**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Data Science: инструментарий и жизненный цикл проекта

Data Science: Tools and Project Lifecycle

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 061294

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Дисциплина направлена на создание среды, в которой обучающийся имеет возможность реализовать свои уже ранее обретенные фундаментальные знания и узкопрофессиональные навыки на практике (в практическом применении) совместно с развитием навыков деловой коммуникации, сбора, анализа и удовлетворения потребностей заказчика в проведении исследований по сбору данных, последующему их анализу, и формированию и представлению выводов.

Одной из задач курса является формирование у обучающихся научно-теоретических и практических знаний в области интеллектуального анализа данных (в частности, в теории вероятностных графических моделей), умений и навыков общения. Помимо этого, в рамках курса в качестве дополнительной подготовки для развития навыков разработки специфических программных продуктов, ориентированных на применение в подготовке научных текстов, предусмотрен модуль подготовки научно-технических публикаций и презентаций в системе LaTeX.

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

Максимальная эффективность программы будет обеспечена при следующем условии: обучающийся владеет базовыми математическими понятиями и навыками программирования на языке высокого уровня.

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

В результате освоения дисциплины обучающийся должен получить знания в области интеллектуального анализа данных (в частности, в теории вероятностных графических моделей), умения и навыки общения.

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

Лекции (18 часов) построены на активных формах обучения, в первую очередь — на диалоге преподавателя и обучающихся), семинары (18 часов) построены на интерактивных формах обучения, в первую очередь — на модерируемой дискуссии студентов в отношении работ, проектов, презентаций докладов друг друга.

Предполагается, также, что самостоятельную работу в предлагаемом курсе обучающиуся выполняют с обязательным использованием компьютера.

**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.) |
| ТРАЕКТОРИЯ 2 СЕМЕСТРА | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 2 | 18 | 18 | 2 |  |  |  |  |  | 2 |  |  |  | 38 |  | 30 |  | 20 | 3 |
|  | 1-8 | 1-8 | 1-8 |  |  |  |  |  | 1-8 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 18 | 18 | 2 |  |  |  |  |  | 2 |  |  |  | 38 |  | 30 |  |  | 3 |
| ТРАЕКТОРИЯ 4 СЕМЕСТРА | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 4 | 18 | 18 | 2 |  |  |  |  |  | 2 |  |  |  | 38 |  | 30 |  | 20 | 3 |
|  | 2-100 | 10-25 | 2-100 |  |  |  |  |  | 2-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 18 | 18 | 2 |  |  |  |  |  | 2 |  |  |  | 38 |  | 30 |  |  | 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 |
| ТРАЕКТОРИЯ 2 СЕМЕСТРА | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 2 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |
| ТРАЕКТОРИЯ 4 СЕМЕСТРА | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 4 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Наименование темы (раздела, части)** | **Наименование подтемы (подраздела, подчасти)** | **Вид учебных занятий** | **Кол-во часов** |
| 1 | Введение | Современный анализ данных. Жизненный цикл проекта: постановка вопроса, разведочный анализ данных, построение модели, интерпретация, представление результатов. | лекции | 2 |
| сам. раб. с использованием  методических материалов | 8 |
| 2 | Постановка вопроса | Технологии коммуникации с заказчиком, выявления потребностей заказчика в обработке данных. Перевод потребностей заказчика на язык данных. Результат: формулировка темы собственного проекта по анализу данных. | лекции | 4 |
| семинары | 4 |
| сам. раб. с использованием  методических материалов | 8 |
| 3 | Построение модели | Разведочный анализ данных. Типы переменных, измерения и шкалы, частотный анализ и описательные статистики. Инструментарий: R, SPSS. | лекции | 6 |
| семинары | 4 |
| Взаимосвязи между переменными: проверка гипотез, статистические тесты, корреляция. Интерпретация: выборка, репрезентативность, величина эффекта. | сам. раб. с использованием  методических материалов | 8 |
| Введение в машинное обучение, регрессия, классификация, кластеризация. Интерпретация результатов. |
| Вероятностные графические модели. Элементы базовых теорий. Алгебраические байесовские сети. Байесовские сети доверия. Прикладные программные библиотеки. Интерпретация результатов |
| 4 | Хранение данных и автоматическая работа с данными | Базы данных. Создание баз данных. Разработка структуры. Заполнение, хранение, основные операции. Разработка приложения. C#. Графический интерфейс. Автоматизация доступа к базам данных и анализа последних. | лекции | 2 |
| семинары | 2 |
| сам. раб. с использованием  методических материалов | 8 |
| 5 | Представление результатов | Подготовка публичного выступления. Технологии проведения публичного выступления. Технологии ведения дискуссии. | лекции | 4 |
| семинары | 8 |
| сам. раб. с использованием  методических материалов | 6 |
| Технологии автоматического формирования документов. Работа с MS Word в C#-приложении. R.Net. Результат: программа с автоматизацией отчетности |
| Отчет. Требования к оформлению документации. Язык LaTeX. Автоматическое создание отчетов из R. Результат: промежуточный отчет по проекту (письменный) |
| Подготовка презентации. Язык LaTeX. Подготовка презентаций. Пакет beamer. Результат: отчет по проекту (выступление) |
| 6 | Промежуточная аттестация | | самост. работа | 30 |
| консультация | 2 |
| экзамен | 2 |
| **Итого** | | | | **108** |

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Успешное освоение дисциплины возможно благодаря посещению семинарских занятий, участию в обсуждении вопросов, подготовленных к занятию, самостоятельной работе, включающей в себя чтение специальной литературы по разделам темы, а также подготовленных преподавателем и студентами электронных материалов. В силу того, что  
дисциплина проектноориентированная, часть лекционных занятий может быть заменена на семинарские, посвященные проектам обучающихся.

**3.1.2 Методическое обеспечение самостоятельной работы**

Основная литература, а также, при необходимости, дополнительные информационные источники, подобранные или согласованные преподавателем.

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

В целях текущего контроля успеваемости предложены задания, направленные на подготовку их выпускной квалификационной работы.

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

Аттестация возможна в двух вариантах:

1. Экзамен по проекту

Защита проекта по анализу выбранного и согласованного на занятиях набора (наборов) данных, состоящего из 1) аналитического отчета (в форме отчета о НИР, статьи, брошюры, иной согласованной формы), 2) презентации по проекту, 3) устного доклада по проекту (4–9 минут).

Максимальный балл по компонентам проекта: 1) текст — 40 баллов, 2) презентация — 25 баллов, 3) устный доклад — 35 баллов. Штраф за каждые полные или неполные 0,5 минуты отклонеения от заданного преподавателем норматива длительности устного доклада — 5 баллов. Допускается до 20 бонусных баллов, если студент особо качественно, особы быстро выполнил работу или предложил удачное решение по тематике дисциплины или при выполнении проекта.

Полученные баллы суммируются, штраф — вычитается.

1. Балльно-рейтинговая система.

Получение обучающимися баллов в процессе изучения дисциплины за работу на семинарских занятиях и самостоятельную работу в течение семестра. В этом случае балльно-рейтинговая система учитывает 1) посещаемость, 2) готовность к занятиям, 3) работу на занятиях, 4) составляющие финального проекта (текст, презентация и устный доклад). Максимальные баллы соответственно 12, 12, 12, 64 (26, 16, 22). Допускается до 20 бонусных баллов, если обучающийся особо качественно, особо быстро выполнил работу или предложил удачное решение по тематике дисциплины или при выполнении проекта. Балльно-рейтинговая система каждый раз в начале чтения дисциплины адаптируется к фактическому распределению занятий по календарным дням и доводится до студентов на одном из трех первых аудиторных (контактных) занятий.

В случае неаттестации обучающегося по балльно-рейтинговой системе, его аттестация проводится на экзамене (повторном экзамене) в форме экзамена по проекту со штрафом 18 баллов; в случае аттестации — набранные баллы рассматриваются как набранные баллы при защите проекта.

На аттестационное мероприятие обучающийся должен явиться к его началу с полностью готовыми материалами проекта. В случае, когда обучающийся не явился в течение 15 минут от начала экзамена, в ведомости проставляется «не явился».

На подготовку к экзамену отводится 60 минут, по согласованию обучающегося и экзаменатора (экзаменационной комиссии) срок может быть сокращен. Экзаменатор (экзаменационная комиссия) может задать вопросы по проекту и по содержанию дисциплины в контексте анализа подготовленного студентом проекта.

Накопление баллов обучающимися во время лекционных занятий, семинаров, по результатам самостоятельной работы и за защиту проекта. Общая сумма баллов: 100. Распределение оценок по баллам:

2: менее 60;

3: от 60 до 74,(9);

4: от 75 до 89,(9);

5: от 90.

Образцы требований к проектам и альтернативные варианты распределения баллов при использовании балльно-рейтинговой системы приводятся ниже. Выбор определяется уровнем подготовки обучающихся, целями (потребностями) их подготовки, индивидуальными запросами обучающихся.

Вариант 1.

Для выполнения итогового проекта обучающимися 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.

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

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 4.

Проект выполняется в рамках более широкого проекта по подготовке выпускной квалификационной работы и (или), прохождения практики, формирования отчета по ней.

**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 ведению семинарских и лекционных занятий привлекаются преподаватели, имеющие ученую степень и опыт работы или преподавательской деятельности, содержательно связанной с дисциплиной.

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

Для подготовки и проведения занятий требуется лаборант или инженер для следующих работ:

1) техническая подготовка каталогов исходных данных в форме, удобной для учебной работы;

2) поддержания работоспособности компьютерного класса, проецирующего оборудования и его оснастки.

**3.3. Материально-техническое обеспечение**

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

Стандартно оборудованная аудитория вместимостью 25 человек для проведения интерактивных занятий: видеопроектор, экран, акустические колонки.

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

Нет.

**3.3.3 Характеристики специализированного оборудования**

Рабочее место преподавателя должно быть оснащено оборудованием не ниже: Pentium IV-800/ОЗУ-256 Мб / Video-32 Мб / Sound card – 16bit /Headphones / HDD 80 Гб / СD-ROM – 48x / Network adapter – 10/100/ Мбс / SVGA – 19”.

**3.3.4 Характеристики специализированного программного обеспечения**

Процесс чтения лекций должен сопровождаться демонстрацией работы программного обеспечения, подготовленного преподавателем и студентами. В качестве среды программирования требуется продукты R Studio, SPSS, MS Visual Studio, MS Access, а также доступ в интернет, в том числе, чтобы пользоваться справочными сервисами и сервисами онлайн работы с документами LaTeX.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Фломастеры цветные, губки.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Van der Loo M. P. J. Learning RStudio for R statistical computing. – Packt Publishing Ltd, 2012. – ЭР по подписке СПбГУ: <https://find.library.spbu.ru/vufind/Record/ocn823719008>.

2. Davis C. SPSS for applied sciences: Basic statistical testing. – Csiro publishing, 2013. ЭР по подписке СПбГУ: <https://find.library.spbu.ru/vufind/Record/ocn828866313>.

3. Тулупьев А.Л., Николенко С.И., Сироткин А.В. Основы теории байесовских сетей: учебник. СПб.: Изд-во СПбГУ, 2019. 399 с. – ЭР ЭБС «Лань» по подписке СПбГУ: <https://proxy.library.spbu.ru:2279/book/120100>

4. Тулупьева Т.В. Лекции по основам технологий деловой коммуникации: учебник. СПб.: Изд-во СПбГУ, 2019. 164 с. – ЭБС «Лань» по подписке СПбГУ: https://proxy.library.spbu.ru:2279/book/120087.

Кроме того, используются указанные преподавателем информационные источники и сервисы, доступные в интернет, доступные на момент чтения дисциплины.

**3.4.2 Список дополнительной литературы**

1. Тулупьев А.Л. Алгебраические байесовские сети: локальный логико-вероятностный вывод: учебное пособие. СПб.: ВВМ, 2019. 138 с.

2. Тулупьев А.Л. Алгебраические байесовские сети: глобальный логико-вероятностный вывод в деревьях смежности: учебное пособие. СПб.: ВВМ, 2019. 108 с.

**3.4.3 Перечень иных информационных источников**

Открытые на момент изучения дисциплины информационные источники в среде интернет по компьютерным инструментам (технологиям), используемым при изучении дисциплины, а также открытые информационные источники в среде интернет, содержащие наборы данных.

**Раздел 4. Разработчики программы**

Тулупьев Александр Львович, д.ф.-м.н., доцент, профессор математико-механического факультета СПбГУ. alexander.tulupyev@gmail.com, +7 931 288-31-77;

Тулупьева Татьяна Валентиновна, к.пс.н., доцент, доцент математико-механического факультета СПбГУ. tvt100a@mail.ru, +7 921 753-54-88;

Абрамов Максим Викторович, к.т.н., старший преподаватель кафедры информатики, mva16@list.ru +7(981) 680-99-29; младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН Харитонов Никита Алексеевич, nikita.kharitonov95@yandex.ru, +7 921 354-15-97.

Разработка программы дисциплины частично поддержана средствами гранта ГК170001610 — «Проект-победитель Грантового конкурса Стипендиальной программы Владимира Потанина 2016/2017».