ЛАБОРАТОРНАЯ РАБОТА 1. ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

**Исследование набора данных**

**Цели и задачи**

Цель лабораторной работы: изучение программных средств для организации рабочего места специалиста по интеллектуальному анализу данных.

Основные задачи:

* установка и настройка среды разработки Python;
* изучение принципов загрузки и очистки данных;
* получение навыков по предварительной обработке данных на языке

Python;

* изучение основных библиотек Python для работы с данными.

**Оборудование и материалы**

Для выполнения лабораторной работы рекомендуется использовать персональный компьютер со следующими программными средствами разработки (выбрать один или несколько програмных продуктов для практической реализации задач лабораторной работы): MS Visual Studio 2013 и выше; среда разработки Java, интерпретатор Python (Google Colab).

**Методика и порядок выполнения работы**

Перед выполнением индивидуального задания рекомендуется выполнить все пункты учебной задачи.

**Постановка задачи.**

Необходимо организовать подготовку данных для построения модели (допустим модели классификации). В качестве данных выбран набор данных об ирисах Фишера. Это, пожалуй, самый известный набор данных, с которого многие начинают исследование алгоритмов интеллектуального анализа данных.

Указанный набор данных предназначен для построения модели классификации. Данные о 150 экземплярах ириса (рис. 1), по 50 экземпляров из трёх видов – Ирис щетинистый (Iris setosa), Ирис виргинский (Iris virginica) и Ирис разноцветный (Iris versicolor). Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):

1. длина наружной доли околоцветника (sepal length);
2. ширина наружной доли околоцветника (sepal width);
3. длина внутренней доли околоцветника (petal length);

4) ширина внутренней доли околоцветника (petal width).

На основании этого набора данных требуется построить правило классификации, определяющее вид растения по данным измерений. Это задача многоклассовой классификации, так как имеется три класса – три вида ириса.



а) б) в)

Рисунок 1– Внешний вид классифицируемых ирисов: а) Iris setosa; б) Iris virginica; в) Iris versicolor

1. Необходимо скачать набор данных из репозитория Center for

Machine Learning and Intelligent Systems (необходим только один текстовый файл с данными измерений): [http://archive.ics.uci.edu/ml/datasets/Iris.](http://archive.ics.uci.edu/ml/datasets/Iris)

Файл Iris.data при просмотре выглядит следующим образом (рис. 2):

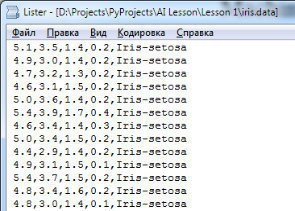


Рисунок 2 – Внешний вид данных файла Iris.data

1. Использовать текстовые редакторы для просмотра и анализа данных из определенных наборов – нерациональный вариант. Поэтому запустим Jupyter Notebook и начнем работать с загруженным набором с использованием среды Python. Используем метод genfromtxt() из пакета scipy.

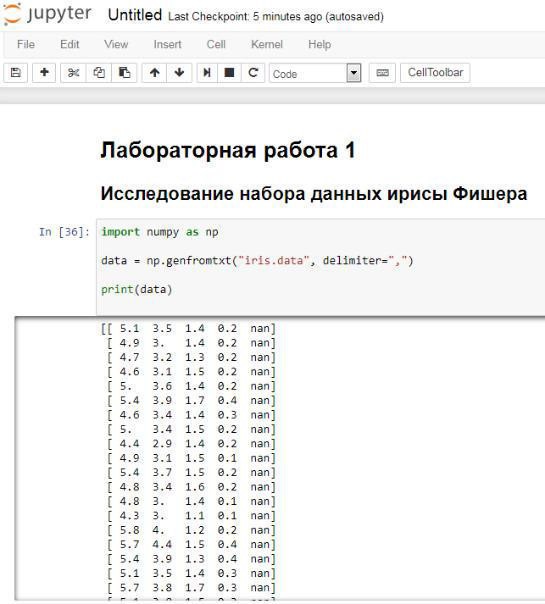


Рисунок 3 – Загрузка данных файла Iris.data

Метод genfromtxt() возвращает массив numpy (тип numpy.ndarray). Следует обратить внимание, что пятый столбец содержит неопределенные значения numpy.NaN (объясните – почему?).

1. Производить вывод всего источника данных – нерациональный путь. В реальных задачах данных может оказаться слишком много, поэтому чаще всего используют подвыборку данных для поверхностного обзора исследуемой обучающей выборки (рис. 4).

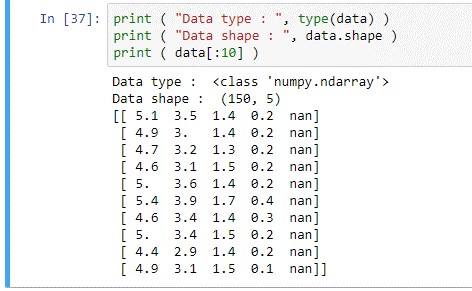


Рисунок 4 – Начальное исследование Iris.data

Из представленного фрагмента видно, что data – это двумерный массив размером 150x5, или можно сказать, что это одномерный массив, каждый элемент которого также одномерный массив размером 5 элементов.

1. В рамках данной задачи необходимо все-таки получить значения пятого столбца. Для этого желательно использовать другой подход (рис. 5):

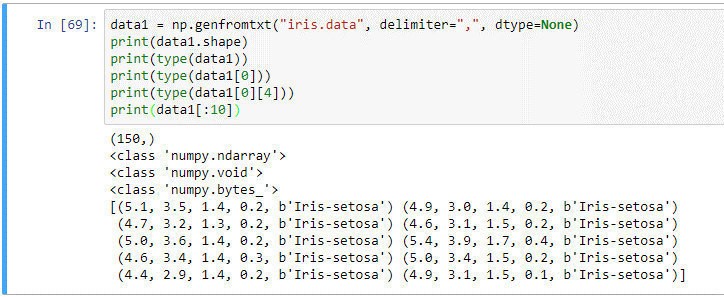


Рисунок 5 – Загрузка данных разного типа в массив

Сразу же желательно (на первоначальных этапах изучения Python) проводить анализ типов различных значений. На рис. 6 представлен еще один вариант загрузки данных в массив numpy.ndarray.

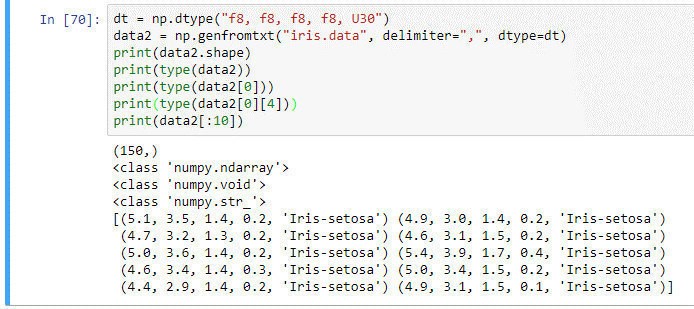


Рисунок 6 – Загрузка даных в массив с типом, определяемым пользователем

Поясните различие в структурах данных, получаемых с использованием представленных листингов.

1. Было загружено 150 элементов данных, но даже при такой маленькой выборке невозможно что-либо сказать о наборе данных. Для получения дополнительной информации необходимо визуализировать загруженные данные. В нашем случае каждый элемент данных представлен вещественными признаками – это существенно упрощает визуализацию (рис. 7). Но сложность заключается в том, что приходится работать с элементами 4-мерного пространства, поэтому строится не графическое представление распределения, а отдельные проекции.
2. Уже из графического распределения на рис. 7 видно, что тип ирисов Setosa хорошо отделяется. На данном графике представлено отображение в плоскости признаков (‘Sepal Width’, ‘Sepal Length’) Но исследователь имеет возможность построить столько графиков, сколько необходимо для глубокого анализа данных. Изменим ячейку In[72] в соответствии с листингом, представленным на рис. 8.

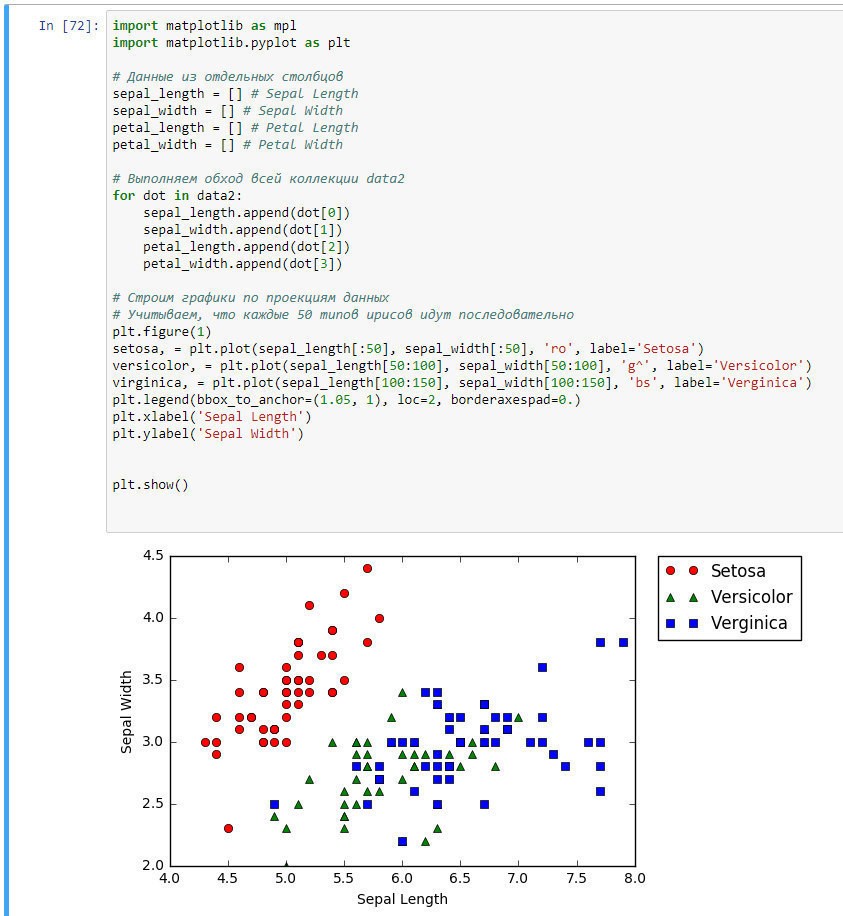


Рисунок 7 – Простейший анализ данных по графическому представлению



Рисунок 8 – Построение простейшего графика для отображения различных проекций данных

Вывод данного кода представлен на рис. 9.

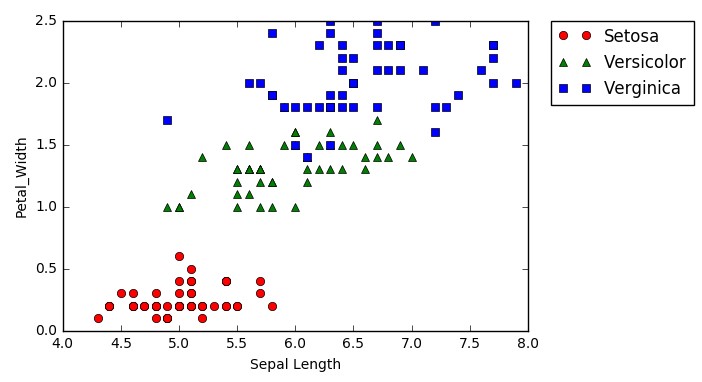
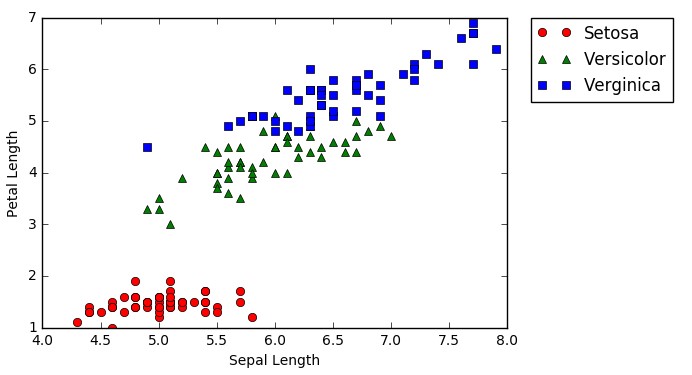
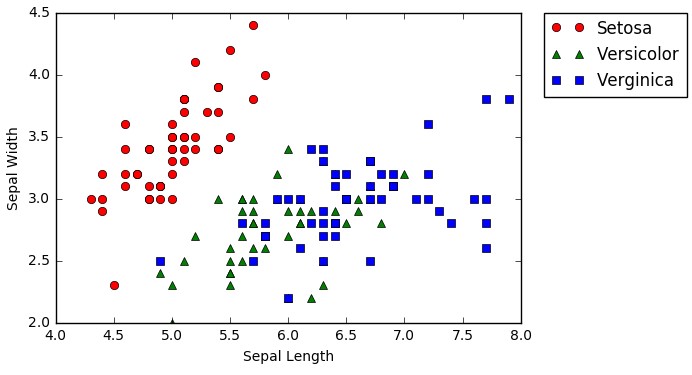


Рисунок 9 – Вывод графика для отображения различных проекций данных

Из графиков 9 уже хорошо видно, что множество Setosa хорощо отделимо, а множества Versicolor и Verginica представляют собой множества, разделение которых является непростой задачей.

Постройте другие проекции исходных данных. Сколько всего различных проекций можно построить для данного набора данных?

**Важные замечания**

1. Несмотря на кажущуюся простоту и «понятность» данных в результате визуализации, исследователь не должен делать поспешных выводов (например, было бы ошибочно делать вывод по рис. 9 о том, что ирисы Setosa те, у которых petal width менее 0,75). Следует помнить, что цель первичного исследования данных – получение представления о структуре и природе данных, а не построение модели предсказания, классификации и т.п.

1. В качестве среды разработки используйте языки программирования Python, Java или C#. По согласованию с преподавателем студент может самостоятельно может выбрать язык программирования и среду разработки (при этом студенту необходимо критически обосновать свой выбор).
2. При выборе набора данных (data set) на ресурсах [9, 10] необходимо согласовать свой выбор с другими студентами группы и преподавателем, так как работа над одинаковыми наборами данных недопустима.
3. В рамках данного лабораторного курса рекомендуется использовать инструментарий Python (библиотеки, среду разработки) для решения поставленных задач.

**Индивидуальное задание**

1. Подберите набор данных на ресурсах [9, 10] и согласуйте свой выбор с преподавателем. Студент может предложить набор данных в соответствии с тематикой исследования.
2. Проведите первичный анализ данных. В результате анализа данных студент должен предоставить следующую информацию о наборе данных:
   1. Описание набора данных, пояснения, позволяющие лучше понять природу данных. Назначение набора данных и возможные модели, которые можно построить на основе данного набора данных (практические задачи, решаемые с использованием данного обучающего набора данных). Описание каждого признака и его тип.
   2. Форма набора данных: количество элементов набора, количество признаков, количество пропущенных значений, среднее значение отдельных признаков, максимальные и минимальные значения отдельных признаков и прочие показатели. Предположения, которые можно сделать, проведя первичный анализ.
   3. Графические представления, позволяющие судить о неоднородности исследуемого набора данных. Построение графиков желательно произвести по нескольким проекциям.

**Содержание отчета и его форма**

Отчет по лабораторной работе должен содержать:

1. Номер и название лабораторной работы; задачи лабораторной работы.
2. Реализация каждого пункта подраздела «Индивидуальное задание» с приведением исходного кода программы, диаграмм и графиков для визуализации данных.
3. Ответы на контрольные вопросы.
4. Экранные формы (консольный вывод) и листинг программного кода с комментариями, показывающие порядок выполнения лабораторной работы, и результаты, полученные в ходе её выполнения.

Отчет о выполнении лабораторной работы сдается преподавателю.

**Контрольные вопросы**

1. Какие инструментальные средства используются для организации рабочего места специалиста Data Science?
2. Какие библиотеки Python используются для работы в области интеллектуального анализа данных? Дайте краткую характеристику каждой библиотеке.
3. Почему при реализации систем интеллектуального анализа данных широкое распространение получили библиотеки Python?

**Список литературы**

Для выполнения лабораторной работы, при подготовке к защите, а также для ответа на контрольные вопросы рекомендуется использовать следующие источники.

**Список основной литературы**

1. Уэс, Маккинли. Python и анализ данных Электронный ресурс /

Маккинли Уэс ; пер. А. А. Слинкин. - Python и анализ данных,2021-04-19. - Саратов : Профобразование, 2017. - 482 с. - Книга находится в премиум-версии

ЭБС IPR BOOKS. - ISBN 978-5-4488-0046-7, экземпляров неограниченно.

1. Сузи, Р.А. Язык программирования Python Электронный ресурс :

учебное пособие / Р.А. Сузи. - Язык программирования Python,2020-07-28. -

Москва : Интернет-Университет Информационных Технологий (ИНТУИТ),

2016. - 350 c. - Книга находится в базовой версии ЭБС IPRbooks. - ISBN 5-

9556-0058-2, экземпляров неограниченно

**Список дополнительной литературы**

3. Стенли, Липпман. Язык программирования С++ Электронный ресурс : Полное руководство / Липпман Стенли, Лажойе Жози ; пер. А. Слинкин. - Язык программирования С++,2021-04-19. - Саратов : Профобразование, 2017.

- 1104 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-

4488-0136-5, экземпляров неограниченно

1. Седжвик, Р. Алгоритмы на С++ / Р. Седжвик. - 2-е изд., испр. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 1773 с., экземпляров неограниченно
2. <https://archive.ics.uci.edu/ml/index.html>– Репозиторий наборов данных для машинного обучения (Центр машинного обучения и интеллектуальных систем).
3. [https://www.kaggle.com](https://www.kaggle.com/) – Портал и система проведения соревнований по проблемам анализа данных.
4. [https://www.mockaroo.com](https://www.mockaroo.com/) – Сайт для генерации наборов данных.