

## Программа учебной дисциплины «Анализ паттернов данных»

Утверждена

Академическим советом ООП

Протокол №17 от «28» мая 2018г.

Автор	Мячин А.Л. <a href="mailto:amyachin@hse.ru">amyachin@hse.ru</a>
Число кредитов	5
Контактная работа (час.)	36
Самостоятельная работа (час.)	154
Курс	1-й курс
Формат изучения дисциплины	без использования онлайн курса

### I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения дисциплины «Анализ паттернов данных» является освоение студентами основ теории кластерного анализа, анализа латентных классов и методов анализа паттернов на примерах прикладных задач с использованием языка Python для поиска закономерностей среди большого количества разнородных данных.

### II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

#### **Тема 1. Введение в анализ данных.**

Статистика, машинное обучение и Data Mining. Многомерное шкалирование, матричные диаграммы рассеяния и система параллельных координат. Анализ данных с использованием языка Python.

#### **Тема 2. Методы классификации.**

Основные задачи и процесс классификации. Линейные модели. Масштабирование признаков и переход в новые признаковые пространства. Практическое применение методов классификации с использованием языка Python.

Деревья решений. Алгоритмы конструирования. Процесс построения композиций деревьев решений. Практическое применение с использованием языка Python.

Анализ латентных классов. Основные свойства наивной классификации. Практическое применение метода анализа латентных классов для обработки результатов анкетирования с использованием языка Python.

#### **Тема 3. Кластерный анализ.**

Иерархические и итеративные методы. Меры сходства. Алгоритм k-средних. Проверка качества кластеризации.

#### **Тема 4. Методы анализа паттернов.**

Основные определения. Методы анализа паттернов, основанные на парном сравнении показателей. Методы анализа паттернов, результат которых не зависит от выбора исходной

последовательности показателей. Использование методов анализа паттернов для исследования электорального поведения. Решение прикладных задач с использованием методов анализа паттернов на базе языка Python.

### III. ОЦЕНИВАНИЕ

Накопленная оценка за четвертый модуль выставляется по следующей формуле:

$$O_{\text{накопл.4}} = 0,5 \cdot O_{\text{д.з.1}} + 0,5 \cdot O_{\text{д.з.2}},$$

где  $O_{\text{д.з.1}}$  – результат выполнения первого домашнего задания в четвертом модуле;  $O_{\text{д.з.2}}$  – результат выполнения второго домашнего задания в четвертом модуле. Оценка  $O_{\text{д.з.1}}$  выставляется в 10-балльной системе (без округлений). Оценка  $O_{\text{д.з.2}}$  выставляется в 10-балльной системе (без округлений). Оценка  $O_{\text{накопл.4}}$  выставляется в 10-балльной системе и округляется до целого числа по правилам арифметики округления.

Итоговая оценка за курс выставляется по следующей формуле:

$$O_{\text{итог}} = 0,7 \cdot O_{\text{экс.}} + 0,3 \cdot O_{\text{накопл.}},$$

где  $O_{\text{экс.}}$  – оценка, полученная студентом за экзаменационную работу. Оценка  $O_{\text{экс.}}$  выставляется в 100-балльной системе, а затем переводятся в 10-балльную систему (округляется до целого числа по правилам арифметики округления). Способ округления итоговой оценки производится по правилам арифметики округления.

Данная программа не предусматривает возможность пересдачи неудовлетворительных оценок, полученных за любую из форм текущего контроля, а также возможность компенсировать оценки за домашние контрольные работы, не полученные вследствие несвоевременной сдачи работы или пропуска семинарского занятия по любой причине. В этом случае за соответствующую форму текущего контроля студенту выставляется 0 баллов. Перевод в 5-балльную шкалу осуществляется по правилу:

Оценка по 10-балльной шкале	Оценка по 5-балльной шкале
0	неудовлетворительно
1	
2	
3	
4	удовлетворительно
5	
6	хорошо
7	
8	отлично
9	
10	

### IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

#### Примеры заданий итогового контроля

1. В файле Data.xlsx находится база данных с гипотетическими результатами выборов в парламент. Написать программу на языке Python, позволяющую загрузить

данные, проверить наличие выбросов в данных, а также найти закономерности используя методы анализа паттернов, основанные на парном сравнении показателей. Описать полученные закономерности.

2. В файле Data.docx находятся результаты социологического опроса. Написать программу на языке Python, позволяющую провести разбиение исходной выборки респондентов с использованием метода анализа латентных классов. Интерпретировать полученные результаты.

## V. РЕСУРСЫ

### 1. Основная литература

1. Миркин Б.Г. Введение в анализ данных: учебник и практикум для ВУЗов. М.: Юрайт, 2015.

2. Чашкин Ю. Р. Математическая статистика: Анализ и обработка данных: учебное пособие для студентов высших учебных заведений. Ростов н/Д: Феникс, 2010.

### 2. Дополнительная литература

1. Толстова Ю.Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками: Учеб. пособие для вузов. М.: Научный мир, 2000.

2. Мячин А. Л. Анализ паттернов: диффузионно-инвариантная паттерн-кластеризация // Проблемы управления. 2016. № 4. С. 2-9. URL: <https://publications.hse.ru/articles/194910575>

### 3. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	<i>Из внутренней сети университета (договор)</i>
2.	Microsoft Office Professional Plus 2010	<i>Из внутренней сети университета (договор)</i>

### 4. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы)

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.