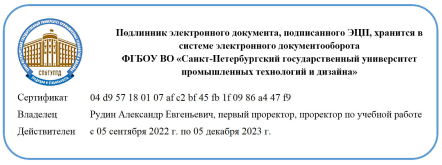
Министерство науки и высшего образования Российской Федерации
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет промышленных технологий и дизайна»
(СПбГУПТД)



УТВЕРЖДАЮ
Первый проректор, проректор по УР
_____А.Е. Рудин
«28»____06____ 2022 года

Рабочая программа дисциплины

Б1.В.14 Основы анализа данных и Data Mining

Учебный план: 2022-2023 09.03.01 ВШПМ Разр IT-сист и мультим прил ОО №1-1-55.plx

Кафедра: 21 Информационных и управляющих систем

Направление подготовки: (специальность) 09.03.01 Информатика и вычислительная техника

Профиль подготовки: (специализация) Разработка IT-систем и мультимедийных приложений

Уровень образования: бакалавриат

Форма обучения: очная

План учебного процесса

Семестр (курс для ЗАО)		Контактная работа обучающихся		Сам. работа	Контроль, час.	Трудоёмкость, ЗЕТ	Форма промежуточной аттестации
		Лекции	Практ. занятия				
7	УП	34	34	47	29	4	Экзамен, Курсовая работа
	РПД	34	34	47	29	4	
Итого	УП	34	34	47	29	4	
	РПД	34	34	47	29	4	

Санкт-Петербург
2022

Рабочая программа дисциплины составлена в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки 09.03.01 Информатика и вычислительная техника, утверждённым приказом Министерства образования и науки Российской Федерации от 19.09.2017 г. № 929

Составитель (и):

кандидат технических наук, Доцент	_____	Белая Т.И.
От кафедры составителя: Заведующий кафедрой информационных и управляющих систем	_____	Горина Елена Владимировна
От выпускающей кафедры: Заведующий кафедрой	_____	Горина Елена Владимировна

Методический отдел:

_____

1 ВВЕДЕНИЕ К РАБОЧЕЙ ПРОГРАММЕ ДИСЦИПЛИНЫ

1.1 Цель дисциплины: Сформировать у студентов компетенции в области применения технологий обработки данных (в том числе big data) и машинного обучения к решению прикладных задач, современных проблем прикладной математики и информатики, проблем обработки и анализа информации.

1.2 Задачи дисциплины:

− формирование у студентов представления о типах задач, возникающих в области интеллектуального анализа данных (Data Mining).

− изучение основных подходов и алгоритмов решения задач анализа данных и особенностей их применения к решению реальных задач.

− получение студентами навыка по выявлению, формализации и успешному решению практических задач анализа данных, возникающие в процессе их профессиональной деятельности.

− получение практического навыка в работе с существующими программными пакетами по анализу данных.

1.3 Требования к предварительной подготовке обучающегося:

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

Программирование

Философия

Социология

Основы системного анализа

Алгоритмы и структуры данных

Правоведение

Теория информации

Вычислительная математика

Языки и методы программирования

Методы программирования

Системы искусственного интеллекта

Имитационное моделирование

2 КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ПК-1: Способен осуществлять проектирование и дизайн информационных систем
Знать: теоретические и методологические основы интеллектуального анализа; методики предобработки первичных данных; особенности методов интеллектуального анализа данных; особенности методов визуального интеллектуального анализа
Уметь: выбирать средства анализа, наиболее эффективные для конкретных данных; применять методы первичной обработки данных; правильно понимать и интерпретировать полученные результаты исследования
Владеть: навыками сбора первичной организации и хранения данных для конкретного исследования; навыками самостоятельного проведения исследований

3 РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Наименование и содержание разделов, тем и учебных занятий	Семестр (курс для ЗАО)	Контактная работа		СР (часы)	Инновац. формы занятий	Форма текущего контроля
		Лек. (часы)	Пр. (часы)			
Раздел 1. Технологии анализа данных	7					О
Тема 1. Введение. Понятие анализа данных. Задачи систем поддержки принятия решений. OLTP и OLAP-системы. Принципы построения информационных хранилищ. Модели информационных хранилищ. Многомерная модель данных. Правила Кодда. Размерностные модели. MOLAP, ROLAP, HOLAP- системы. Витрины данных. ETL (ExtractingTransformatingandLoading) – средство извлечения, обработки и загрузки данных. Добыча данных. Добыча данных в управлении качеством. DataMining. Стандарты DataMining. Стандарт CWM, CRISP, PMML. Жизненный цикл процесса анализа данных. Классификация методов DataMining. Модели DataMining. Понятие данные и знания. Процесс обнаружения знаний. Классификация задач DataMining. Методы анализа данных. Разведочный анализ данных. Очистка и фильтрация данных. Статистические диаграммы. «Ящичные» диаграммы. Диаграммы «ствол-листья». Задачи классификации и регрессии. Использование статистических пакетов для интеллектуального анализа данных. Понятие бизнес-аналитики. Средства бизнес-аналитики. Средства легкой бизнес-аналитики. QlikView, QlikSence.		4		8	ИЛ	
Тема 2. Методология KDD. Задачи предобработки данных. Технология ETL. Просмотр данных. Очистка данных. Оценка качества данных. Заполнение пропущенных данных. Аномальные и предельные данные. Использование ящечной диаграммы. Выявление дубликатов и противоречий. Корреляционный анализ. Использование факторного анализа при предобработке данных. Трансформация данных. Квантование. Сэмплинг. Группировка данных.		4		8	ИЛ	
Раздел 2. Интеллектуальный анализ данных						О

Тема 3. Постановка задач кластерного анализа. Определение кластера. Параметры кластера. Меры близости. Метрики кластерного анализа. Базовые алгоритмы кластеризации. Иерархическая кластеризация. Дендограммы. Метод К-средних. Профили кластеров. Взаимосвязь кластерного и регрессионного анализа. Использование пакета Deductor для решения задач кластерного анализа. Кластерный анализ в средствах интеллектулаьного анализа MicrosoftOffice. Практическое занятие 1: Подготовка данных к анализу и обработке		8	8	8	ИЛ	
Тема 4. Основные положения непараметрической и нечисловой статистики. Таблицы сопряженности. Таблица сопряженности 2x2. Таблицы флагов и заголовков. Непараметрические и нечисловые критерии. Канонический анализ. Корреляционная матрица. Коэффициенты канонической корреляции. Меры избыточности переменных. Задачи ассоциации. Ассоциативные правила. Поддержка и достоверность ассоциативных правил. Лифт. Алгоритмы построения ассоциативных правил. Рекомендации по генерации правил. Алгоритм apriori. Использование пакета Deductor для построения ассоциативных правил. Практическое занятие 2: Решение задач корреляционного анализа		8	8	8	ИЛ	
Раздел 3. Проверка статистических гипотез						О,Пр
Тема 5. Ошибки первого и второго рода. Уровень значимости и мощность критерия. Описание гипотез и критерии их проверки. Простые и сложные гипотезы. Проверка гипотез о равенстве средних и дисперсий двух нормально распределенных генеральных совокупностей. Хи-квадрат критерий Пирсона: проверка гипотезы о соответствии наблюдаемых значений предполагаемому распределению вероятностей (дискретному или непрерывному). Проверка гипотез о вероятностной природе данных (стационарности, нормальности, независимости, однородности). Практическое занятие 3: Проверка гипотез		6	8	7	ИЛ	

Тема 6. Функциональная и статистическая зависимости. Корреляционная таблица. Групповые средние. Понятие корреляционной зависимости. Эмпирическая ковариация. Выборочный коэффициент корреляции, его свойства. Основные задачи теории корреляции: определение формы и оценка тесноты связи. Виды корреляционной связи (парная и множественная, линейная и нелинейная). Линейная корреляция. Уравнения прямых регрессии для парной корреляции. Определение параметров прямых регрессии методом наименьших квадратов. Значимость коэффициентов по критерию Стьюдента. Практическое занятие 4: Проверка зависимостей		4	10	8	ИЛ	
Итого в семестре (на курсе для ЗАО)		34	34	47		
Консультации и промежуточная аттестация (Экзамен, Курсовая работа)		4,5		24,5		
Всего контактная работа и СР по дисциплине		72,5		71,5		

4 КУРСОВОЕ ПРОЕКТИРОВАНИЕ

4.1 Цели и задачи курсовой работы (проекта): Целями написания курсового проекта являются: закрепление и углубление знаний по анализу данных, полученных студентами в рамках изучаемой дисциплины; формирование умений применять теоретические знания при решении конкретных практических задач; приобретение и закрепление навыков самостоятельной работы.

4.2 Тематика курсовой работы (проекта): Основные сферы применения Data mining.
Статистические методы анализа данных. Описательная статистика.
Статистические методы анализа данных. Проверка гипотез.
Статистические методы анализа данных. Определение размера выборки.
Основные методы выявления вероятностной природы данных.
Выявление связей и закономерностей в данных.
Основные методы анализа временных рядов и их применение при обработке данных.
Метод опорных векторов и его применение для анализа данных.
Методы построения деревьев решений и их применение для анализа данных.
Методы кластер-анализа и их применение для обработки данных.
Применение корреляционно-регрессионного анализа для решения экономических задач.
Применение факторного анализа для решения экономических задач.
Применение методов анализа временных рядов для решения экономических задач.
Применение методов построения деревьев решений в экономике.
Анализ данных с использованием баз и хранилищ данных, SQL и OLAP технологии.
Инструменты анализа данных в системе Statistica.
Инструменты анализа данных в системе Excel.
Применение методов дисперсионного анализа в экономике.
Применение ковариационного анализа для решения экономических задач.
Обзор и основные характеристики отечественных и зарубежных статистических пакетов.
Ряды распределения. Понятие, статистические характеристики ряда и способы построения.

4.3 Требования к выполнению и представлению результатов курсовой работы (проекта):
Язык и стиль изложения основного текста курсовой работы должен быть точным, логически последовательным. В курсовой работе должны обеспечиваться: - самостоятельный, творческий характер изложения; - смысловая законченность, целостность и связность текста; - точность использования основного терминологического фонда управленческой науки; - объективность, четкость и доступность изложения материала.

Курсовая работа должна иметь следующую структуру: 1) титульный лист; 2) СОДЕРЖАНИЕ; 4) ВВЕДЕНИЕ; 5) основная часть; 6) ЗАКЛЮЧЕНИЕ; 7) СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ; 8) ПРИЛОЖЕНИЯ, в том числе перечень принятых сокращений и терминов (при необходимости).

Результаты курсовой работы оформляются согласно ГОСТу 7.32-2017 «Отчет о научно-исследовательской работе. Структура и правила оформления»

Студенту предоставляется слово для доклада (время доклада – 5 мин). Приветствуется научный стиль

изложения, лаконизм и содержательность выводов по работе. В докладе должны быть отражены следующие основные моменты: - цель и задачи работы; - обоснование выбора языка и среды программирования; - изложение основных результатов работы; - краткие выводы по тем результатам работы, которые определяют ее практическую значимость, степень и характер новизны элементов. Доклад может сопровождаться презентацией (MS PowerPoint). После доклада студенту-автору работы задаются вопросы. Докладчику может быть задан любой вопрос по содержанию работы. Общая длительность защиты одной работы – не более 15 минут. Оценка за курсовой проект ставится с учетом: соответствия работы заданию; новизны результатов работы; практической значимости результатов работы; качества оформления; качества защиты работы студентом. Оценками курсового проекта могут быть: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

5. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

5.1 Описание показателей, критериев и системы оценивания результатов обучения

5.1.1 Показатели оценивания

Код компетенции	Показатели оценивания результатов обучения	Наименование оценочного средства
ПК-1	1. Называет основные понятия и виды интеллектуального анализа данных, формулирует основные принципы организации анализа данных, характеризует свойства исходных данных 2. Определяет вид анализа для конкретного исследования, подготавливает данные для исследования, формулирует гипотезу исследования 3. Использует специализированное программное обеспечения для решения конкретной задачи анализа данных, интерпретирует полученные результаты	Вопросы для устного собеседования Практические задания Курсовая работа

5.1.2 Система и критерии оценивания

Шкала оценивания	Критерии оценивания сформированности компетенций	
	Устное собеседование	Письменная работа
5 (отлично)	Полный, исчерпывающий ответ, явно демонстрирующий глубокое понимание предмета и широкую эрудицию в оцениваемой области, умение использовать теоретические знания для решения практических задач.	Работа выполнена в указанные преподавателем сроки, пояснительная записка оформлена в соответствии с требованиями, недочетов в оформлении нет. Работа написана грамотным русским языком.
4 (хорошо)	Ответ полный и правильный, основанный на проработке всех обязательных источников информации. Подход к материалу ответственный, но допущены в ответах несущественные ошибки, которые устраняются только в результате собеседования Ответ стандартный, в целом качественный, основан на всех обязательных источниках информации. Присутствуют небольшие пробелы в знаниях или несущественные ошибки.	Работа выполнена в указанные преподавателем сроки, пояснительная записка оформлена в соответствии с требованиями, содержит незначительные погрешности в оформлении Работа написана грамотным русским языком.
3 (удовлетворительно)	Ответ воспроизводит в основном только лекционные материалы, без самостоятельной работы с рекомендованной литературой. Демонстрирует понимание предмета в целом при неполных, слабо аргументированных ответах. Присутствуют неточности в ответах, пробелы в знаниях по некоторым темам, существенные ошибки, которые могут быть найдены и частично устранены в результате собеседования Ответ неполный, основанный только на лекционных материалах. При понимании сущности предмета в целом –	Работа выполнена позже указанных преподавателем сроков, пояснительная записка оформлена в соответствии с требованиями, содержит значительное количество небольших недочётов. Содержание пояснительной записки выполнено формально, низкий уровень уникальности. Работа написана грамотным русским языком.

	пробелы в знаниях сразу по нескольким темам, существенные ошибки, устранение которых в результате собеседования затруднено.	
2 (неудовлетворительно)	Неспособность ответить на вопрос без помощи экзаменатора. Незнание значительной части принципиально важных элементов дисциплины. Многочисленные существенные ошибки. Непонимание заданного вопроса. Неспособность сформулировать хотя бы отдельные концепции дисциплины. Попытка списывания, использования неразрешенных технических устройств или пользования подсказкой другого человека (вне зависимости от успешности такой попытки).	Работа не выполнена. Работа выполнена позже установленных сроков, оформлена с грубыми нарушениями требований, уровень уникальности предельно низкий. Наличие ошибок в тексте пояснительной записки. Работа списана.

5.2 Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности

5.2.1 Перечень контрольных вопросов

№ п/п	Формулировки вопросов
Семестр 7	
1	Понятие Большие данные. Роль цифровой информации в 21 веке. Проблемы анализа и обработки большого объема данных.
2	Базовые принципы обработки больших данных.
3	Определение модели. Свойства модели.
4	Аналитический подход к моделированию.
5	Информационный подход к моделированию.
6	Лица, участвующие в информационном моделировании. Общая схема анализа.
7	Определение тиражирования знаний. Процесс построения модели.
8	Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
9	Методика извлечения знаний Knowledge Discovery in Databases (KDD). Этапы KDD.
10	Data Mining. Постановка основных задач.
11	Машинное обучение. Бизнес-решения с помощью алгоритмов Data Mining.
12	Классификация ПО в области Data Mining и KDD. Типовая схема системы на базе аналитической платформы.
13	Формальная постановка задачи кластеризации. Цели кластеризации.
14	Основные шаги алгоритма k-means. Условие остановки алгоритма k-means. Преимущества и недостатки алгоритма k-means.
15	Кластеризация с помощью самоорганизующейся карты Кохонена
16	Этапы проведения классификации. Обзор методов классификации и регрессии.
17	Задачи линейной и логистической регрессии.
18	Определение дерева решений. Структура дерева решений. Выбор атрибута разбиения в узле.
19	Алгоритм ID3. Алгоритм C4.5.

5.2.2 Типовые тестовые задания

не предусмотрено

5.2.3 Типовые практико-ориентированные задания (задачи, кейсы)

1 Найдите в сети Интернет два сайта, на которых используются системы прогнозирования.

2 Найдите в сети Интернет два сайта, на которых используются рекомендательные системы.

3 Пользуясь системой SCOPUS, проанализируйте динамику количества публикаций за пять лет по направлениям Deep Learning, Big Data, Recommender Systems, Social Network Analysis.

4 Пользуясь системой SCOPUS, найдите пять публикаций с наибольшей цитируемостью публикаций за последние десять лет по направлениям Deep Learning, Big Data, Recommender Systems, Social Network Analysis.

5 Пользуясь системами SCOPUS, Web of Science, E-library (РИНЦ), выявите нескольких ведущих ученых в сфере анализа данных.

5.3 Методические материалы, определяющие процедуры оценивания знаний, умений, владений (навыков и (или) практического опыта деятельности)

5.3.1 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»

5.3.2 Форма проведения промежуточной аттестации по дисциплине

Устная	×	Письменная	×	Компьютерное тестирование		Иная	

5.3.3 Особенности проведения промежуточной аттестации по дисциплине

При проведении экзамена время, отводимое на подготовку к ответу, составляет не более 40 мин. Для выполнения практического задания обучающему предоставляется необходимая справочная информация.

Время, отводимое на защиту курсовой работы не более 20 минут.

Сообщение результатов обучающемуся производится непосредственно после устного ответа.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1 Учебная литература

Автор	Заглавие	Издательство	Год издания	Ссылка
6.1.1 Основная учебная литература				
Мельниченко, А. С.	Математическая статистика и анализ данных	Москва: Издательский Дом МИСиС	2018	http://www.iprbookshop.ru/78563.html
Любимцева, О. Л.	Блочное планирование эксперимента и анализ данных	Нижний Новгород: Нижегородский государственный архитектурно-строительный университет, ЭБС АСВ	2018	http://www.iprbookshop.ru/80885.html
Воронов, В. И., Воронова, Л. И., Усачев, В. А.	Data Mining - технологии обработки больших данных	Москва: Московский технический университет связи и информатики	2018	http://www.iprbookshop.ru/81324.html
Замятин, А. В.	Интеллектуальный анализ данных	Томск: Издательский Дом Томского государственного университета	2020	https://www.iprbookshop.ru/116889.html
6.1.2 Дополнительная учебная литература				
Чубукова, И. А.	Data Mining	Москва, Саратов: Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа	2020	https://www.iprbookshop.ru/89404.html
Брюс П., Брюс Э.	Практическая статистика для специалистов Data Science	Санкт-Петербург: БХВ-Петербург	2018	https://ibooks.ru/reading.php?short=1&productid=358886
Маккинли, Уэс, Слинкина, А.	Python и анализ данных	Саратов: Профобразование	2019	https://www.iprbookshop.ru/88752.html
Шнарева, Г. В., Пономарева, Ж. Г.	Анализ данных	Симферополь: Университет экономики и управления	2019	https://www.iprbookshop.ru/89482.html
Истомина, А. П.	Анализ данных качественных исследований	Ставрополь: Северо-Кавказский федеральный университет	2018	https://www.iprbookshop.ru/92674.html

6.2 Перечень профессиональных баз данных и информационно-справочных систем

1. https://elibrary.ru - Научная электронная библиотека eLIBRARY.RU (ресурсы открытого доступа)
2. https://www.rsl.ru - Российская Государственная Библиотека (ресурсы открытого доступа)
3. https://link.springer.com - Международная реферативная база данных научных изданий Springerlink (ресурсы открытого доступа)
4. https://zbmath.org - Международная реферативная база данных научных изданий zbMATH (ресурсы открытого доступа)

6.3 Перечень лицензионного и свободно распространяемого программного обеспечения

MicrosoftOfficeProfessional

Microsoft Windows

R

Python

Notepad++

Microsoft Visual Studio Community

Microsoft Visual Studio Code

Microsoft Visual C++ 2010 Express

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

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