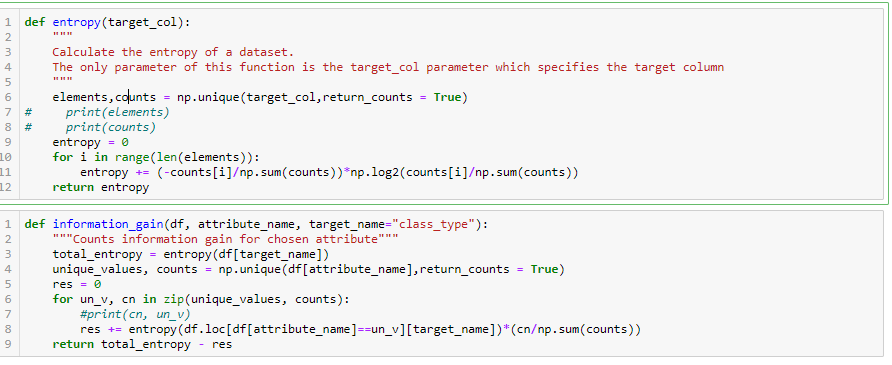
**Цель работы:** построить дерево решений для поставленных задач, применить алгоритм классификации с помощью дерева решений.

**Формальная постановка задачи:** при выполнении данной работы необходимо ознакомиться с алгоритмом ID3 и реализовать деревья решений для наборов данных zoo и iris.

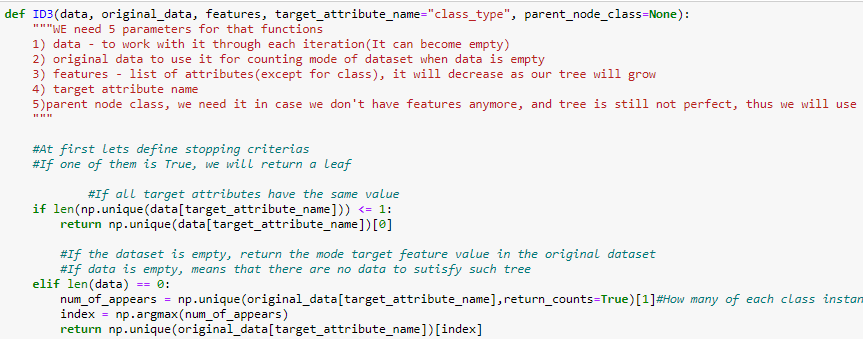
Также необходимо расклассировать тестовую выборку в каждом наборе и оценить точность (accuracy).

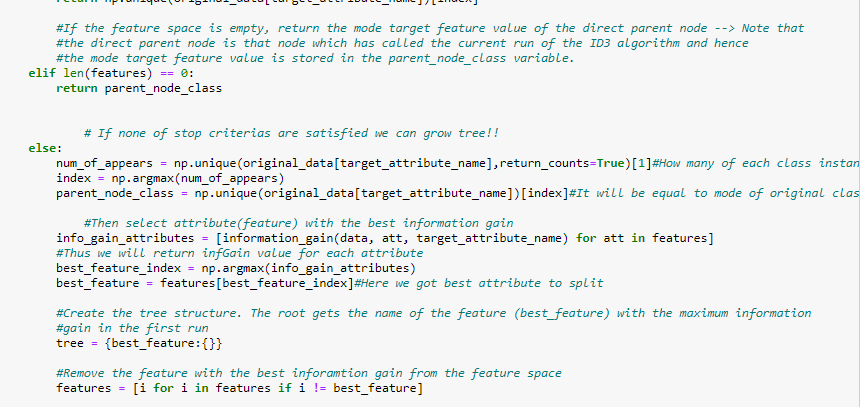
**Код алгоритма ID3 на языке Python:**

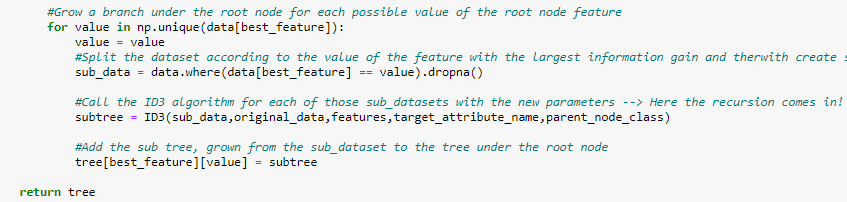
Функции для Entropy и Information Gain:

****

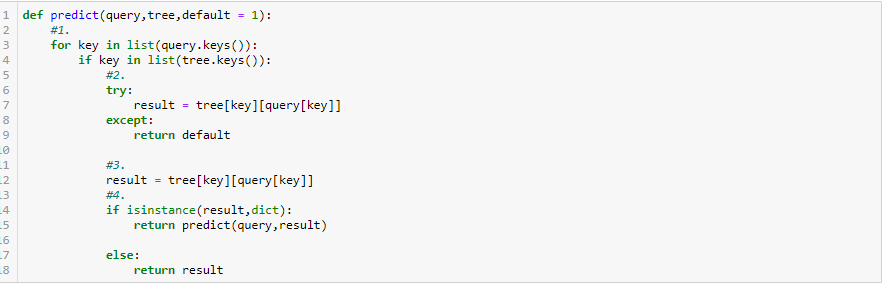
**Сам Алгоритм:**

****

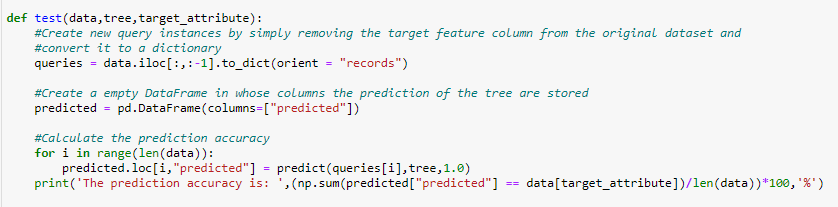
****

****

**Функция для предсказания тестовой выборки на основе дерева**

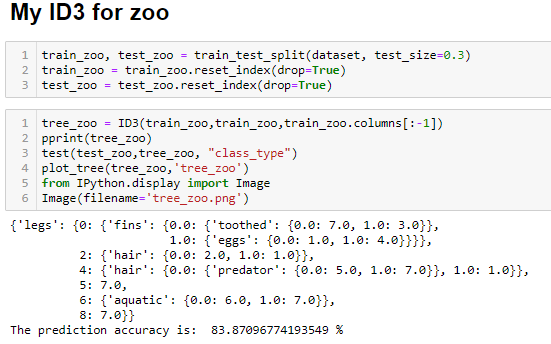
****

**Функция для оценки точности**

****

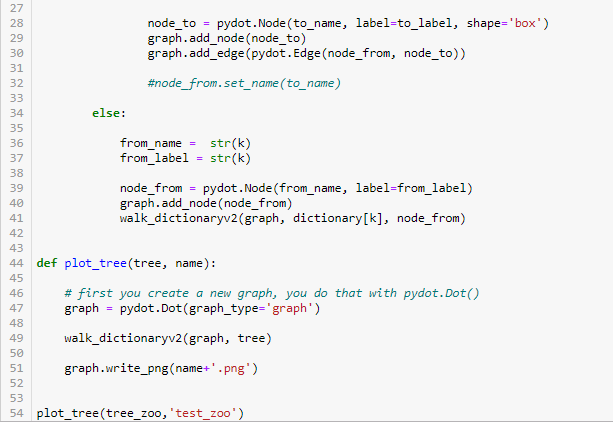
Таким образом, при анализе полученных результатов можно понять какие атрибуты будут наиболее информативны, при построении деревьев для обеих выборок.

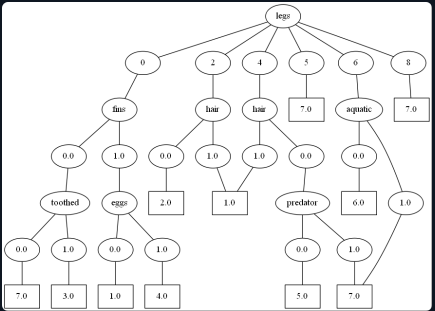
Результаты выполнения данного кода приведены ниже.



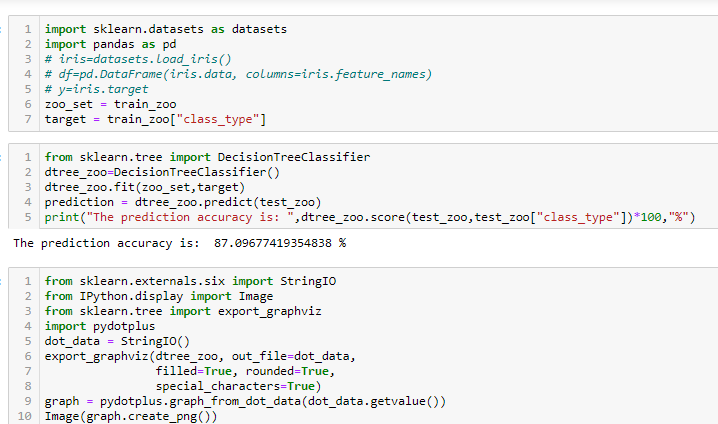
Для выборки зоопарк он выдает хорошую точность(80-90%). Однако вывод дерева производится в виде словарей. Я модифицировал это так, чтобы оно рисовало деревья.

Визуализация деревьев из словарей:

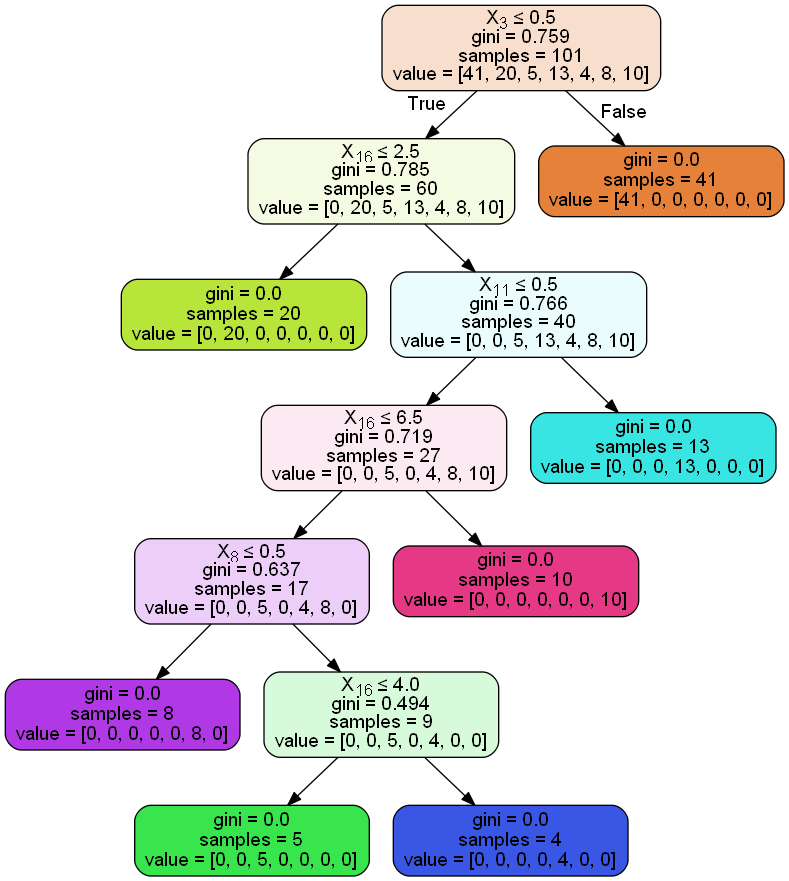


В результате вот такое дерево строится для зоопарка. 

Теперь сравним с древом решений из sklearn:

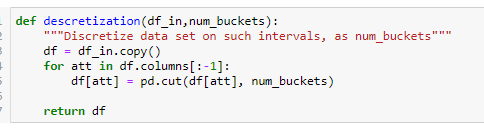


Как можно заметить точность отличается незначительно. И в результате мы получи вот такое вот дерево:

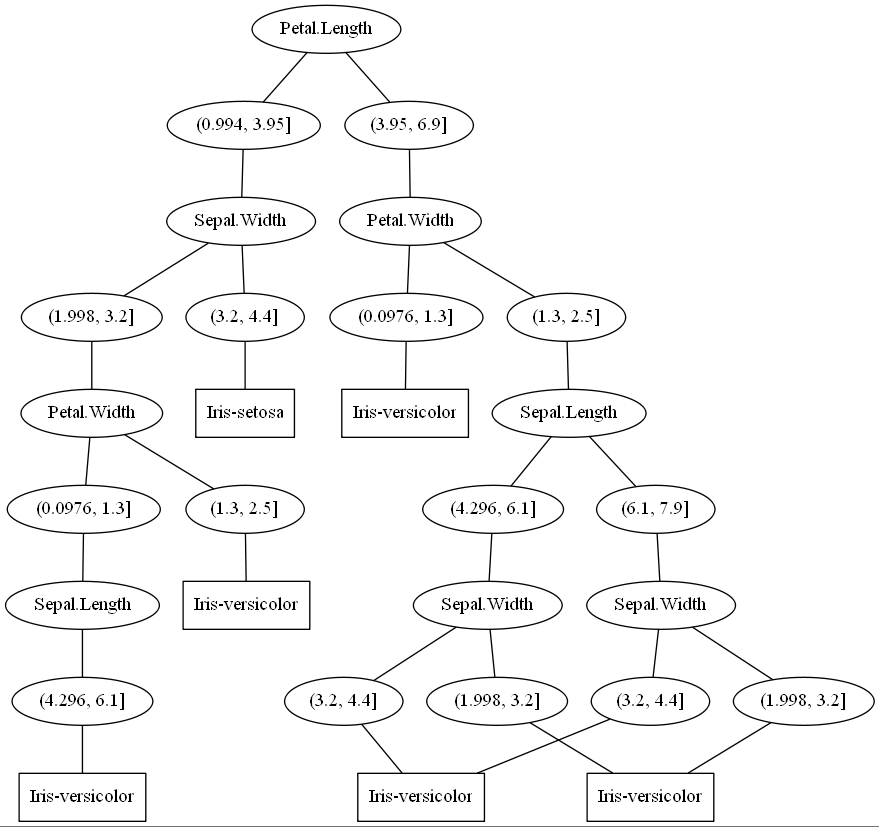
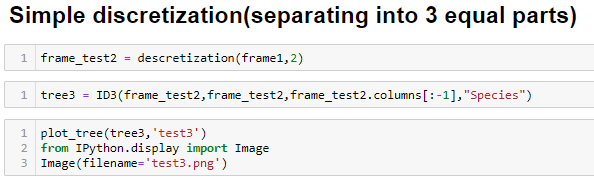


Как можно заметить, существенным преимуществом моего алгоритма является то, что он относится к данным как к категориям, в то время как алгоритм Sklearn работает с ними как с continious valued. И также мой алгоритм строит небинарные деревья.

Теперь рассмотрим его работу для Ириса

Очевидно что для ириса данные нам придется дискретизировать. 

Эта функция просто разбивает данные на равное количество участков, в зависимости от того что мы ввели.

Продемонстрирую работу моего алгоритма при разбиении на 3 промежутка. 

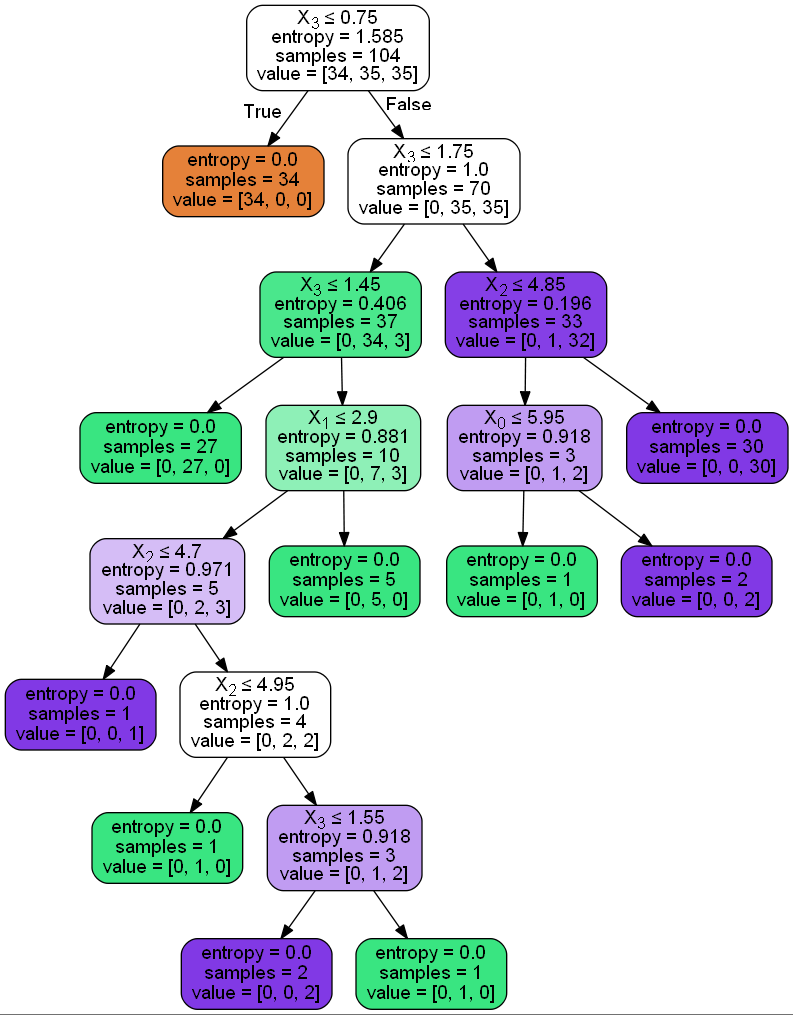
По итогу тестирования, я могу сказать что лучшую точность считает при разбиении на 3 и 6 промежутков:





Теперь сравним с результатами Sklearn:



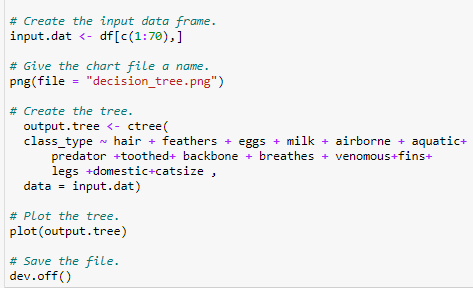


**Применения языка программирования R для построения дерева решений**

При реализации вышеупомянутых решений, было решено использовать уже готовые решения и средства языка R.

Код программы для выборки Зоопарк, сначала применяя библиотеку

party, а затем rpart –



Результаты выполнения программы приведены на рисунке 4:

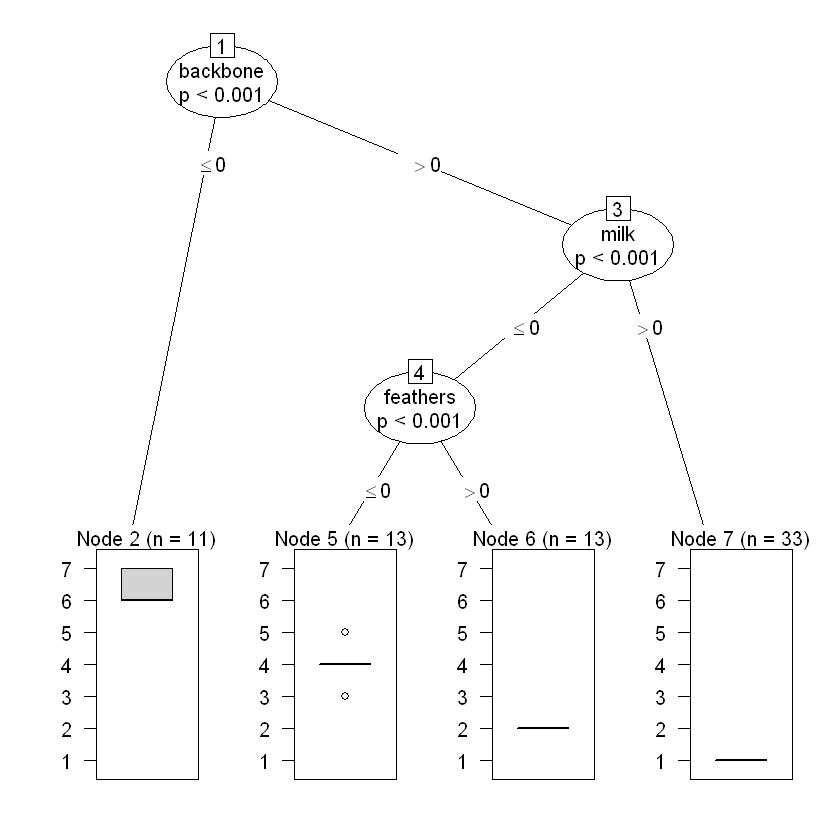
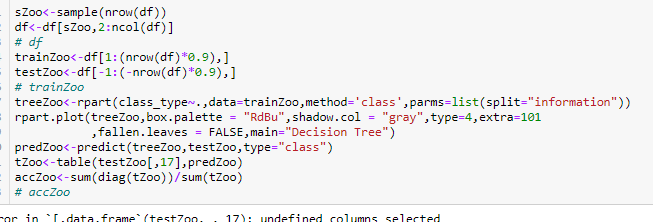
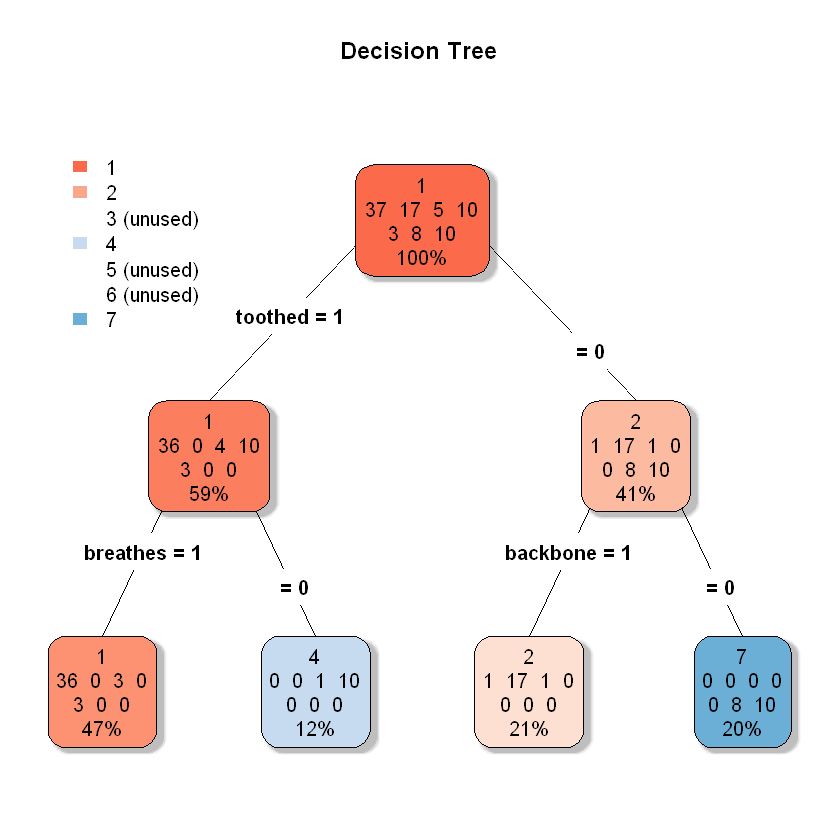


Рисунок 4 – Выполнение программы средствами языка R





**Выводы**

Данная работа дает возможность на примере алгоритма ID3 закрепить основные понятия по теме деревья решений, а также использовать на практике различные виды языков программирования и готовых модулей к ним для дальнейшего сравнения по критериям удобства, скорости и точности работы.

При рассмотрении данных методов были изучены основные понятия об построении деревьев решений, алгоритмах Gini и InfoGain. Также, рассмотренные примеры («Zoo» и «Iris») помогли лучше понять данную тему на белее простых примерах.

Более того, в ходе данной работы ознакомились с наличием алгоритмов CART и С4.5, а также с основными отличиями от ID3.