**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Data Science: основы обработки и анализа данных

Data Science: Fundamentals of Data Processing and Analysis

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 056375

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Целью учебных занятий по дисциплине является формирование у обучающихся знаний о системе математико-статистических методов, использующихся при анализе данных, полученных в результате практических исследований в различных областях науки, и навыков работы в различных исследовательских ситуациях.

Задачами учебных занятий по дисциплине являются:

• сформировать у обучающихся систему знаний об основах измерения и количественного описания данных в различных областях знаний (психология, социология, эпидемиология);

• сформировать умения эффективно пользоваться методами статистического вывода и многомерными методами и моделями в зависимости от исследовательской ситуации;

• сформировать пользовательские умения в компьютерной программе для статистического анализа данных SPSS либо приложении MS Excel.

Предложенная программа курса характеризуется акцентом на практическом освоении методов и приемов экспериментального исследования. Таким образом, результатом учебных занятий по дисциплине должно стать целостное представление о принципах и способах планирования, реализации и анализа результатов экспериментального исследования в различных областях науки и практики. Данный курс взаимосвязан математической статистикой, психологией, социологией.

**1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)**

Программа курса рассчитана на первоначальное знакомство с принципами проведения экспериментального исследования в гуманитарных науках, однако содержит элементы статистического и теоретико-вероятностного вывода. Поэтому для максимально эффективного усвоения материала дисциплины обучающийся

• должен владеть базовыми математическими понятиями, достаточными для работы с формулировками математических утверждений;

• должен иметь навыки работы с компьютерными программами.

**1.3. Перечень результатов обучения (learning outcomes)**

В результате освоения дисциплины обучающиеся должны:

• знать основные теоретико-методологические принципы экспериментального исследования;

• обладать навыками планирования и проведения экспериментального исследования, количественного и качественного анализа результатов эксперимента и их интерпретации;

• уметь обоснованно использовать статистический вывод для обработки результатов исследования;

• обладать навыками количественного анализа данных при помощи компьютерной программы для статистического анализа данных SPSS либо приложения MS Excel

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Аудиторная учебная работа: лекции – 48 часов, лабораторные работы – 16 часов. В качестве активных и интерактивных форм учебных занятий предполагается: тренинговые упражнения по применению изученных методов на предложенных данных, самостоятельный выбор данных из предложенных публичных ресурсов и применение изученных методов на выбранной базе данных, представление результатов обработки данных и интерактивное обсуждение полученных результатов, защита проекта по обработке данных.

**Раздел 2. Организация, структура и содержание учебных занятий**

**2.1. Организация учебных занятий**

**2.1.1 Основной курс**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся | | | | | | | | | | | | | | | | | | |
| Код модуля в составе дисциплины,  практики и т.п. | Контактная работа обучающихся с преподавателем | | | | | | | | | | | | Самостоятельная работа | | | | Объём активных и интерактивных  форм учебных занятий | Трудоёмкость |
| лекции | семинары | консультации | практические  занятия | лабораторные работы | контрольные работы | коллоквиумы | текущий контроль | промежуточная  аттестация | итоговая аттестация | под руководством преподавателя | в присутствии  преподавателя | сам. раб. с использованием  методических материалов | текущий контроль (сам.раб.) | промежуточная аттестация (сам.раб.) | итоговая аттестация  (сам.раб.) |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 5 | 48 |  | 2 | 16 |  |  |  |  | 2 |  |  |  | 4 |  | 36 |  | 19 | 3 |
|  | 2-100 |  | 2-100 | 10-25 |  |  |  |  | 2-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 48 |  | 2 | 16 |  |  |  |  | 2 |  |  |  | 4 |  | 36 |  |  | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 5 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

Модуль М1. Основные теоретико-методологические принципы экспериментального исследования.

1.1. Интеграция социологического (психологического) и математического знания.

1.1.1. Возможность и необходимость применения математических методов в социологии и психологии.

1.1.2. Основные направления использования математики в психологии и социологии (обеспечение объективности получаемых знаний, измерение и обобщение результатов измерений, статистическая проверка гипотез)

1.1.3. Методы математической статистики в теоретической и практической психологической и социологической работе

1.2. Генеральная совокупность и выборка.

1.2.1. Понятие генеральной совокупности и выборки.

1.2.2. Способы отбора репрезентативной выборки (рандомизированный и стратифицированный отбор).

1.2.3. Объем выборки и статистическая достоверность.

1.2.4. Зависимые и независимые выборки.

1.3. Измерения и шкалы.

1.3.1. Общее понятие о процедуре измерения в психологическом исследовании.

1.3.2. Измерительные шкалы (номинативная, ранговая, интервальная, отношений)

1.4. Представление и первичная обработка результатов исследования

1.4.1. Таблицы и графики распределения частот.

1.4.2. Таблицы сопряженности номинативных признаков (сравнение теоретического и эмпирического распределений, анализ таблиц сопряженности с числом градаций больше двух, анализ таблиц сопряженности для повторных измерений, анализ последовательностей)

1.4.3. Первичные описательные статистики (меры центральной тенденции, квантили распределения, меры изменчивости).

1.4.4. Нормальный закон распределения и его применение.

1.4.5. Частотный анализ и описательные статистики в SPSS и R

Модуль М2. Статистические основы проведения исследований.

2.1. Проблема статистического вывода.

2.1.1. Научные и статистические гипотезы.

2.1.2. Математическая модель проверки статистической гипотезы.

2.1.3. Уровень статистической значимости; статистическое решение и вероятность ошибки.

2.1.4. Содержательная интерпретация статистического решения

2.2. Корреляционный анализ.

2.2.1. Общее понятие о корреляционном анализе и коэффициенте корреляции.

2.2.2. Корреляция метрических переменных.

2.2.3. Частная корреляция.

2.2.4. Корреляция ранговых переменных.

2.3. Параметрические критерии сравнения выборок.

2.3.1. Математическая модель сравнения дисперсий.

2.3.2. Критерий t-Стьюдента для одной выборки.

2.3.3. Критерий t-Стьюдента для независимых выборок.

2.3.4. Критерийt-Стьюдента для зависимых выборок.

2.3.5. Вычисление t-критерия в SPSS и R

2.3.6. Критерий Манна-Уитни в SPSS и R

2.4. Дисперсионный анализ (ANOVA)

2.4.1. Назначение и общие понятия дисперсионного анализа (градации номинативного признака, межгрупповой и внутригрупповой фактор)

2.4.2. Виды дисперсионного анализа (однофакторный, многофакторный, с повторными измерениями, многомерный)

2.4.3. Математическая идея и основные допущения ANOVA

2.4.5. ANOVA в SPSS и R

2.5. Многомерные методы анализа данных.

2.5.1. Общее понятие многомерных методов анализа статистических данных

2.5.2. Факторный анализ

2.5.3. Кластерный анализ

2.5.4. Многомерное шкалирование

2.5.5. Использование SPSS и R при многомерном анализе статистических данных.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Основными видами занятий при изучении дисциплины «Основы обработки и интеллектуального анализа данных» являются: лекции, лабораторные работы, самостоятельная работа обучающихся.

Данный курс в значительной мере подразумевает на выходе уверенное владение основами обработки и интеллектуального анализа данных, которые закрепляются в виде выполнения лабораторных работ и в виде самостоятельной работы.

В рамках лекционного материала основное внимание уделяется изложению теоретических основ курса. Для закрепления теоретического материала все методы обработки и интеллектуального анализа данных сразу же применяются на имеющихся эмпирических данных.

Целью лабораторных работ является закрепление теоретических знаний, выработка навыков решения задач с использованием современных методов обработки и интеллектуального анализа данных. Для реализации этих занятий разрабатываются специальные задания, содержание которых отражает основные этапы и методику обработки и интеллектуального анализа данных и подбирается в зависимости от уровня изначальной подготовленности группы обучающихся.

**3.1.2 Методическое обеспечение самостоятельной работы**

Самостоятельная работа обучающегося, как вид деятельности, стимулирующий активность, самостоятельность, познавательный интерес с целью поиска необходимой информации, приобретения знаний, использования этих знаний для решения учебных, научных и профессиональных задач, представляет собой важную составляющую учебного процесса. Время, отводимое на самостоятельную работу, должно использоваться обучающимися для наиболее полного освоения учебной дисциплины. Следовательно, организация эффективной внеаудиторной самостоятельной работы в процессе обучения требует, с одной стороны, создание условий, призванных обеспечить рациональное и планомерное управление учебной деятельностью, протекающей в отсутствие преподавателя, и тщательной подготовки целого ряда учебных пособий, снабженных методическими указаниями, с другой стороны.

Роль преподавателя в организации самостоятельной работы состоит в координации действий обучающихся в освоении дисциплины, в методическом и организационном обеспечении учебного процесса. Взаимодействие между преподавателем и обучающимся осуществляется в форме консультаций, как очных, так и дистанционных с использованием современных социальных сетей для организации самостоятельной работы обучающихся. Преподаватели также оказывают помощь обучающимся по планированию и организации самостоятельной работы.

**3.1.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания**

Текущий контроль осуществляется в форме коротких опросов и тестов, углубленных вопросов по темам занятий, дополнительных вопросов, и т.д.

**3.1.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)**

Вариант 1. Отчет только по курсу «Основы обработки и интеллектуального анализа данных».

Для выполнения итогового проекта обучающимся выбираются данные для обработки.

Требования к данным.

Объем выборки более 200 элементов

1. Не менее бинарных 2-х переменных
2. Как минимум одна переменная, по которой можно разбить выборку на несколько (>3) групп
3. Не менее 3-х метрических (непрерывных) переменных

Источники данных.

1. [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/)
2. [CMU StatLib Datasets Archive](http://lib.stat.cmu.edu/datasets/)
3. [Airport, airline and route data](http://openflights.org/data.html)
4. [World Bank Data](http://data.worldbank.org/)
5. <http://r-dir.com/reference/datasets.html>
6. <http://getthedata.org/>
7. <http://www.umass.edu/statdata/statdata/>

По результатам обработки и интеллектуального анализа данных составляется отчет, в котором содержится интерпретация результатов следующих видов анализа:

1. описательные статистики
2. сравнение групп
3. корреляционный анализ
4. регрессионный анализ
5. кластерный анализ.

Система оценивания, таблица начисления баллов:

|  |  |  |
| --- | --- | --- |
| Задание | Баллы | Вид отчетности |
| Описание выбранных данных, описательные статистики, корреляционный анализ и результаты сравнения групп в SPSS | 8 | отчет |
| 8 | презентация |
| Описательные статистики и результаты сравнения групп в R | 8 | отчет |
| Результаты регрессионного анализа | 8 | отчет |
| Результаты кластерного анализа | 8 | отчет |
| Итоговый проект | 30 | отчет |
| презентация |
| Посещаемость (1 занятие = 1 балл) | 30 |  |
|  |  |  |
| ИТОГО: | 100 |  |

Перевод баллов в оценку;

От 60 до 74,9 – 3;

От 75 до 90 – 4;

От 90, 1 – 5.

Вариант 2. Комбинация двух дисциплин «Основы обработки и интеллектуального анализа данных» и «Разработка комплексов программ».

Итоговый проект

Программа с ГПИ, которая позволяет работать с выбранными данными, формировать заготовку статистического отчета определенной структуры и заготовку презентации. Экспорт части данных в БД.

Подробнее о функциональности:

1) интерфейс для выбора переменных для анализа

п.2-5, 8: сложный вариант с пользовательским выбором переменных, простой – с заранее выбранными переменными, в интерфейсе только выбор «включить в анализ»

2) описательные статистики

3) сравнение групп

4) корреляцию

5) регрессию

6) кластеризацию

7) автоматически сформировать по результатам анализа заготовку отчета в Word определенной структуры (таблички со значимыми различиями, характеристикики кластеров и т.д.) со свободными местами для текста

8) автоматически сформировать по результатам анализа заготовку презентации в LaTeX определенной структуры (таблички со значимыми различиями, хар-ки кластеров и т.д.) со свободными местами для текста

Статистические методы – сравнение групп (t-критерий, хи-квадрат), корреляция, регрессия, кластеризация

Технологии – R, R.Net, C#, автоматизация с Word (отчет) и Excel, БД, LaTeX (презентация)

Отчетность – работающая программа, заготовки отчета и презентации, готовый отчет, готовая презентация с выступлением

Возможное применение – регулярные отчеты по изменяющейся базе.

Требования к используемым данным

• Объем выборки более 200 элементов

• Не менее бинарных 2-х переменных

• Как минимум одна переменная, по которой можно разбить выборку на несколько (>3) групп

• Не менее 3-х метрических (непрерывных) переменных

где искать

1. [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/)
2. [CMU StatLib Datasets Archive](http://lib.stat.cmu.edu/datasets/)
3. [Airport, airline and route data](http://openflights.org/data.html)
4. [World Bank Data](http://data.worldbank.org/)
5. <http://r-dir.com/reference/datasets.html>
6. <http://getthedata.org/>
7. <http://www.umass.edu/statdata/statdata/>

Система оценивания, таблица начисления баллов:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Задание | Баллы | Вид отчетности |
| РКП, ООиИАД | Связь с БД из программы C# | 2 | программа |
| РКП, ООиИАД | Работа с Word и Excel из программы на C# | 4 | программа |
| ООиИАД, РКП | Описание выбранных данных, описательные статистики и результаты сравнения групп в SPSS | 5 | отчет |
| 10 | презентация |
| ООиИАД, РКП | Описательные статистики и результаты сравнения групп в R | 4 | отчет |
| ООиИАД | 6 | отчет |
| ООиИАД | Результаты кластерного анализа | 6 | отчет |
| ООиИАД, РКП | Использование функций R.Net | 1 | программа |
| РКП | Презентация по C# | 8 | презентация |
| РКП, ООиИАД | Чтение по C# | 2 | запись в обсуждении |
| РКП ,ООиИАД | чтение по LaTeX (3 задания по 1 баллу) | 3 | запись в обсуждении |
| РКП, ООиИАД | HW по ЛаТеХ - сведения о студентах | 1 | сообщение ВК |
| РКП, ООиИАД | HW по ЛаТеХ - установка ПО | 1 | сообщение ВК |
| РКП, ООиИАД | HW-02 по ЛаТеХ | 2 | Загрузка по инструкции |
| РКП, ООиИАД | HW-03 по ЛаТеХ | 2 | Загрузка по инструкции |
| РКП, ООиИАД | HW-04 по ЛаТеХ | 3 | Загрузка по инструкции |
| РКП | HW-05 по ЛаТеХ | 3 | Загрузка по инструкции |
| ООиИАД, РКП | Итоговый проект | 23 | программа |
| отчет |
| презентация |
| ООиИАД, РКП | Посещаемость (1 занятие = 1 балл) | 14 |  |
|  |  |  |  |
|  | ИТОГО: | 100 |  |

Перевод баллов в оценку;

От 60 до 74,9 – 3;

От 75 до 90 – 4;

От 90, 1 – 5.

Вариант 3. По выбору обучающегося возможна разработка индивидуальной образовательной траектории и сдача экзамена в устной форме по билетам, составленным из вопросов, перечисленных в п. 2.2 данной программы.

Для комбинированного курса индивидуальная программа выглядит следующим образом.

**Проект для индивидуального плана**

Программа с ГПИ, которая позволяет работать с выбранными данными, формировать заготовку статистического отчета определенной структуры и заготовку презентации. Экспорт части данных в БД.

Подробнее о функциональности:

1. загрузка произвольного набора данных (можно ограничить файлами определенного формата, например, .csv, .dat и т.д.)
2. интерфейс для выбора переменных для анализа (с пользовательским выбором переменных). Если переменная не подходит по типу для анализа, то возможны две реализации: а) выдать ошибку и попросить выбрать другие, б) давать первоначально выбор только из переменных, подходящих по типу
3. описательные статистики
4. сравнение групп
5. корреляция
6. регрессия
7. кластеризация
8. автоматическое формирование по результатам анализа заготовки отчета в Word определенной структуры (таблички со значимыми различиями, хар-ки кластеров и т.д.) со свободными местами для текста (пример отчета – в конце документа)
9. автоматическое формирование по результатам анализа заготовки презентации в LaTeX определенной структуры (таблички со значимыми различиями, хар-ки кластеров и т.д.) со свободными местами для текста
10. заполнение корреляционной и регрессионной таблицы в Excel(как это было в дз про Word и Excel – см. обсуждения).
11. экспорт какой-либо подвыборки в БД, т.е. возможность задать условие в интерфейсе и последующий выбор строк / столбцов.

План работы

1. Реализовать программу
2. Выбрать конкретные данные
3. Доработать вручную отчет и презентацию для представления по выбранным данным
4. Сдать отчет
5. Выступить с презентацией

Статистические методы – сравнение групп (t-критерий, хи-квадрат), корреляция, регрессия, кластеризация

Технологии – Ваш выбор (материалы в группе представляются для R, R.Net, C#, автоматизация с Word и Excel, Access, LaTeX, но можно заменить R/C# на Python, например)

Отчетность – работающая программа, заготовки отчета и презентации, готовый отчет, готовая презентация с выступлением

Возможное применение – регулярные отчеты по изменяющейся базе.

Вариант 1.

План работы обучающегося над проектом по обработке и интеллектуальному анализу данных:

1. Выбрать данные
2. Выполнить первый этап анализа данных (описательные статистики, сравнение групп, корреляционный анализ)
3. Выступить с презентаций, обсудить результаты
4. Выполнить второй этап анализа данных (регрессионный и кластерный анализ, математическое моделирование)
5. Сдать полный отчет
6. Выступить с презентацией, защитить полученные результаты

Вариант 2.

План работы

A. Выбрать данные

B. Выполнить анализ

C. Реализовать программу

D. Доработать вручную отчет и презентацию для представления

E. Сдать отчет

F. Выступить с презентацией

Вариант 3. По выбору обучающегося возможна сдача экзамена в устной форме по билетам, составленным из вопросов, перечисленных в п. 2.2 данной программы.

**3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса**

Оценка обучающимися содержания и качества учебного процесса по дисциплине осуществляется в установленном в СПбГУ порядке.

**3.2. Кадровое обеспечение**

**3.2.1 Образование и (или) квалификация штатных преподавателей и иных лиц, допущенных к проведению учебных занятий**

К преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению и либо прошедшие обучение в аспирантуре не менее года, либо имеющие ученую степень.

**3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом**

Специальных требований нет.

**3.3. Материально-техническое обеспечение**

**3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий**

В аудитории, где проводятся занятия, необходимо наличие досок и средств письма на них, а также технических средств, указанных в пункте 3.3.2.

**3.3.2 Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования**

В аудитории, где проводятся занятия, для преподавателя и каждого студента должен быть персональный компьютер, на котором установлена компьютерная программа для обработки данных SPSS и приложение MS Excel новейших версий. Компьютер преподавателя должен быть подключён к проекционному оборудованию.

**3.3.3 Характеристики специализированного оборудования**

Специальных требований нет.

**3.3.4 Характеристики специализированного программного обеспечения**

Специальных требований нет.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Специальных требований нет.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. StatSoft, Inc. (2012). Электронный учебник по статистике. Москва, StatSoft. WEB: <http://www.statsoft.ru/home/textbook/default.htm>.

2. Кутейников А.Н. Математические методы в психологии. СПб: Речь, 2008.

3. Наследов А.Д. Математические методы психологического исследования. СПб, 2004.

4. Сидоренко Е.В. Методы математической обработки в психологии. – СПб: Речь 2006.

**3.4.2 Список дополнительной литературы**

1. Бурлачук Л.Ф., Морозов С.М. Словарь-справочник по психодиагностике. – СПб: Питер, 2005.

2. Мастицкий С. Э., Шитиков В. К. (2014) Статистический анализ и визуализация данных с помощью R. - Электронная книга, 400 с.  
<http://www.ievbras.ru/ecostat/Kiril/R/Mastitsky%20and%20Shitikov%202014.pdf>

3. <https://compscicenter.ru/courses/math-stat/2013-spring/>

4. <https://compscicenter.ru/courses/introdatamining/2011-autumn/>

5. https://ru.wikibooks.org/wiki/%D0%98%D0%B7\_SPSS\_%D0%B2\_R

**3.4.3 Перечень иных информационных источников**

1. UCI Machine Learning Repository

2. CMU StatLib Datasets Archive

3. Airport, airline and route data

4. World Bank Data

5. <http://r-dir.com/reference/datasets.html>

6. <http://getthedata.org/>

7. http://www.umass.edu/statdata/statdata/

**Раздел 4. Разработчики программы**

Тулупьева Татьяна Валентиновна к.пс.н., доцент, доцент кафедры информатики. tvt100a@mail.ru