

Министерство науки и высшего образования Российской Федерации
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет промышленных технологий и дизайна»
(СПбГУПТД)

УТВЕРЖДАЮ
Директор ВШТЭ



Рабочая программа дисциплины

Б1.О.42

Работа с большими данными

Учебный план: _____ ФГОС3++b010302-1_24-14.plx

Кафедра: 16 Прикладной математики и информатики

Направление подготовки:
(специальность) 01.03.02 Прикладная математика и информатика

Профиль подготовки:
(специализация) Прикладная математика и информатика

Уровень образования: бакалавриат

Форма обучения: очная

План учебного процесса

Семестр (курс для ЗАО)		Контактная работа обучающихся		Сам. работа	Контроль, час.	Трудоё мкость, ЗЕТ	Форма промежуточной аттестации
		Лекции	Практ. занятия				
7	УП	17	34	56,75	0,25	3	Зачет
	РПД	17	34	56,75	0,25	3	
Итого	УП	17	34	56,75	0,25	3	
	РПД	17	34	56,75	0,25	3	

Рабочая программа дисциплины составлена в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки 01.03.02 Прикладная математика и информатика, утверждённым приказом Минобрнауки России от 10.01.2018 г. № 9

Составитель (и):

старший преподаватель

Леонова Н.Л.

От кафедры составителя:

Заведующий кафедрой прикладной математики и информатики

Ремизова И.В.

От выпускающей кафедры:

Заведующий кафедрой

Ремизова И.В.

Методический отдел:

1 ВВЕДЕНИЕ К РАБОЧЕЙ ПРОГРАММЕ ДИСЦИПЛИНЫ

1.1 Цель дисциплины:

формирование целостного представления о современных технологиях анализа и обработки структурированных или неструктурированных массивов данных большого объема.

1.2 Задачи дисциплины:

- приобретение необходимых знаний современных средств сбора и анализа данных в предметной области;

- изучение и освоение на практике современных программных средств анализа больших данных, таких как MS Power BI, SQL, MongoDB, Hadoop MapReduce, которые в дальнейшем студенты могут эффективно использовать в своей профессиональной деятельности

1.3 Требования к предварительной подготовке обучающегося:

Предварительная подготовка предполагает создание основы для формирования компетенций, указанных в п. 2, при изучении дисциплин:

Теория игр

Компьютерное моделирование

Программирование на языке Python

Основы системного анализа

Дискретная математика

Численные методы

Математический анализ

Теория вероятностей и математическая статистика

Информатика

Языки и методы программирования

Алгебра и геометрия

Основы алгоритмизации

2 КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ОПК-1: Способен применять фундаментальные знания, полученные в области математических и (или) естественных наук, и использовать их в профессиональной деятельности

Знать: основные понятия термина Большие Данные; жизненный цикл данных; основные способы хранения и анализа данных; методы извлечения знаний из данных (кластеризация, классификация, прогнозирование, регрессия, нейронные сети).

Уметь: анализировать современные потоки данных; находить, извлекать и структурировать данные; работать с программными средствами для хранения и анализа данных; разрабатывать и адаптировать программные компоненты работы с данными для нужд предприятия.

Владеть: терминологией Больших данных, Data Mining; основными способами хранения и анализа данных; навыками извлечения знаний из данных (кластеризация, классификация, прогнозирование, нейронные сети).

ОПК-3: Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности

Знать: архитектуру системы обработки больших данных; основные принципы приема данных, сбора данных, анализа данных.

Уметь: использовать параллельные алгоритмы для работы с данными; пользоваться возможностями лямбда-архитектуры.

Владеть: методами работы с программными платформами для больших данных; приемами обработки данных в реальном времени; навыками работы с оборудованием для больших данных

3 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Наименование и содержание разделов, тем и учебных занятий	Семестр (курс для ЗАО)	Контактная работа		СР (часы)	Форма текущего контроля
		Лек. (часы)	Пр. (часы)		
Раздел 1. Теоретические основы работы с большими данными	7				
Тема 1. Ведение в анализ больших данных Основные термины и задачи анализа больших данных; этапы анализа данных; измерения и шкалы; виды данных; источники данных; методики сбора данных.		4	6	10	
Тема 2. Методы анализа больших данных. Жизненный цикл анализа больших данных, стандарты. Когнитивный анализ данных. Визуализация больших данных. Способы понижения размерности данных. описательная статистика, преобразование и нормирование данных; работа с временными рядами: анализ, метрики качества, сглаживание, определение трендов, построение моделей для временных рядов с сезонными составляющими; интеллектуальный анализ данных: основные понятия и алгоритмы. Байесовские методы в анализе и обработке данных. Кластеризация данных.		5	10	20	О
Раздел 2. Инструментарий аналитики и обработки больших данных					
Тема 3. Технологии хранения и обработки больших данных. Технологии хранения больших данных: базы данных, модели данных; подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных.		4	6	8	
Тема 4. Современные программные средства анализа больших данных. Инструменты обработки данных; языки программирования Python и R в аналитике больших данных; NoSQL системы: основные характеристики, технологии, примеры NoSQL систем; система документного хранилища данных MongoDB: основы MongoDB и управление базами, выборка данных из коллекций, фильтры, сортировка и агрегирование, редактирование данных и индексы.		4	12	18,75	О
Итого в семестре (на курсе для ЗАО)		17	34	56,75	
Консультации и промежуточная аттестация (Зачет)		0,25			

Всего контактная работа и СР по дисциплине		51,25	56,75	
--	--	-------	-------	--

4 КУРСОВОЕ ПРОЕКТИРОВАНИЕ

Курсовое проектирование учебным планом не предусмотрено

5. ОПИСАНИЕ ПОКАЗАТЕЛЕЙ, КРИТЕРИЕВ, СИСТЕМЫ И СРЕДСТВ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

5.1 Показатели оценивания

Код компетенции	Показатели оценивания результатов обучения
ОПК-3	1. излагает базовые методы манипуляции и визуализации данных. Основные классы задач машинного обучения 2. способен обрабатывать данные, выявлять скрытые зависимости и закономерности в больших данных, извлеченных непосредственно или полученных в результате эксперимента; 3. обрабатывает результаты экспериментов с использованием современных средств работы с большими данными
ОПК-1	1. ориентируется в методиках и алгоритмах анализа и обработки больших данных с использованием современных языков программирования высокого уровня; 2. анализирует и выбирает современные информационные технологии и программные средства, в том числе отечественного производства, наиболее соответствующие решаемым задачам профессиональной деятельности; 3. способен применять современные информационные системы и технологии для реализации моделей многократного использования при обработке больших данных методами машинного и глубокого обучения.

5.2 Система и критерии оценивания

Шкала оценивания	Критерии оценивания сформированности компетенций	
	Устное собеседование	Письменная работа
Зачтено	Обучающийся: • ответил на поставленные вопросы; • выполнил практическое задание и представил результаты; возможно допуская несущественные ошибки	
Не зачтено	Обучающийся: • не выполнил практическое задание; • не ответил на вопросы преподавателя, или допустил существенные ошибки в ответе.	

5.3 Типовые контрольные задания и/или иные материалы, используемые для оценки результатов обучения

5.3.1 Типовые контрольные вопросы

№ п/п	Формулировки вопросов
Семестр 7	
1	Дайте определение и приведите примеры для термина Data science.
2	Дайте определение и приведите примеры для термина Big data.
3	Дайте определение и приведите примеры для термина Data lake.
4	Дайте определение и приведите примеры для термина Data mining.
5	Дайте определение и приведите примеры для термина Machine learning.
6	Дайте определение и приведите примеры для термина Deep learning.
7	Дайте определение и приведите примеры для термина Business intelligence.
8	Как Data Science применяется в различных сферах деятельности?
9	В чем суть управления качеством данных?
10	Суть ERP-систем. Российские ERP-системы.

11	Неспециализированные программы для обработки данных / больших данных.
12	Специализированные статистические программы
13	Языки программирования для обработки данных / больших данных.
14	Фреймворки для обработки данных / больших данных.
15	Аналитические платформы для обработки данных / больших данных.
16	Характеристики и источники больших данных. Культура сбора данных.
17	Аналитика для неструктурированных данных
18	Основы архитектуры хранения и обработки больших данных, виды обработки и масштабирования.
19	Методики сбора данных.
20	Пакетная и динамическая загрузка данных.
21	Обзор технологий хранения больших данных.
22	Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных.
23	Описательная статистика
24	Преобразование и нормирование данных.
25	Работа с временными рядами: анализ и метрики качества, сглаживание и определение трендов, построение моделей для временных рядов с сезонными составляющими.
26	Интеллектуальный анализ данных: основные понятия и алгоритмы
27	«Жизненный цикл» проекта по аналитике больших данных. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом. Перечислить основные роли исполнителей проекта
28	Охарактеризовать конструкции языка R Перечислить типы языка R, привести примеры.
29	Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.
30	Классификация и кластеризация – суть и назначение. Метрики. Постановка задачи кластеризации. Методы кластеризации на графах. Отличие от задачи классификации. Привести примеры использования алгоритмов кластеризации.
31	Парадигма Map Reduce. Описать принцип работы. Нарисовать схему. Перечислить слабые и сильные стороны. Обозначить области применимости. Привести примеры использования.
32	Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры и инструменты для визуализации
33	Научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.
34	OLAP и OLTP системы. Разница.
35	Репликация и шардинг.

5.3.2 Типовые тестовые задания

не предусмотрено

5.3.3 Типовые практико-ориентированные задания (задачи, кейсы)

Подобрать данные для таблицы проанализировать их взаимное влияние,отобразить корреляцию:

- Роста ВВП на прирост населения
- Прироста населения на динамику безработицы
- Прирост людей с высшим образованием на рост промышленного производства
- Прирост людей с высшим образованием на развитие науки
- Прирост людей с высшим образованием на динамику доходов на душу населения
- Динамику безработицы на динамику преступности
- С помощью регрессионного анализа найдите зависимые переменные и поясните влияние на них независимых переменных.
- С помощью функции `predict()` и `help()` постройте прогноз по столбцу, соответствующему варианту.

5.4 Методические материалы, определяющие процедуры оценивания результатов обучения

5.4.1 Условия допуска обучающегося к промежуточной аттестации и порядок ликвидации академической задолженности

Проведение промежуточной аттестации регламентировано локальным нормативным актом СПбГУПТД «Положение о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся»

5.4.2 Форма проведения промежуточной аттестации по дисциплине

Устная ☒ Письменная ☐ Компьютерное тестирование ☐ Иная ☐

5.4.3 Особенности проведения промежуточной аттестации по дисциплине

Возможность пользоваться справочными таблицами, калькулятором, компьютером
Время на подготовку ответа 20 минут.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1 Учебная литература

Автор	Заглавие	Издательство	Год издания	Ссылка
6.1.1 Основная учебная литература				
Воронова, Л. И., Воронов, В. И.	Big Data. Методы и средства анализа	Москва: Московский технический университет связи и информатики	2016	http://www.iprbookshop.ru/61463.html
Тарасов, А. С., Орлова, И. А., Геращенко, Е. С., Потапова, В. Ю., Никифоров, М. Б.	Статистическая обработка экспериментальных данных. Язык R в задачах планирования эксперимента	Рязань: Bookjet	2018	https://www.iprbookshop.ru/121854.html
6.1.2 Дополнительная учебная литература				
Плас Дж. Вандер	Python для сложных задач: наука о данных и машинное обучение	Санкт-Петербург: Питер	2018	https://ibooks.ru/reading.php?short=1&productid=356721
Уэс Маккинли, Слинкин А. А.	Python и анализ данных	Саратов: Профобразование	2017	http://www.iprbookshop.ru/64058.html

6.2 Перечень профессиональных баз данных и информационно-справочных систем

Электронно-библиотечная система IPRbooks [Электронный ресурс]. URL: <http://www.iprbookshop.ru/>
Электронная библиотека ВШТЭ СПб ГУПТД [Электронный ресурс]. URL: <http://nizrp.narod.ru>
Электронно-библиотечная система «Айбукс» [Электронный ресурс]. URL: <https://www.ibooks.ru/>

6.3 Перечень лицензионного и свободно распространяемого программного обеспечения

MicrosoftOfficeProfessional 2013
MicrosoftWindows 8

6.4 Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Аудитория	Оснащение
Компьютерный класс	Мультимедийное оборудование, компьютерная техника с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду
Лекционная аудитория	Мультимедийное оборудование, специализированная мебель, доска