**Технологии анализа данных.
Методические указания к практическим занятиям**

**Содержание**

[Основы работы с пакетом KNIME 2](#_Toc414527283)

[Лабораторная работа № 1. Поиск шаблонов 4](#_Toc414527284)

[Лабораторная работа № 2. Деревья решений 6](#_Toc414527285)

[Лабораторная работа № 3. Кластеризация 8](#_Toc414527286)

# Основы работы с пакетом KNIME

KNIME представляет собой свободно распространяемый прикладной программный пакет [<http://www.knime.org/>] с графическим интерфейсом, поддерживающий цикл интеллектуального анализа данных (доступ к данным различных форматов, трансформация данных, аналитические функции, визуализация и подготовка отчетов).

Идеологической основой KNIME является понятие потока работ (workflow). Поток работ графически изображает процесс преобразования исходных данных в результаты (см. Рис. 1). Изображение состоит из узлов (прямоугольников) и стрелок. Узел инкапсулирует некоторую операцию над данными, стрелки показывают путь данных.



**Рис. 1. Простой поток работ**

С помощью руководства [Начало работы в KNIME](http://tech.knime.org/getting-started) выполните пошаговое построение простого потока работ и настройку его узлов.

Изучите краткое [Руководство пользователя KNIME](http://tech.knime.org/workbench). Проверьте себя на понимание основных терминов KNIME: поток работ (workflow), рабочее пространство (workspace), узел (node), порт узла (port), статус узла (node status), соединение и настройка узла (connecting and configuring node).

## Полезные ссылки

* Официальный сайт разработчиков KNIME: <http://www.knime.org/>
* Документация KNIME
* [Начало работы в KNIME](http://tech.knime.org/node/106/) (на англ. языке)
* [Демонстрационные примерs и видео](http://www.knime.org/node/147/) (на англ. языке)
* [Дистрибутивы для установки](http://www.knime.org/downloads/) (на англ. языке)

# Лабораторная работа № . Поиск шаблонов

**Количество баллов: 5**.

*Цель*. Построение потока работ, выполняющего решение задачи анализа рыночной корзины и поиска ассоциативных правил. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, преобразовать загруженные данные в специализированный тип данных пакета KNIME, найти частые наборы и ассоциативные правила, вывести результаты.

1. Создайте поток работ, приведеный на Рис. 2.



Рис. . Поток работ для решения задачи анализа рыночной корзины

1. Выполните настройку узла “File Reader” так, чтобы:
* в качестве файла исходных данных фигурировал baskets.csv (см. архив [baskets.zip](https://elearning.sp.susu.ac.ru/mod/resource/view.php?id=305), прилагаемый к заданию);
* первая строка файла трактовалась как содержащая названия столбцов (“Read column headers”);
* в качестве разделителя столбцов фигурировала запятая;
* узел обрабатывал неполные строки (кнопка “Advanced”, вкладка “Short lines”).
1. Выполните настройку узла “Create Collection Column” так, чтобы:
* все строки исходного файла попали в выходную коллекцию данных;
* установите флаги “Create collection of type set”, “Ignore missing values”, “Remove aggregated columns from table”.
1. Выполните поток работ, предварительно настроив узел “Association Rule Learner”, указав различные значения поддержки (minimum support). Объясните полученные результаты (как данный параметр влияет на решение задачи?).
2. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

# Лабораторная работа № 2. Деревья решений

**Количество баллов: 5.**

*Цель.* Построение потока работ, выполняющего решение задачи классификации посредством построения дерева решений. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

1. Создайте поток работ, приведеный на Рис. 2.



Рис. . Поток работ для решения задачи классификации

1. Выполните настройку узла “File Reader” так, чтобы:
* в качестве файла исходных данных фигурировал [marks.csv](https://elearning.sp.susu.ac.ru/mod/resource/view.php?id=305);
* первая строка файла трактовалась как содержащая названия столбцов (“Read column headers”);
1. Выполните настройку узла “Decision Tree Learner” так, чтобы поле FINALMARK трактовалось как признак класса.
2. Выполните поток работ. Убедитесь, что построенное дерево решений показывает зависимость итоговой оценки только от атрибутов, отражающих мнение учителя (имеющих название вида TEACHER\_xx), и ее независимость от атрибутов, отражающих персональные данные ученика (имеющих название вида PUPIL\_xx).
3. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

**Факультативное задание (1 балл)**

1. Измените созданный поток работ, как показано на Рис. 4.



Рис. . Поток работ для решения задачи классификации
с дополнительными узлами

1. Настройте узел “Color Manager”, указав столбец FINALMARK как раскрашенный. Выполните поток работ. Сравните результаты при различных настройках узла “Decision Tree Learner” (контекстное меню узла): “Decision Tree View” и “Decision Tree View (simple)”.
2. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

# Лабораторная работа № 3. Кластеризация

**Количество баллов: 5.**

*Цель.* Построение потока работ, выполняющего решение задачи кластеризации. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

1. Создайте поток работ, приведеный на Рис. 5.



Рис. . Поток работ для решения задачи кластеризации

1. Выполните настройку узла “File Reader” так, чтобы в качестве файла исходных данных фигурировал [basketball.csv](https://elearning.sp.susu.ac.ru/mod/resource/view.php?id=309).
2. Выполните поток работ. Сравните результаты при различных параметрах: количество кластеров, цвета для отображения кластеризуемых объектов.
3. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.