# 

# 

# 

# Отчет по Лабораторной работе №1

## 

## 

## 

## 

## 

## 

## Курс: Интеллектуальные системы.

## Выполнил: Демянчук А.И. (АПИМ-17)

### 1. Описание структуры исходных данных и задачи в терминах предметной области и машинного обучения

Задача относится к области сельского хозяйства. Данные представляют собой набор характеристик трех различных сортов пшеницы (Kama, Rosa и Canadian) по 70 элементов каждого вида в выборке. Визуализация внутренней структуры зерна была проведена при помощи техники мягкого рентгеновского излучения. Изображения структуры зерен были сохранены на рентгеновские пластины Kodak. Зерна пшеницы были получены с экспериментальных полей Института Агрофизики Польской Академии Наук (г. Люблин).

#### Формальное описание задачи в терминах машинного обучения

Задачей является: разработать алгоритм классификации зерен пшеницы трех сортов (Kama, Rosa и Canadian) на основе семи параметров измерений зерна.

Класс задачи: задача классификации.

Исходные признаки:

1. Поверхность (A)

2. Периметр (P)

3. Компактность (C = 4\*pi\*A/P^2)

4. Длина зерна

5. Ширина зерна

6. Коэффициент асимметрии

7. Длина бороздки зерна

Целевая переменная:

Сорт зерна

1. Cama

2. Rosa

3. Canadian

### 2. Результаты предварительного анализа и визуализации исходных признаков и целевой переменной

Все признаки представлены вещественными переменными.

Классы: три дискретных класса. Тип данных Object в признаках Compactness и Kernel Width наводит на мысль о пропущенных данных. Некоторые данные в признаках Compactness и Kernel Width заменены на знаки вопроса. Эти данные найдены и заменены.

Проведен корреляционный анализ признаков. Площадь, периметр, длина зерна и длина бороздки имеют прямую зависимость. Кроме того при построении boxplot отмечено, что площадь, периметр и длина зерна лучше всего делят данные на классы.

### 3. Подготовка данных. Обучение и тестирование моделей

Проведена подготовка данных, обучение и тестирование моделей.

Лучший критерий разбиения - энтропия.

Количество признаков (max features) - лучший результат при использовании всех семи признаков.

Оптимальная глубина дерева - 7.

К отчету приложен файл Lab1.ipynb с детальной визуализацией вышесказанного.

Проведено сравнение метода дерева решений с методом логистической регрессии.

### 4. Выводы

Проведен анализ задачи классификации и решение ее с использованием метода машинного обучения деревья решений. Входные данные (признаки) полностью числовые с минимальными пропусками в данных (не более 2%). При первичном анализе выделены признаки, которые оптимально делили данные на классы (площадь, периметр и длина зерна). Проведено пробное построение деревьев решений с использованием различных параметров. Лучший результат показало дерево, в котором использованы следующие параметры: критерий - энтропия, максимальной число признаков - 7, максимальная глубина дерева - 7. При выбранных параметрах достигнута точность классификации - 0.914. Кроме деревьев решений для сравнения использован метод решения задач классификации: логистическая регрессия. Этот метод был выбран с учетом картины входящих данных (небольшое количество признаков и тип признаков - вещественные). Со стандартными методами оптимизации и регуляризации достигнута высокая точность классификации - 0.957. Значение model.score больше чем cros\_val\_score.mean, потому что несмотря на использование кросс-валидации, обучение модели (fit) в итоге дает нам только лучшие из полученных весов. Метод model.score считает среднее значение model.predict(X) по отношению к y и значит находит score при использовании лучшего веса. cros\_val\_score же считает среднее всех score полученных при кросс-валидации.