**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Data Science: комплексы программ

Data Science: Software Systems

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 056347

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Основным методологическим принципом построения программы курса, равно как и всей концепции обучения в области информатики, информационных технологий и систем, является принцип поэтапного системного накопления знаний и формирования необходимых компетенций по модели: от простого и/или знакомого к сложному и/или не-знакомому, а основной методологической стратегией прохождения отдельных разделов программы является ступенчатость и цикличность, предусматривающие постепенный возврат к ранее усвоенному материалу на более высоком языковом и концептуальном уровне. Системообразующей деятельностью в рамках учебной дисциплины является самостоятельная работа как основная форма приобретения, закрепления и развития обучающимся соответствующих навыков, умений, а также системы операциональных знаний; роль преподавателя-информатора меняется на роль преподавателя-координатора, преподавателя-консультанта, а также преподавателя-фасилитатора.

Отдельные параметры курса (выбор языков программирования, выбор интегрированных сред разработки, выбор тем и элементов тем, формирование перечня обязательных к выполнению домашних заданий, лабораторных работ, упражнений, формирование обязательного для прочтения списка литературы, выбор тем для занятий и перечень вопросов к экзамену, распределение тем для лекционных, лабораторных, практических, самостоятельных занятий и др.) могут варьироваться преподавателем по порядку и режиму изучения, степени сложности и объему в зависимости от уровня подготовки обучающихся и в зависимости от сложившихся реалий в соответствующем срезе профессиональной сферы. Часть учебной деятельности, предусмотренной настоящим РПУД, может быть реализована или повторно выполнена (пройдена) на соответствующих практических занятиях.

Основной целью дисциплины является освоение обучающимся ряда технологий, языков, средств разработки и информационных систем в степени, достаточной для компетентного участия в проектах по разработке и применению комплексов программ (в том числе информационно-аналитических систем), включая такие виды деятельности как анализ, кодирование и документирование. Сопутствующей целью выступает формирование у обучающегося навыков совместного, комплексного применения таких технологий, языков, средств разработки и информационных систем. Еще одной целью выступает подготовка слушателей к освоению профессиональных дисциплин «Технологии разработки программного обеспечения», «Программная инженерия» и ряда спецкурсов, а также к выполнению курсовых работ, выпускной квалификационной работы, научной (научно-исследовательской) работы.

Результатами курса являются решение ряда задач, необходимых для достижения перечисленных выше целей:

• сформировать у обучающихся навыки применения знаний о языке программирования высокого уровня и соответствующей интегрированной среде разработки (С# и MS Visual Studio, либо иных по выбору преподавателя) в разработке комплексов программ;

• сформировать у обучающихся навыки применения знаний о языке SQL и соответствующей СУБД (MS Access, либо иной по выбору преподавателя) в разработке отвечающих за интенсивную работу с данными компонент комплексов программ и интеграцию таких компонент с другими компонентами комплекса;

• сформировать у обучающихся базовый уровень владения методами автоматизации приложений MS Office (на примере MS Excel и MS Word), используемых для автоматического формирования отчетов в соответствующих компонентах комплексов программ;

• сформировать у обучающихся базовый уровень владения методами генерации текстов на языках HTML и LaTeX, используемых для формирования отчетов в соответствующих компонентах комплексов программ;

• сформировать у обучающихся умения самостоятельно искать и использовать литературу (и более широко — доступные информационные источники) по специальности, базовую функциональность интегрированных сред разработки для реализации комплексов программ, написанных на указанном выше языке программирования высокого уровня;

• сформировать способность и готовность применять подход, ставший стандартом в ряде профессиональных сообществ, в подготовке публикаций, технической документации и презентаций на языке LaTeX;

• сформировать у обучающихся базовый уровень владения языком R и соответствующей вычислительной средой для решения базовых задач анализа данных и математического моделирования, возникающих при проектировании, реализации и эксплуатации комплексов программ, или сопутствующих им и решению релевантных научных задач;

• сформировать способность и готовность аргументированно, последовательно, логически верно и содержательно ясно строить устную и письменную речь, использовать навыки публичной речи, ведения дискуссии и полемики, в том числе с разработкой и последующим использованием компьютерных презентаций, иллюстративных компьютерных программ, опорных конспектов;

• сформировать способность и готовность применять системный подход и математические методы в формализации решения прикладных задач;

• сформировать способность и готовность эффективно искать необходимую профессиональную информацию в сети Интернет и в специализированных базах и банках данных с использованием возможностей современных поисковых систем.

• развить у обучающихся логическое мышление, а также способность строить аргументацию на статистических выводах.

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

Программа курса нацелена на обучающихся 3-го курса и рассчитана на их глубокие знания и навыки программирования как технологической и математической дисциплины.

Максимальная эффективность программы будет обеспечена при условии, что обучающийся:

• владеет базовыми математическими понятиями, достаточными для работы с формулировками математических утверждений и построения алгоритмических конструкций;

• владеет навыками работы в ОС Windows 7 (графический пользовательский интерфейс, иерархия папок и документов, операции с папками и файлами, запуск программ);

• владеет навыками чтения, письма, устной речи на русском языке в объеме 11 классов средней школы;

• владеет достаточным объемом лексики английского языка и умеет читать техническую литературу (в том числе содержание электронных информационных источников) на английском языке; умеет пользоваться доступны-ми в среде Интернет электронными словарями (translate.ru, multitran.ru и др.);

• владеет навыками создания, разработки, модификации, сохранения документов в текстовом формате, в форматах MS Word, MS Excel, MS Power-Point; навыками работы с документами в pdf-формате;

• владеет навыками работы в Интернет, в том числе с порталами электронной почты, порталом vk.com, порталом СПбГУ и математико-механического факультета;

• владеет навыками поиска (в т.ч. умеет использовать поисковые инструменты Google, Yahoo, Yandex и др.) и получения информации из среды Интернет;

• владеет навыками проектирования и разработки реляционных СУБД, использования наиболее часто использующихся конструкций языка SQL;

• владеет навыками программирования на хотя бы одном языке программирования высокого уровня (С++, С#, Java и пр.) и хотя бы одной соответствующей интегрированной среды разработки;

• владеет навыками анализа задач, возникающих в автоматизации бизнеспроцессов и в научных исследованиях, а также навыками алгоритмизации решения указанных задач в рамках своих профессиональных компетенций;

• успешно освоил дисциплины «Информатика», «Программирование», «Дискретная математика», «Базы данных и СУБД», «Структуры и алгоритмы компьютерной обработки данных», предусмотренные в предшествующих семестрах компетентностно-ориентированным планом направления подготовки.

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

Обучаемые приобретают следующие знания

• информационные системы;

• системы управления базами данных;

• языки запросов;

• языки разметки;

• технологии автоматического формирования документов;

• технологии автоматического формирования документов;

• система (инструментарий) для статистического анализа данных.

Данные знания формируются при использовании в качестве примеров компонент MS Office (Word, Excel, Access, Power Point), языков C#, Java, SQL, HTML, LaTeX, системы R, а также средств разработки приложений MS Visual Studio, Netbeans и, при необходимости, одного–двух доступных на момент обучения текстовых редакторов.

Умения

• разрабатывать, компилировать и отлаживать программы на перечисленных выше языках с помощью перечисленных выше средств;

• организовывать в программах, разработанных на языках C# и Java, обмен данными с базой данных MS Access, автоматическое формирование HTML и LaTeX-документов, автоматическое формирование MS Word и MS Excel-документов (только С#-программы);

• осуществлять обмен данными между системами R, MS Excel, MS Office;

• выполнять первичный статистический анализ (расчет нескольких характеристик выборки, построение диаграмм) с помощью MS Excel и R.

Навыки

• работы с информацией из различных источников для решения профессиональных задач;

• применения в профессиональной деятельности знаний и умений, полученных при обучении по дисциплине «Data Science: комплексы программ»;

• взаимодействия с коллегами, работы в коллективе.

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

В качестве основных интерактивных форм (общее количество 19 часов) предполагается

• проведение лекций (48 часов), основная часть лекционного занятия проводится в интерактивных формах; при этом до 75% лекционных занятий проводится в активных формах;

• лабораторных занятий (16 часов).

Самостоятельная работа с использованием методических материалов (4 часа). В зависимости от степени подготовки обучающихся и их профессиональной направленности число часов, отведенных на лабораторные работы и практикумы (проводимых в различном режиме), может быть увеличено за счет числа часов, отведенных на проведение лекций.

**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**

**Базовый курс Основная траектория Очная форма обучения**

Период обучения: **Семестр 5**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п.п.** | **Наименование темы (раздела, части)** | **Вид учебных занятий** | **Кол-во часов** |
| 1 | Тема 1. Язык программирования C#. | лекции | 12 |
| лабораторные работы | 6 |
| по методическим материалам | 4 |
| 2 | Тема 2. Введение в базы данных. | лекции | 12 |
| лабораторные работы | 4 |
| по методическим материалам | 3 |
| 3 | Тема 3. Технологии обмена данными между MS Access и программами на языках программирования C#. | лекции | 6 |
| лабораторные работы | 2 |
| по методическим материалам | 2 |
| 4 | Тема 4. Технологии автоматического формирования документов. | лекции | 6 |
| лабораторные работы | 2 |
| по методическим материалам | 2 |
| 5 | Тема 5. Язык LaTeX. | лекции | 6 |
| лабораторные работы | 1 |
| по методическим материалам | 2 |
| 6 | Тема 6. Язык программирования и программная среда вычислений R. | лекции | 6 |
| лабораторные работы | 1 |
| по методическим материалам | 2 |

**Тема 1. Язык программирования C# .**

Среда разработки MS Visual Studio. Основные типы, выражения C#. Представление конструкции if-else, switch в C#. Представление конструкций for, while, foreach в C#. Понятие пространства имен, методов в C#. Классы в C#. Интерфейсы, наследование, полиморфизм. Массивы. Описание исключений в C#. Параметры по умолчанию. Свойства. Абстрактные классы, виртуальные и невиртуальные методы. Множественное наследование. Partial. Sealed классы. Статический класс и статические члены. Статическая инициализация. Nullable. Generic классы. Ковариантность. Контрвариантность. Индексаторы. Перегрузка операторов. Yield return. String builder. Коллекции. Unsafe код. Атрибуты. Сериализация. Отложенная инициализация. Файлы. Enum. Bitarray. Dynamic. Linq. Анонимные методы и лямбда-выражения. Многопоточность. Делегаты и события. (Содержание и объем темы зависят от объема и уровня подготовки группы студентов/слушателей.)

**Тема 2. Введение в базы данных.**

Понятие баз данных. Модели данных. Реляционные БД: Отношение, Поле, Запись, Связи и ключевые поля. СУБД: понятие и функции. Функции SQL. Правила синтаксиса и основные запросы SQL. Создание, изменение и удаление базовых таблиц БД. Добавление новых данных: запрос INSERT. Обновление существующих данных: запрос UPDATE. Удаление существующих данных: запрос DELETE. Синтаксис запроса SELECT: Предложения SELECT и FROM. Синтаксис запроса SELECT: Предложение WHERE. Синтаксис запроса SELECT: Агрегирующие функции. Синтаксис запроса SELECT: Предложение GROUP BY. Синтаксис запроса SELECT: Предложение HAVING. Синтаксис запроса SELECT: Сортировка результатов запроса. Соединения таблиц (оператор JOIN). Запросы с вложенными запросами. Понятие нормализации. Приведение базы данных к нормальной форме. Первая, вторая, третья нормальные формы. Независимость данных и приложений. Ограничение доступа.

**Тема 3. Технологии обмена данными между MS Access и программами на языках программирования C#.**

MS Access: достоинства и недостатки. Создание и управление базой данных MS Access, ее элементами. Принцип и технология разделения интерфейса и данных в MS Access. Объекты MS Access. Особенности диалекта SQL в MS Access. Работа с MS Access в C#-приложении.

**Тема 4. Технологии автоматического формирования документов.**

Работа с MS Word в C#-приложении: технология COM. Объекты MS Word. Работа с MS Word в C#-приложении: открытие/создание/ сохранение файла, работа с текстом, создание таблиц. Работа с MS Excel в C#-приложении. Объекты MS Excel. Запуск MS Excel из C#-приложения. Создание книги, вывод информации. Работа с MS Excel в C#-приложении: вставка формул. Основы HTML: основные понятия. Структура HTML-документа. Основы HTML: теги и атрибуты. Основы HTML: вывод текстовой информации: параграф, заголовок. Основы HTML: форматирование текста. Основы HTML: ссылки, графические элементы.

**Тема 5. Язык LaTeX.**

LaTeX как издательская система и как язык гипертекстовой разметки. MiKTeX, инсталляция среды, компиляция программ. Структура файла на LaTeX. Виды документов. Опции \documentclass. Подразделы. Базовые приемы верстки документов. Продвинутые приемы верстки документов. Таблицы, рисунки, листинги. Верстка математических формул. Формирование библиографии, предметного указателя, списков таблиц и рисунков. Особенности хранения и обработки библиографических ссылок. Подготовка презентаций. Пакет beamer.

**Тема 6. Язык программирования и программная среда вычислений R.**

Модель данных R. Управляющие структуры. Базовые приемы статистического анализа. Пакеты. Примеры.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Успешное освоение дисциплины возможно благодаря посещению лекций, участию в обсуждении вопросов, подготовленных к занятию, подготовке презентаций, проектов, отчетов по тематике курса.

Для организации взаимодействия с обучающимися, предоставления отчетности и организации самостоятельной работы создается группа Вконтакте. В группе размещаются все необходимые материалы: презентации лекций, задания, информация о системе оценивания.

Предусматривается три варианта освоения дисциплины, выбор которых зависит от уровня подготовки обучающихся, индивидуальных целей и мотивации обучающихся, организационных возможностей. Первый вариант — классический, обучение ведется только по данной дисциплине. Второй вариант — более интенсивный, основан на проектном методе и представляет собой комбинацию двух дисциплин «Data Science: комплексы программ» и «Data Science: основы обработки и анализа данных», что позволяет сформировать у обучающихся комплексные навыки, которые более применимы в профессиональной деятельности и более востребованы на современном рынке труда. Рабочие программы указанных дисциплин комплементарны друг другу. Формальные аспекты совместного комплексного изучения указанных дисциплин отражены в п. 3.1.3. Третий вариант — индивидуальная образовательная траектория в рамках дисциплины, форма работы по которой и промежуточная отчетность согласуется при обращении студента к преподавателю по электронной почте. Требования по итоговой отчетности отражены в п. 3.1.3. Такая индивидуальная образовательная траектория должна быть согласована с преподавателем до установленного им срока (например, в течение первого месяца обучения в семестре, в котором проводятся занятия по дисциплине); в противном случае считается, что студент движется по общей образовательной траектории в рамках дисциплины.

Для успешного освоения дисциплины, в частности — для достижения высокого уровня мастерства презентации результатов проектов перед заказчиком или в учебных ситуациях, рекомендуется параллельное изучение дисциплины «Технологии деловой коммуникации».

**3.1.2 Методическое обеспечение самостоятельной работы**

Самостоятельная работа обучающихся, 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, которой отводится не менее половины учебного времени при очной форме обучения. Время, отводимое на самостоятельную работу, должно использоваться студентами 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.

К числу методических пособий относятся общие методические рекомендации и указания по самостоятельной работе. 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. 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)**

Поскольку дисциплина имеет практическую направленность и ориентирована, в том числе, на успешное представление результатов деятельности обучающимся в учебной или условно деловой ситуации в форме доклада с презентацией, подготовка ответов на экзамене проводится при без ограничения доступа к материалам дисциплины, литературе и иным информационным источникам.

Вариант 1. Отчет только по дисциплине «Data Science: комплексы программ».

Экзамен выставляется на основе баллов, набранных обучающимся при выполнении заданий в течение семестра и по защите итогового проекта.

В течение семестра для выполнения итогового проекта обучающимися выбираются (при согласовании с преподавателем) данные (dataset) для хранения, редактирования и обработки; особый акцент делается на демонстрацию навыков разработки и применения комплексов программ, предназначенных для хранения, редактирования, подготовки к последующему анализу и собственно анализа данных. Кроме того, еще один акцент делается на формирование и представление отчета, использование технологий автоматизации MS Office при формировании его отдельных частей, использовании LaTeX в подготовке презентации к докладу. Наконец, во внимание принимаются иные итоги освоения дисциплины, указанные в разделе 1 настоящей программы.

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/>

Для внесения, редактирования, обработки и иных операций с данными и фрагментами отчета (при необходимости — презентации) разрабатывается комплекс программ с привлечением (где применимо и обосновано) MS Visual Studio, MS Access, MS Word и иных компонент MS Office, технологий автоматизации MS Office, технологий взаимодействия с реляционной БД, технологий R.NET и иных — по усмотрению студента.

По результатам обработки и 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** |
| 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 |
|
| 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 | 35 | программа |
| отчет |
| презентация |
| Посещаемость | 14 | распределение максимальных баллов по дням занятий устанавливается и объявляется в течение месяца с начала семестра, т.к. определяется календарем |
|  |  |  |
| 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 «Data Science: основы обработки и анализа данных» и «Data Science: комплексы 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(или) 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.

Требования к результатам разработки аналогичны варианту 1 с учетом адаптации к согласованным условиям индивидуальной работы.

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.

В случае вариантов 1 и 2 экзаменационная оценка выставляется по результатам работы и защиты проекта в течение семестра на основе таблиц баллов в разделе 3.1.3. В случае варианта 3 экзамен проводится по результатам представления проекта и его защиты согласно содержанию утвержденной индивидуальной траектории обучения в рамках дисциплины.

По выбору обучающегося в качестве опции для получения дополнительных баллов возможна сдача экзамена в устной форме по билетам, составленным из вопросов, приведенном ниже. Ответ студента производится в форме доклада с презентацией, подготовленной с помощью LaTeX и, при необходимости, с демонстрацией иллюстративных примеров программ. Максимальное число баллов – 30.

Примерный список вопросов к экзамену

1. Среда разработки MS Visual Studio.

2. Основные типы, выражения C#.

3. Представление конструкции if-else, switch в C#.

4. Представление конструкций for, while, foreach в C#.

5. Понятие пространства имен, методов в C#.

6. Классы в C#.

7. Интерфейсы, наследование, полиморфизм.

8. Массивы.

9. Описание исключений в C#.

10. Параметры по умолчанию.

11. Свойства.

12. Абстрактные классы, виртуальные и невиртуальные методы.

13. Множественное наследование. Статический класс и статические члены. Статическая инициализация.

14. Generic классы.

15. Ковариантность. Контрвариантность.

16. Индексаторы.

17. Enum.

18. Linq.

19. Анонимные методы и лямбда-выражения.

20. Делегаты и события.

21. Понятие баз данных. Модели данных.

22. Реляционные БД: Отношение, Поле, Запись, Связи и ключевые поля. СУБД: понятие и функции.

23. Правила синтаксиса и основные запросы SQL. Создание, изменение и удаление базовых таблиц БД.

24. Добавление новых данных: запрос INSERT. Обновление существующих данных: запрос UPDATE. Удаление существующих данных: запрос DELETE.

25. Синтаксис запроса SELECT.

26. Соединения таблиц (оператор JOIN).

27. Понятие нормализации. Приведение базы данных к нормальной форме. Первая, вторая, третья нормальные формы.

28. Работа с MS Access в C#-приложении. Объекты MS Access.

29. Работа с MS Word в C#-приложении: технология COM. Объекты MS Word.

30. Работа с MS Excel в C#-приложении. Объекты MS Excel.

31. \*Основы HTML: основные понятия. Структура HTML-документа.

32. \*Основы HTML: теги и атрибуты.

33. \*Основы HTML: вывод текстовой информации: параграф, заголовок.

34. \*Основы HTML: форматирование текста.

35. \*Основы HTML: ссылки, графические элементы.

36. LaTeX как издательская система и как язык гипертекстовой разметки.

37. Структура файла на LaTeX. Виды документов.

38. Модель данных R.

39. Управляющие структуры.

40. Базовые приемы статистического анализа.

**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 чтению лекций привлекаются преподаватели, имеющие базовое образование и/или ученую степень соответствующую профилю преподаваемой дисциплины.

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

Для проведения лекционно-практических, практических занятий, лабораторных работ дополнительно привлекаются аспиранты соответствующих специальностей, либо ассистенты, старшие преподаватели, доценты, иные лица, имеющие профессиональные знания в темах, рассматриваемых в рамках дисциплины.

**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, электронного проектора, 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 обучающийсяа должен быть персональный компьютер, на котором установлены Java SE Development Kit и интегрированная среда разработки NetBeans новейших версий, среда Microsoft Visual Studio, инструменты Microsoft Office, среда программирования R. Компьютер преподавателя должен быть подключён к проекционному оборудованию. (Требования к программному обеспечению могут варьироваться в зависимости от предусмотренных в п. 1.1. изменений программы.)

**3.3.3 Характеристики специализированного оборудования**

Специальных требований нет.

**3.3.4 Характеристики специализированного программного обеспечения**

При использовании электронных документов каждый обучающийся во время занятий и самостоятельной подготовки должен быть обеспечен рабочим местом в компьютерном классе с выходом в интернет и корпоративную сеть факультета.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Специальных требований нет.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Котельников И. А., Чеботаев П. З. LATEX по-русски. – 3-е издание, перераб. и доп.— Новосибирск: Сибирский хронограф, 2004. – 496 с.

**3.4.2 Список дополнительной литературы**

1. Бекаревич Ю., Пушкина Н. Самоучитель Access 2010. СПб.: БХВ-Петербург, 2012. 432 с.

2. Сеннов А. Access 2010. Учебный курс. СПб.: Питер, 2010. 288 с.

3. Васильев А. C#. Объектно-ориентированное программирование: Учебный курс. СПб.: Питер, 2012. 320 с.

**3.4.3 Перечень иных информационных источников**

1. The Java Tutorials. <http://docs.oracle.com/javase/tutorial>

2. Visual Studio Library <http://msdn.microsoft.com/ru-ru/library/vstudio/>

3. Руководство по JDBC v.1 http://www.javaportal.ru/java/tutorial/tutorialJDBC/ intro.html

4. Trail: JDBC(TM) Database Access <http://docs.oracle.com/javase/tutorial/jdbc/>

5. Работа с Word в C# <http://wladm.narod.ru/C_Sharp/comword.html>

6. Работа с Excel в C# <http://wladm.narod.ru/C_Sharp/comexcel.html>

7. The R Project for Statistical Computing <http://www.r-project.org/>

8. Статистическое программирование на R: Часть 1. Купаемся в изобилии статистических возможностей <http://www.ibm.com/developerworks/ru/library/l-r1/>

9. Другие ресурсы сети Интернет.

**Раздел 4. Разработчики программы**

Тулупьев Александр Львович, д.ф.-м.н., доцент, профессор. alexander.tulupyev@gmail.com +7(931)288-31-77.

Тулупьева Татьяна Валентиновна, к.псих.н., доцент , доцент. tvt100a@mail.ru +7(921)753-54-88.

Абрамов Максим Викторович, старший преподаватель, mva16@list.ru +7(981)680-99-29.

Суворова Алена Владимировна, к.ф.-м.н. , c.н.с. suvalv@gmail.com.

Торопова Александра Витальевна, м.н.с. alexandra.toropova@gmail.com