Программа учебной дисциплины «Анализ и визуализация данных в Python»

Утверждена

Академическим советом ООП

Протокол № 38 от «21» июня 2018 г.

Автор	Ульянов В.В., д.ф.м.н., vulyanov@hse.ru
Число кредитов	4
Контактная работа (час.)	40
Самостоятельная работа (час.)	112
Курс	4
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения курса "Анализ и визуализация данных в Python" является формирование компетенций, связанных с решением задач по сбору, анализу и визуализации социологических данных. Курс направлен на изучение основ программирования в Python для применения полученных знаний и навыков при анализе и визуализации в Python как тренировочных, так и реальных данных, включая По итогам курса студенты должны научиться самостоятельно принимать решение о целесообразности использования возможностей Python для решения прикладных задач в исследовательской практике.

В результате освоения дисциплины студент должен:

знать

- основы программирования, включая стандартные алгоритмы, и их реализацию в Python (базовые структуры данных, в частности типы данных, логические выражения, условные операторы, организация множественного ветвления, циклы, последовательности (строки и списки) и словари в Python);

уметь

- строить модели, оценивать их качество и сравнивать различные модели средствами Python;

- проводить преобразование и очистку данных, «восстановление отсутствующих»

значений, построение графиков (гистограмм, графиков плотностей и диаграмм рассеяния), агрегирование данных, таблицы сопряженности, корреляционный и квантильный анализ, построение предсказательных моделей, оценка качества моделей.

владеть:

- навыками анализа реальных социологических данных в Python.

Изучение дисциплины **«Анализ и визуализация данных в Python»** базируется на следующих дисциплинах:

- Теория вероятностей и математическая статистика
- Методология и методы социологического исследования
- Прикладное программное обеспечение
- Анализ данных в социологии

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

- Обладать базовыми знаниями теории вероятностей и математической статистики;
- Знать основные методы статистического анализа социологических данных.

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

- Выполнение и защита квалификационной работы
- Организация, подготовка и презентация социологического исследования.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в Python и его основные библиотеки и модули

Общая характеристика языка Python. Базовые команды, библиотеки и модули Python, включая библиотеки SciPy, NumPy (основные пакет для выполнения научных и инженерных расчетов на Python), Matplotlib (библиотека для визуализации данных двумерной графикой), Pandas (программная библиотека на языке Python для обработки и анализа данных, в частности для работы с числовыми таблицами; работа Pandas с данными строится поверх библиотеки NumPy) и Scikit-Learn (предоставляет реализацию целого ряда алгоритмов для обучения с учителем и обучения без учителя).

Тема 2. Типы данных в Python, подготовка данных к построению моделей

Чтение и запись данных. Форматы файлов. Типы хранения данных: векторы, двухмерные таблицы, матрицы, массивы. Переформатирование данных: очистка, преобразование, слияние, изменение формы. Способы работы с пропущенными данными. Агрегирование данных и групповые операции. Построение таблиц сопряженности и квантильный анализ. Понятие функции, ссылки на функции из разных пакетов, создание собственной функции. Циклы for, while, repeat.

Тема 3. Визуализация данных и результатов их анализа в Python

Построение графиков, статическая и интерактивная визуализации. Работа с библиотекой Matplotlib. Изменение масштаба. Нанесение рисок, меток и надписей. Добавление пояснительных надписей. Аннотации и рисование в подграфике. Использование символов и цветов. Сохранение графиков в файле. Функции построения графиков в библиотеке Pandas, включая линейные и нелинейные графики функций, столбиковые диаграммы, гистограммы, графики плотностей распределения вероятностей, «ящики с усами» и диаграммы рассеяния.

Тема 4. Предсказательное моделирование в Python

Работа с библиотекой Scikit-Learn. Построение моделей, «обучающихся с учителем». Разделение данных на обучающие и тестовые. Нормировка обучающих данных. Модели классификации. Бинарная и многоклассовая классификация. Предварительная выборка атрибутов. Понижение размерности данных. Наивный байесовский классификатор. Регрессионные модели: линейная множественная регрессия, логистическая регрессия, регуляризованная регрессия. Построение моделей, «обучающихся без учителя». Подходы для построения моделей кластеризации. Анализ «рыночной корзины». Поиск ассоциативных правил. Ансамбли моделей. Применение пакета jug для работы с большими данными.

Тема 5. Анализ качества построенных моделей в Python

Определение переобученности модели. Реализация перекрестной проверки в Python. Критерии согласия модели. Оценки, основанные на хи-квадрат статистике, информационные критерии. Возможности в Python по отклонению «плохих» моделей и выбору лучших.

I. ОЦЕНИВАНИЕ

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале. Преподаватель оценивает самостоятельную работу студентов: правильность и своевременность выполнения домашних работ, задания для которых выдаются на

практических занятиях. За нарушение срока сдачи работы на 1 неделю оценка за нее снижается на 50%, на 2 недели – на 100%.

Итоговая оценка по дисциплине складывается из двух компонент:

– Три домашних задания по разделам «Визуализация данных», «Построение моделей», и «Анализ качества моделей» – 80%.

– Письменный тест на знание теоретических понятий – 20%

Итоговая оценка за курс округляется математически (например, 5,5 округляется до 6, а 5,4-до 5). Исключение составляют итоговые оценки менее 4-х баллов. В этом случае даже 3,9 округляется до 3.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примерный список типов вопросов к контрольным и зачётам по курсу:

- Дать определение понятия, которое встречается в курсе.

- Дано текстовое описание алгоритма и его реализация на языке Python с ошибкой. Найти и исправить ошибку.

- Решить с помощью языка Python и его библиотек задачу, связанную с анализом данных.

Студентам предлагается 3 домашних задания, включающих загрузку данных, разведочный анализ данных, визуализацию описательных и аналитических статистик, построение заданного типа модели, и анализ качества модели, предполагающие использование библиотек Python.

V. РЕСУРСЫ

5.1 Основная литература

Основы алгоритмизации и программирования на Python : учеб. пособие / С.Р. Гуриков. — М. : ФОРУМ : ИНФРА-М, 2017. — 343 с. — (Высшее образование: Бакалавриат). - Режим доступа: http://znanium.com/catalog/product/772265

Sneeringer, Luke. Professional Python, John Wiley & Sons, Incorporated, 2015. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4187169.

5.2. Дополнительная литература

Arbuckle, Daniel. Python Testing : Beginner's Guide, Packt Publishing Ltd, 2010. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-

ebooks/detail.action?docID=1126728.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*
3.	Python software Foundation Python	*Разрешенное*

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

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.1	База данных Европейского социального исследования ESS	URL:http://EuropeanSocialSurvey.org
2.2	База данных по Калифорнийского университета в Ирвайне по машинному обучению	URL: URL: http://archive.ics.uci.edu/ml/index.php
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.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 самостоятельных занятий по дисциплине оснащены ПЭВМ с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ