

Программа учебной дисциплины «Анализ данных в HR и социальных исследованиях»

Утверждена

Академическим советом ООП

Протокол № 2.6-06.7/5 от 06.07.2018

Автор	Дегтярёв Павел Александрович
Число кредитов	3
Контактная работа (час.)	30
Самостоятельная работа (час.)	84
Курс	4
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

1. Цели освоения дисциплины

Целями освоения дисциплины «Анализ данных в HR и социальных исследованиях» являются:

- формирование представления о круге профессиональных и академических задач, связанных с обработкой данных в HR и социальных науках;
- формирование представления о современных методах анализа данных;
- формирование навыков использования языка R для решения задач по анализу данных;
- формирование навыков решения ключевых задач в сфере анализа данных в HR и социальных науках: обработка сырых данных, поиск внутренних взаимосвязей в структуре данных и снижение размерности, проверка гипотез о взаимосвязях, обработка естественных языков;
- формирование навыков представления результатов анализа данных.

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

Уметь формировать план проведения исследования данных, включая постановку исследовательской проблемы и задач

Иметь навыки программирования в среде R, включая задачи построения циклов и написания функций

Уметь интегрировать данные в наиболее распространённых форматах в программную среду R

Уметь проводить обработку сырых данных (выделение субвыборок, фильтрация, импутация, преобразование переменных при помощи математических и логических вычислительных операций и т.д.) в соответствии с принципами tidyverse (пакеты dplyr, tidyr, tibble)

Уметь решать задачи поиска внутренних взаимосвязей и сокращения размерности при помощи методов эксплораторного факторного и кластерного анализа

Уметь решать задачи проверки гипотез о взаимосвязи при помощи методов корреляции, различных форм регрессии и построения деревьев решений

Уметь решать задачи обработки естественных языков в рамках парадигм bag-of-words и word2vec

Уметь представлять результаты анализа данных при помощи графических средств пакета ggplot2

3. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к общепрофессиональным дисциплинам направления и является обязательной.

Изучение данной дисциплины базируется на следующих дисциплинах:

1. Математические и статистические методы в психологии
2. Психологический практикум (Раздел «Измерения»);
3. Подбор и оценка персонала;
4. Психологический практикум: Психодиагностика;
5. Индустриальная и организационная психология

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

1. Знания особенностей организации и проведения психодиагностических исследований;
2. Знание специфики психологического измерения;
3. Знания основных типов личности, темперамента, характера, концепций и моделей интеллекта.
4. Знания основ математической статистики.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

1. Обучение и развитие персонала
2. Курсовая работа
3. Выпускная квалификационная работа

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Обработка сырых данных. Принципы tidyverse. Загрузка данных в программную среду R; выделение субвыборок; фильтрация; импутация; преобразование переменных при помощи математических и логических вычислительных операций.

Тема 2. Поиск внутренних взаимосвязей в структуре данных и снижение размерности. Поиск внутренних взаимосвязей в структуре данных; эксплораторный факторный анализ; иерархический кластерный анализ; кластерный анализ k-средних; кластеризация нечётких множеств; partitioning around medoids; оценка ширины силуэтов; обоснование конструктивной валидности шкал психологических измерительных методик; оценка надёжности согласованности результатов психологических измерений.

Тема 3. Проверка гипотез о взаимосвязи. Корреляция; линейная регрессия; биномиальная регрессия; мультилинейная регрессия; мультиномиальная регрессия; ординальная регрессия; регуляризация; кросс-валидизация; random forest.

Тема 4: Обработка естественных языков Обработка естественных языков; контент-анализ; семантический анализ; bag-of-words; word2vec.

Тема 5 Представление результатов анализа данных

R Markdown; представление результатов анализа данных в графической форме; ggplot2.

III. ОЦЕНИВАНИЕ

Преподаватель оценивает работу студентов на практических занятиях: активность на практических занятиях, участие в активных формах работы, дискуссиях, работу по разбору кейсов. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за работу на семинарских занятиях определяется перед промежуточным или итоговым контролем – О аудиторная.

Накопленная оценка за текущий контроль учитывает результаты студента по текущему контролю следующим образом:

$O_{\text{накопленная}} = 0,4 * O_{\text{домашнее задание}} + 0,6 * O_{\text{аудиторная}}$

$O_{\text{аудиторная}} = 0,5 * O_{\text{посещение занятий}} + 0,5 * O_{\text{активность на занятии}}$

Результатирующая оценка за дисциплину рассчитывается следующим образом:

$O_{\text{рез}} = 0,4 * O_{\text{экзамен}} + 0,6 * O_{\text{накопленная}}$

Способ округления оценки итогового контроля в форме экзамена: арифметический (например, оценка 4,4 округляется до 4, а оценка 4,5 до 5). Студент не имеет возможности пересдать низкие результаты за текущий контроль или работу на занятиях, самостоятельную работу, если за любой из этих видов работы была выставлена оценка. В случае если студент пропустил срок сдачи отчетности по текущему контролю или самостоятельной работе по уважительной причине (к уважительным причинам может быть отнесена болезнь в указанный период, смерть близкого родственника, повестка в военкомат или нахождение под стражей, подтвержденные соответствующими справками) допускается сдача форм контроля не позднее чем через две недели с момента окончания действия справки, если другие сроки не установлены отдельно. В случае, если студент не сдает работу в указанный срок, ему выставляется оценка «0».

Условия пересдач На пересдаче студенту не предоставляется возможность получить дополнительный балл для компенсации оценки за текущий контроль. Первая пересдача проводится преподавателем, отвечающим за чтение дисциплины на факультете. При выставлении результирующей оценки учитывается накопленная оценка за текущий контроль и аудиторную работу $O_{\text{рез}} = 0,4 * O_{\text{экзамен}} + 0,6 * O_{\text{накопленная}}$ $O_{\text{накопленная}} = 0,4 * O_{\text{домашнее задание}} + 0,6 * O_{\text{аудиторная}}$ Вторая пересдача проводится в присутствии комиссии, включающей не менее трех преподавателей, при выставлении результирующей оценки учитывается накопленная оценка за текущий контроль, самостоятельную и аудиторную работу $O_{\text{рез}} = 0,4 * O_{\text{экзамен}} + 0,6 * O_{\text{накопленная}}$

$O_{\text{накопленная}} = 0,4 * O_{\text{домашнее задание}} + 0,6 * O_{\text{аудиторная}}$ Способ округления результирующей оценки по учебной дисциплине: арифметический (например, оценка 4,4 округляется до 4, а оценка 4,5 до 5). Накопленная оценка может быть аннулирована только по решению комиссии в исключительных случаях (студент пропустил по уважительной причине более 70% занятий). В этом случае результирующая оценка рассчитывается по формуле: $O_{\text{рез}} = 0,4 * O_{\text{домашнее задание}} + 0,6 * O_{\text{экзамен}}$

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примеры вопросов, задач/кейсов, заданий для экзамена:

1. Предложите механизм вычисления итогового балла по опросу исходя из шкал, которые Вы получили в одном из решений в факторном анализе
2. Предложите корректное факторное решение для выбранной базы данных, обоснуйте факторное решение.

VI. РЕСУРСЫ

5.1 Основная литература

1. Garrett Grolemond, Hadley Wickham. R for Data Science. <http://r4ds.had.co.nz>
2. The Elements of Statistical Learning (2nd edition). Hastie, Tibshirani and Friedman (2009). Springer-Verlag. 763 pages.
https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf
3. Package 'cluster'.
<https://cran.r-project.org/web/packages/cluster/cluster.pdf>
4. Package 'psych'. <https://cran.r-project.org/web/packages/psych/psych.pdf>

5.2 Дополнительная литература

1. Garrett Grolemond, Hadley Wickham. R for Data Science. <http://r4ds.had.co.nz>
2. Package 'dplyr'. <https://cran.r-project.org/web/packages/dplyr/dplyr.pdf>
3. Package 'tidyr'. <https://cran.r-project.org/web/packages/tidyr/tidyr.pdf>
4. Package 'tibble'. <https://cran.r-project.org/web/packages/tibble/tibble.pdf>

5.3 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>

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
<i>Профессиональные базы данных, информационно-справочные системы</i>		
1.	Электронно-библиотечная система Юрайт	<i>Из внутренней сети университета (договор)</i> URL: https://biblio-online.ru/

5.5 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 лабораторных и самостоятельных занятий по дисциплине оснащены микрофоном и звукоусиливающей аппаратурой, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.