Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра информационных технологий

**ОТЧЕТ**

о выполнении лабораторной работы №2

по дисциплине «Машинное обучение»

Выполнили: ст. гр. 4ИТ

Хижний Е.Г.

Проверил: ст. препод.

Харченко А.В.

Краснодар

2022

Лабораторная работа №2.

**КЛАССИФИКАЦИЯ. ПОСТРОЕНИЕ**

**МОДЕЛИ КЛАССИФИКАЦИИ. СРАВНЕНИЕ РАЗЛИЧНЫХ**

**АЛГОРИТМОВ КЛАССИФИКАЦИИ И ВЫБОР ОПТИМАЛЬНОГО**

**В ORANGE**

**Цель и задача работы:**изучить основные методы классификации с использованием приложения «Orange Data Mining». Осуществить классификацию тестовых данных, используя разные алгоритмы. Научиться сравнивать результаты работы алгоритмов классификации и выбирать наиболее подходящий.

**Порядок выполнения работы**

1) Загрузить данные с помощью виджета File.

2) Осуществить классификацию данных с помощью алгоритмов Classification Tree, Logistic Regression, Naive Bayes, SVM, CN2 Rule Induction, Nearest Neighbors, Random Forest Classification

3) Осуществить кросс-валидацию с помощью виджета Test & Score, используя различные виды разбиения входных данных на тестовые и проверочные (cross validation, random sampling, leave one out).

4) Проверить различные варианты выбора тестовой и обучающей выборки в виджете Test & Score.

5) Обосновать выбор наилучшего алгоритма для классификации исходных данных. Использовать виджеты Confusion Matrix и ROC Analysis.

6) Для алгоритма Classification Tree вывести дерево решений в графическом виде.

7) Вывести ошибки классификации для разных алгоритмов на точечную диаграмму вместе с результатами правильной классификации.

8) Вывести ошибки классификации для разных алгоритмов в виде таблицы.

9) Вывести и проанализировать ROC-кривые для разных алгоритмов.

10) Осуществить классификацию данных файла, используя виджет Predictions. Вывести полученные результаты в виде таблицы.

11) Обосновать выбор оптимального алгоритма классификации. Подготовить отчёт.

**Ход работы**

В качества набора данных выбран стандартный файл zoo.bsk, содержащий информацию о 101 животном и их характеристиках.

Для добавления файла нажмем на виджет File, показанный на рис.1.

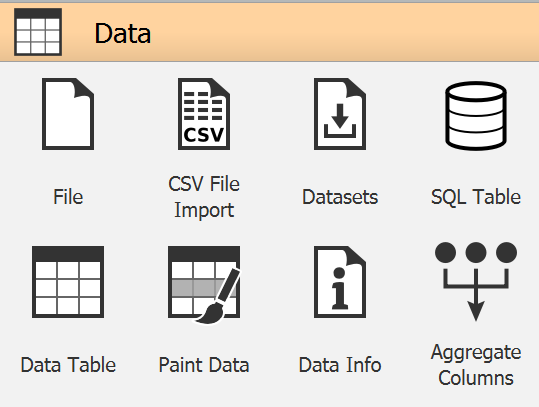


Рисунок 1 – Поле виджетов

На рис. 2 показано окно данного виджета, выберем необходимый файл и нажмем “Apply”.

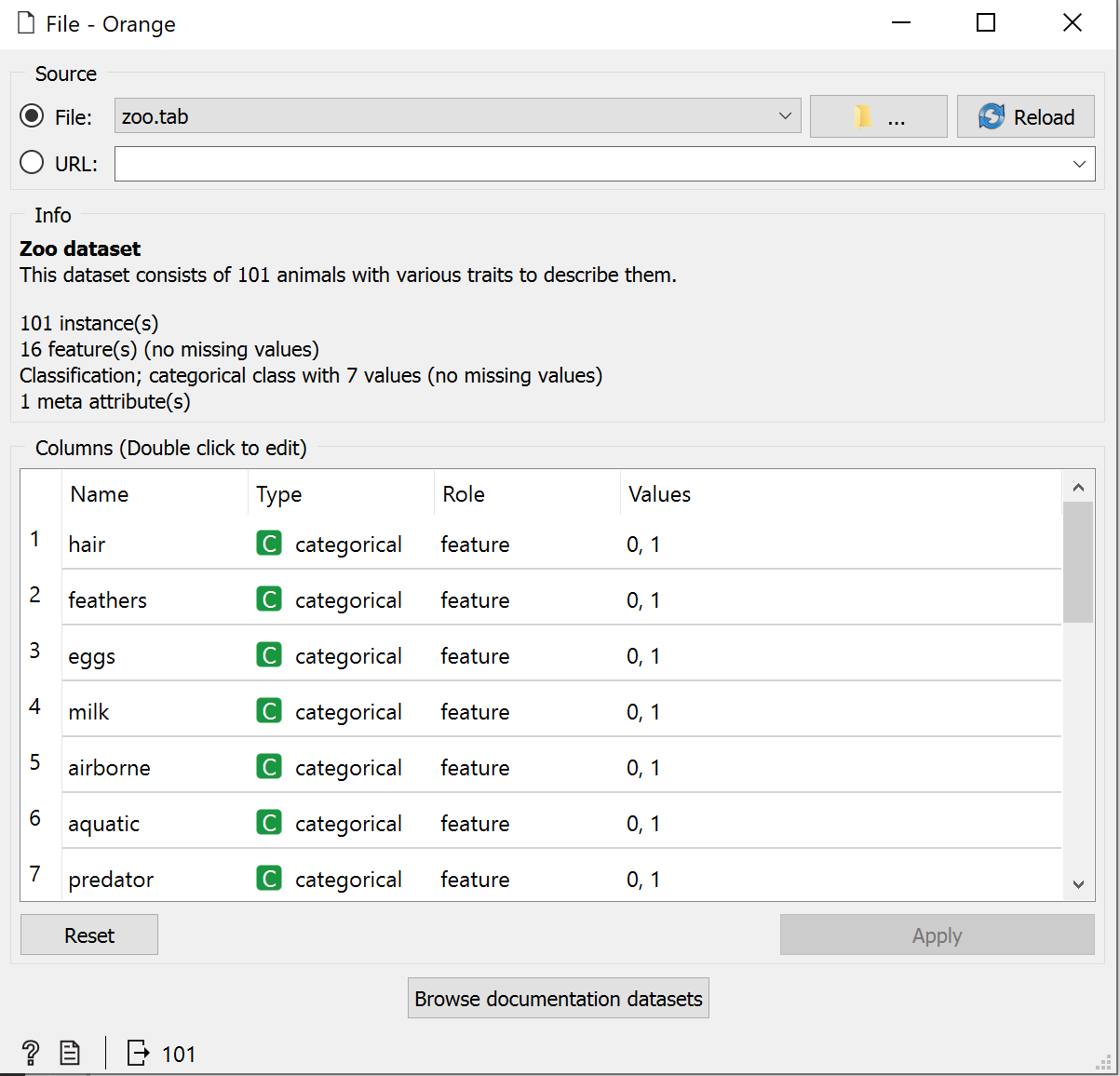


Рисунок 2 – Настройка виджета File

Осуществим необходимую классификацию алгоритмом классификации “Tree”, в нем выберем стандартные параметры, как показано на рис. 3. Ссвяжем с данными, добавим также виджет ”Tree Viewer” для просмотра полученного дерева. От Tree и File передадим данные в Test And Score и к Confusion Matrix, как показано на рис. 4.

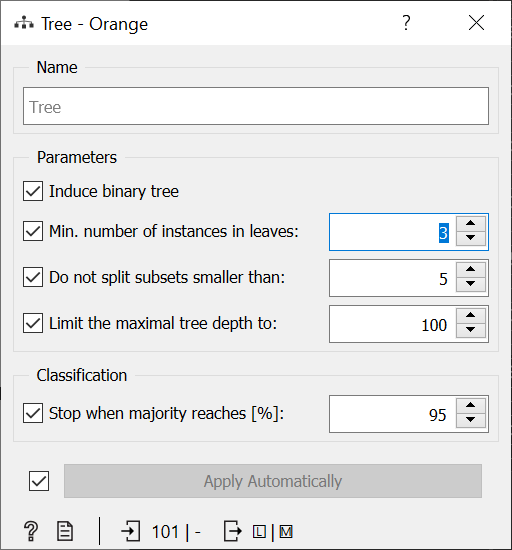


Рисунок 3 – Параметры виджета “Tree”

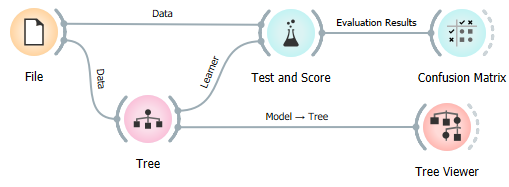


Рисунок 4 – Алгоритм классификации “Tree”

На рис.5 показано граф графическое представление дерева в Tree Viewer. По данному представлению можно определить к какому виду принадлежит животное

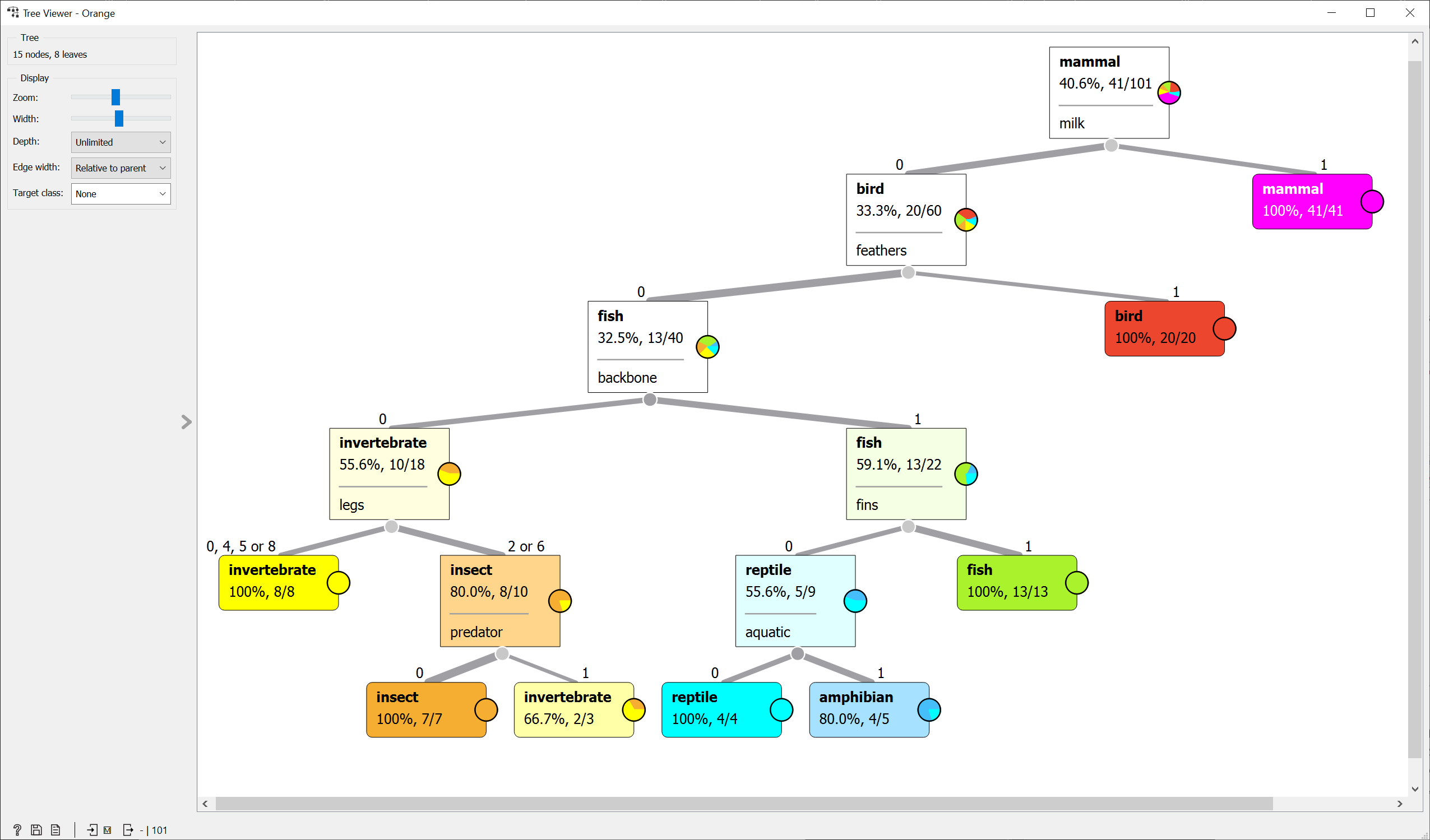


Рисунок 5 – Алгоритм классификации “Tree”

Добавим алгоритмы Logistic Regression, Naive Bayes, SVM, CN2 Rule Induction, Nearest Neighbors, Random Forest Classification. Соединим их с Test and Score, который позволяет оценить точность методов, используя при этом разные способы разбиения начальных данных. От него проведем связь с Confusion Matrix, который позволяет оценить ошибки классификации и ROC Analysis, график истинно положительного показателя против ложноположительного показателя теста Новая схема показана на рис. 6.

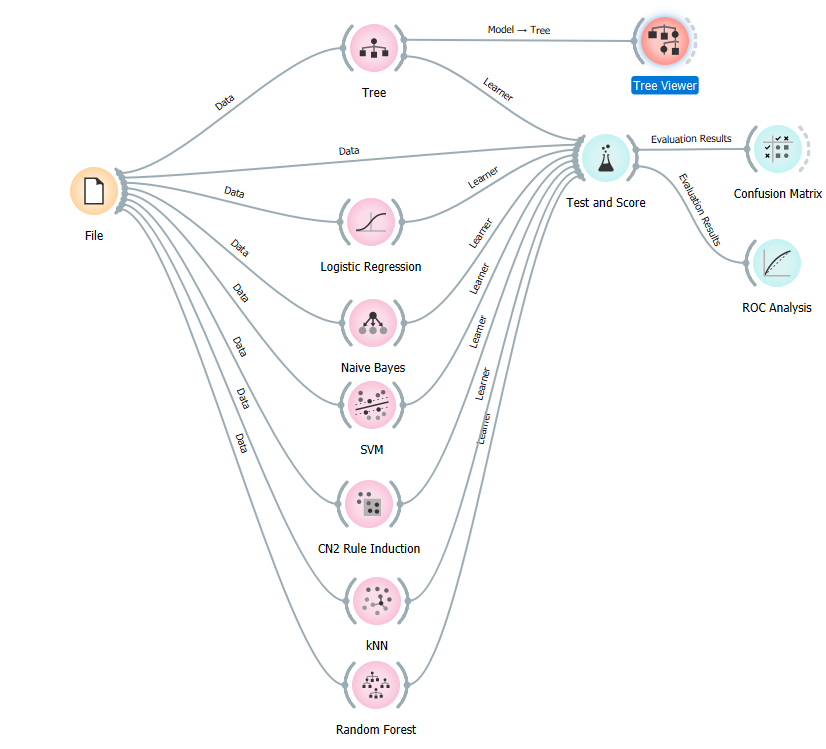


Рисунок 6 – Новая классификация

В Test and Score выберем методы cross validation, random sampling, leave one out. Результаты показаны на рис. 7–9.

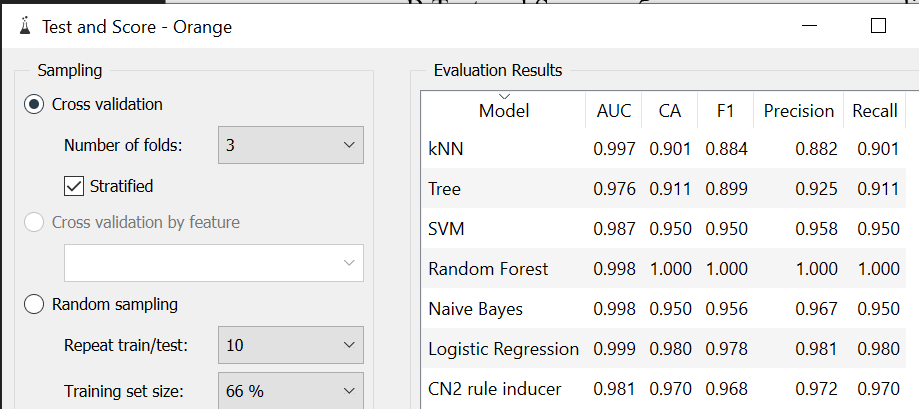


Рисунок 7 – cross validation

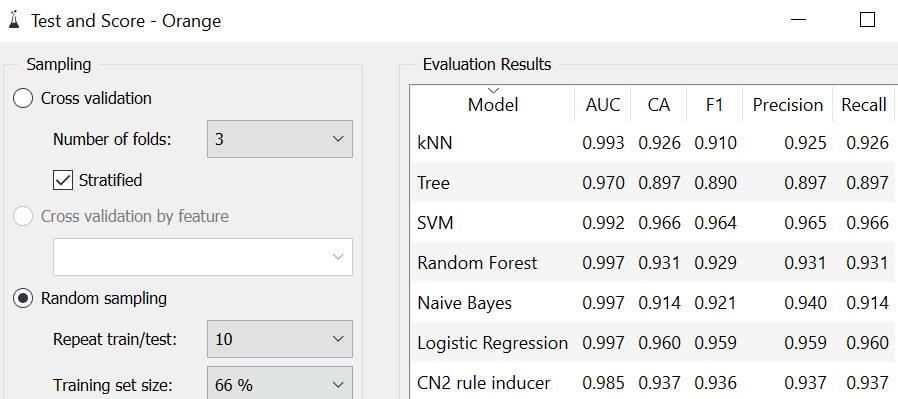


Рисунок 8 – random sampling

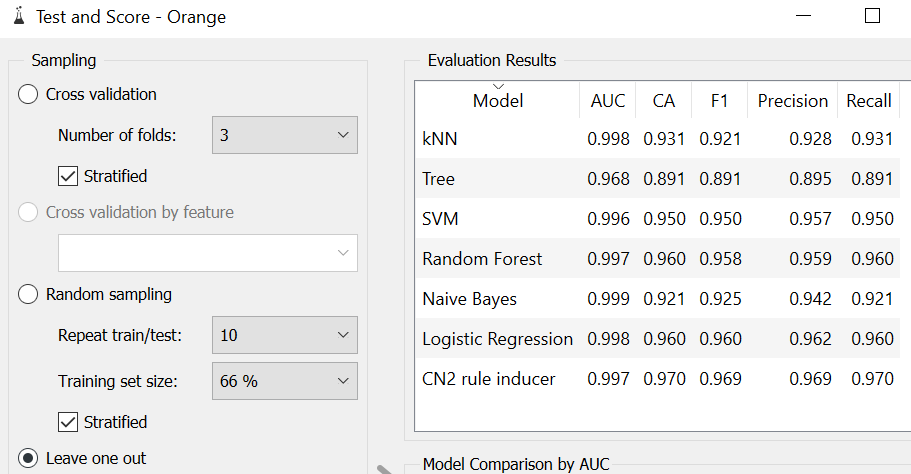


Рисунок 9 – leave one out

По всем показателям Лучшие результаты у Random Forest, Naive Bayes, Logistic Regression.

Просмотрим Confusion Matrix.

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Рисунок 10 – Tree

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 11 –Logistic Regression

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Рисунок 12 – Naive Bayes

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Рисунок 13 –SVM

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описаниеРисунок 14 –CN2

Рисунок 15 – Предсказания Random Forest

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Рисунок 16 –kNN

Метод SN2 сделал 3 ошибки, став лучшим методам. Также можно выделить Logistic Regression и Random Forest с 4-мя ошибками. Худшим оказался метод KNN с 26 ошибками классификации.

Рассмотрим Roc Analysis (рис. 17).

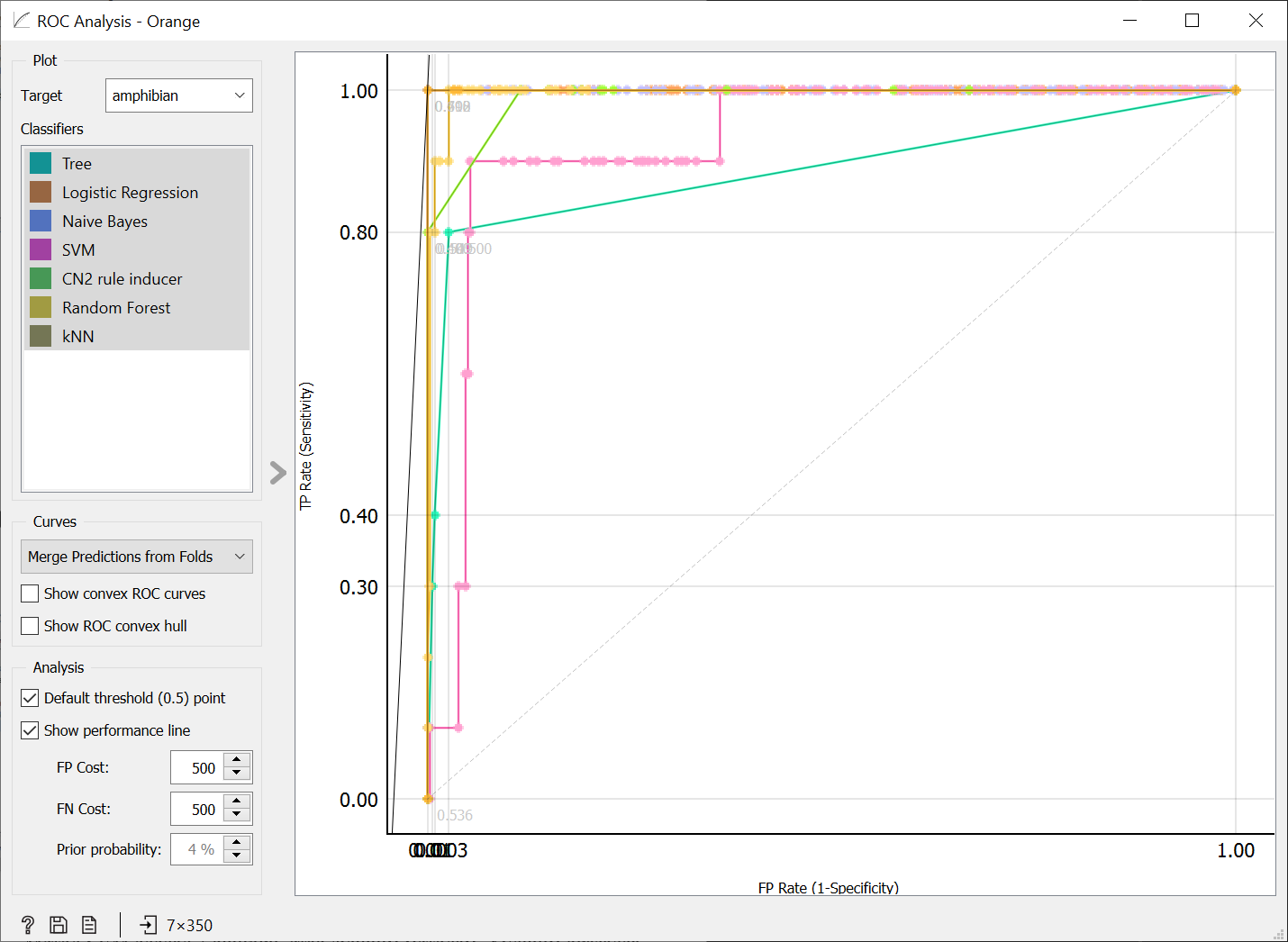


Рисунок 17 –ROC Analysis

Осуществим классификацию данных файла, используя виджет Predictions. Выведем полученные данные в таблицу (рис. 18).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 18 – Виджет Predictions

**Вывод:** в данной лабораторной работе былиизучены основные методы классификации с использованием приложения «Orange Data Mining». Осуществлена классификация тестовых данных, используя разные алгоритмы. Logistic Regression является самым точным алгоритмом для zoo.tab.