

РАБОЧАЯ ПРОГРАММА
модуля(курса)
«Анализ данных и элементы машинного обучения»

1. АННОТАЦИЯ

Дисциплина «Анализ данных и элементы машинного обучения» является одной из основных дисциплин программы. Включает учебный материал, способствующий формированию у слушателей теоретических знаний и практических навыков по методологии обработки данных средствами языка Python.

Цель дисциплины – освоение методов проведения полного цикла работ по анализу данных от сбора данных до интерпретации полученных результатов и подготовки соответствующих отчетов. В строительстве. Дисциплина рассматривается как один из основных курсов программы профессиональной переподготовки «Аналитика больших данных в строительстве».

2.СОДЕРЖАНИЕ

№ п/п	Наименование модуля, разделов и тем	Содержание обучения (по темам в дидактических единицах), наименование и тематика лабораторных работ, практических занятий (семинаров), самостоятельной работы с указанием кол-ва часов, используемых образовательных технологий и рекомендуемой литературы
<i>1.</i>	<i>2.</i>	<i>3.</i>
1	Лекции	8 ч
2	Тема 1. Технологии сбора информации и больших объемов данных	Проблема нарастающего объема данных. Скорость обработки больших объёмов данных. Последовательная, параллельная и распределенная обработка данных. Масштабируемость систем хранения и обработки данных 0,5 ч
3	Тема 2. Роль математической статистики и теории вероятностей при анализе данных, средства статистического анализа данных	Типы статистических данных (числовые, ординальные, номинальные). Генеральная совокупность, выборка. Репрезентативность выборки. Способы обеспечения сопоставимости данных. Обзор существующих открытых библиотек данных. 0,5 ч
4	Тема 3. Подготовка данных к статистическому анализу	Одномерные и многомерные данные. Понятие многомерной случайной величины. Эмпирическая плотность и эмпирическая функция распределения, примеры их построения. Визуализация эмпирических распределений: построение гистограмм, графиков функций распределения. 0,5 ч
5	Тема 4. Возможности визуального анализа данных	Числовые характеристики одномерных случайных величин. Статистические числовые характеристики. Анализ данных на основе статистических числовых характеристик: характеристик центрального положения, разброса. 0,5 ч
6	Тема 5. Статистические числовые характеристики многомерных данных	Числовые характеристики многомерной случайной величины. Ковариация, коэффициент корреляции. Построение ковариационной матрицы. Корреляционный анализ. Выявление зависимости между величинами на

		основе ковариационной матрицы. 0,5 ч
7	Тема 6. Выявление линейной зависимости. Элементы корреляционного и регрессионного анализа	Простая линейная регрессия. Многомерная регрессия. Полиномиальная регрессия. Построение простой и многомерной линейной регрессии. 0,5 ч
8	Тема 7. Кластерный анализ (кластеризация) данных	Кластеризация данных, задача и методы решения (метод к-средних, иерархическая кластеризация). Определение оптимального количества кластеров. 0,5 ч
9	Тема 8. Оценивание неизвестных параметров закона распределения данных	Оценивание неизвестных параметров распределения. Точечные оценки. Параметрическое и непараметрическое оценивание. Метод максимального правдоподобия и метод моментов. 0,5 ч
10	Тема 9. Основы Python. Введение в Python и Jupyter Notebook. Python без программирования.	1.1. Основная цель - привить навыки работы в среде Jupyter Notebook и Google Colab. Язык программирования Python. Среда Jupyter Notebook. Установка дистрибутива Anaconda (или другой эквивалентной системы). Работа в Google Colab (или другой эквивалентной среде). 1.2. Python без программирования. Основная цель - дать представление и привить навыки по выполнению арифметических операций и простейших операций по обработке текста в Python (Jupyter Notebook). Целые числа, вещественные числа, строки. Функции ввода и вывода в Python. 1 ч
11	Тема 10. Основы Python. Контейнеры в языке Python. Основные конструкции языка Python.	2.1. Контейнеры в языке Python. Основная цель - дать представление о списках (list), кортежах (tuple) и словарях (dict) и привить навыки работы с ними. Использование контейнеров для обработки текстовых данных. 2.2. Основные конструкции языка Python. Основная цель - дать представление о синтаксических конструкциях языка питон для организации ветвлений и циклов и привить навыки работы с ними. Булевский (логический) тип данных. Логические выражения. Ветвления (конструкции if, if else, if elif else). Цикл for. Цикл while. Команды break и continue. 1 ч
12	Тема 11. Основы Python. Функции, модули и библиотеки в языке Python.	3.1. Функции, модули и библиотеки в языке Python. Основная цель - познакомить слушателей с концепцией функции, модуля и библиотеки, формальных параметров, областью видимости переменных и сформировать навыки по работе с функциями, модулями и библиотеками и созданию своих функций и библиотек. 1 ч
13	Тема 12. Основы Python. Работа с текстовыми данными. Работа с табличными данными.	4.1 Работа с текстовыми данными. Основная цель - привить навыки работы с простейшими функциями работы с текстовыми строками и текстовыми файлами. 4.2. Работа с табличными данными. Основная цель - дать представление об особенностях работы с табличными данными и привить навыки работы с функциями библиотеки Pandas. Библиотека Pandas и работа с табличными данными. Обзор библиотеки Pandas. Возможности библиотеки для решения задач обработки данных. 1 ч

14	Практические занятия (семинары)	Библиотеки данных для анализа. Начало работы с Python. Чтение и обработка Excel файлов средствами Python. 2 ч
		Визуализация эмпирических распределений средствами Python. 2 ч
		Анализ многомерных данных: выявление зависимости. 2 ч
		Инструменты Python для построения регрессии. 4 ч
		Средства Python кластеризации данных. 4 ч
		Оценивание неизвестных параметров закона распределения данных. 2 ч
16	Самостоятельная работа	Выполнение домашних заданий по теме занятия. 18 ч
17	Промежуточная аттестация	Устный опрос 2 ч

3. ОЦЕНКА КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ МОДУЛЯ

Промежуточная аттестация представляет собой дифференцированный зачет в форме проведения лабораторной работы и устного опроса по её итогам.

Примерный перечень вопросов к зачету:

- Перечислить способы представления выборочных значений.
 - Назовите основные библиотеки Python, используемые для сбора, анализа данных.
 - Дать определение одномерных и многомерных данных.
 - Дать определение статистической (выборочной, эмпирической) функции распределения
 - С помощью каких графиков в Python визуализируются многомерные данные?
 - Определить, основные статистические числовые характеристик (выборочное среднее, дисперсию и стандарт)
 - Привести способы вычисления статистической ковариации и выборочного коэффициента корреляции, определить понятие несмещенной выборочной ковариации
 - Какая библиотека в Python содержит функции для вычисления ковариации и коэффициента корреляции
 - Перечислить способы выявления статистическая зависимости двух случайных величин
 - Определить простую линейную регрессию. Перечислить Ключевые различия между корреляцией и линейной регрессией
 - Описать метод использования регрессии для прогнозирования. Определить, что такое подогнанные значения и остатки
 - Какими средствами в Python строится линейная регрессия?
 - Перечислить способы диагностики качества регрессионной модели
 - Какими средствами в Python проверяется качество построенной регрессионной модели?
 - Перечислить критерии качества статистических оценок
 - Описать Метод моментов
 - Описать Метод максимального правдоподобия
 - Какими средствами реализуются в Python реализуется метод моментов?
 - Описать метод интервального оценивания неизвестных параметров распределения.
 - Описать методологию построения доверительных интервалов для математического ожидания
 - Описать методологию построения доверительных интервалов для неизвестной дисперсии

- Описать методологию построения доверительных интервалов для неизвестной вероятности события

Формы и методы контроля и оценки результатов освоения модуля

№ п/п	Наименование процедуры	Основные показатели оценки	Формы и методы контроля и оценки
1	Промежуточный контроль. Модуль 4. Обработка больших данных в строительстве	Владеет навыками использования инструментов и технологий для анализа больших данных в строительстве На высоком/среднем/низком уровне	Дифференцированный Зачет/Лабораторная работа

Критерии оценки

№ п/п	Наименование процедуры	Основные показатели оценки	Формы и методы контроля и оценки
1	Промежуточная аттестация. Модуль 4. Обработка и анализ больших данных в строительстве	Зачтено. Уровень знаний в объеме, соответствующем программе подготовки. Фрагментарные, либо сформированные, но содержащие отдельные пробелы знания об инструментах и технологиях для анализа больших данных в строительстве Не зачтено. Уровень знаний ниже минимальных требований. Имели место грубые ошибки. Отсутствие знаний теоретического материала об инструментах и технологиях для анализа больших данных в строительстве	Дифференцированный Зачет/Лабораторная работа

4. УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ МОДУЛЯ

4.1 Учебно-методическое и информационное обеспечение программы:

Для эффективного освоения компетенций, формируемых учебной дисциплиной важно использование в учебном процессе активных и интерактивных форм проведения занятий.

Изучение учебной дисциплины предполагает наличие аудиторной и самостоятельной видов работ слушателей. В ходе практических занятий рассматриваются примеры работы с анализом больших данных в строительстве, практические задачи, наиболее сложные ситуации из практики с целью наиболее полного овладения умениями и навыками.

Лекции по учебной дисциплине призваны формировать знания, предусмотренные учебной программой, и включают теоретическую подготовку к процессу обработки и анализа больших данных в строительстве, на базе которой строятся прикладные аспекты.

Освоение дисциплины предполагает значительный объем самостоятельной внеаудиторной работы, которую слушатели должны выполнять как индивидуально, так и в малых группах. Наряду с проработкой основной литературы (глав базового учебника) предусмотрено самостоятельное чтение дополнительной литературы (статей и других научных публикаций), а также проведение анализа кейсов, которые обсуждаются в ходе дискуссий на практических занятиях.

Практические занятия в малых группах и самостоятельная внеаудиторная работа направлены на выработку навыков использования аналитики больших данных в строительстве и формирования профессиональных компетенций, установленных в соответствии с целями и задачами дисциплинами.

Для достижения поставленных целей преподавания дисциплины реализуются следующие средства, способы и организационные мероприятия:

- изучение теоретического материала дисциплины на лекции с использованием компьютерных технологий;
- самостоятельное изучение теоретического материала дисциплины с использованием Internet-ресурсов, информационных баз, электронных библиотек, методических разработок, специальной и научной литературы;
- закрепление теоретического материала при проведении практических занятий с использованием учебного и научного оборудования, выполнения проблемно-ориентированных, поисковых, творческих заданий.

Самостоятельная работа слушателей включает:

1. Изучение учебной литературы по курсу.
2. Решение практических ситуаций и задач
3. Изучение источников управленческой информации
4. Работу с ресурсами Интернет
5. Решение практических ситуаций в виде творческих заданий
6. Изучение практических материалов деятельности конкретных предприятий
7. Изучение статистикой информации
8. Подготовку к зачету.

Цель самостоятельной работы - подготовка современного компетентного специалиста и формирование способностей и навыков к непрерывному самообразованию и профессиональному совершенствованию.

4.2. Содержание комплекта учебно-методических материалов.

Слушателям предоставляются презентации и раздаточный материал.

4.3. Используемые образовательные технологии. Краткое описание.

Применяются методы активного и интерактивного обучения.

4.4. Литература

Основная литература:

1. Федоткин М.А. Основы прикладной теории вероятностей и статистики. — М.: Высшая школа. 2006. - 168 с.
2. Теория вероятностей и математическая статистика. Авторы: Федоткин М.А., Пройдакова Е.В.: Электронный управляемый курс. – Нижний Новгород: Нижегородский госуниверситет, 2014. (Идентификационный номер в электронном каталоге фонда электронных образовательных ресурсов ННГУ: 789Е.14.08)
3. Золотых Н.Ю. Машинное обучение. Курс лекций. Нижний Новгород: ННГУ, 2007. <http://www.uic.nnov.ru/~zny/m>
4. Курс «Машинное обучение» – <https://www.intuit.ru/studies/courses/13844/1241/info>
5. Практическая статистика для специалистов Data Science: Пер. с англ./ П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018..

Программное обеспечение и Интернет-ресурсы

1. Операционная система Microsoft Windows.
 2. Open-source среда Spyder.
 3. Веб-интерактивная вычислительная среда Jupyter Notebook (для поддержки языка Рут)
- Лекции и практические занятия проводятся с использованием возможностей мультимедийного класса. Использование в учебном процессе активных и интерактивных форм проведения занятий (компьютерных симуляций, деловых и ролевых игр, разбор конкретных ситуаций).

Дополнительная литература и источники:

1. Макшанов А. В., Журавлев А. Е., Тындыкаръ Л. Н. Большие данные. Big Data [Электронный ресурс]: - Санкт-Петербург: Лань, 2021. - 188 с. – Режим доступа: <https://e.lanbook.com/book/165835>
2. Парамонов И. Ю., Смагин В. А., Косых Н. Е., Хомоненко А. Д. Методы и модели исследования сложных систем и обработки больших данных [Электронный ресурс]: монография. - Санкт-Петербург: Лань, 2020. - 236 с. – Режим доступа: <https://e.lanbook.com/book/126938>
3. Железнов М. М. Методы и технологии обработки больших данных [Электронный ресурс]: учебно-методическое пособие. - Москва: МИСИ – МГСУ, 2020. - 46 с. – Режим доступа: <https://e.lanbook.com/book/145102>
1. Рындина С. В. Цифровая трансформация бизнеса: использование аналитики на основе больших данных [Электронный ресурс]: учебное пособие. - Пенза: ПГУ, 2019. - 182 с. – Режим доступа: <https://e.lanbook.com/book/162301>

4.5. Материально-техническое обеспечение

№ п.п.	Наименование модуля (тем, разделов)	Материально-техническая база Материально-технические условия для реализации программ (наличие лабораторий, производственных участков и т.п. по профилю программы профессиональной переподготовки)
	<p>Тема 1. Технологии сбора информации и больших объемов данных</p> <p>Тема 2. Роль математической статистики и теории вероятностей при анализе данных, средства статистического анализа данных</p> <p>Тема 3. Подготовка данных к статистическому анализу</p> <p>Тема 4. Возможности визуального анализа данных</p> <p>Тема 5. Статистические числовые характеристики многомерных данных</p> <p>Тема 6. Выявление линейной зависимости. Элементы корреляционного и регрессионного анализа</p> <p>Тема 7. Кластерный анализ (кластеризация) данных</p> <p>Тема 8. Оценивание неизвестных параметров закона распределения данных</p> <p>Тема 9. Основы Python. Введение в Python и Jupyter Notebook. Python без программирования.</p> <p>Тема 10. Основы Python. Контейнеры в языке Python. Основные конструкции языка Python.</p> <p>Тема 11. Основы Python. Функции, модули и библиотеки в языке Python.</p> <p>Тема 12. Основы Python. Работа с текстовыми данными. Работа с табличными данными.</p>	<p>Помещения представляют собой учебные аудитории для проведения учебных занятий, предусмотренных программой, оснащенные оборудованием и техническими средствами обучения: компьютер преподавателя с возможностью подключения к сети Интернет, экран для демонстрации и проектор, компьютеры для студентов с возможностью подключения к сети Интернет.</p> <p>Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду организации.</p> <p>Каждый обучающийся в течение всего периода обучения обеспечен индивидуальным неограниченным доступом к электронно-библиотечным системам (электронным библиотекам) («Консультант студента», «Лань», «Znanium», «Юрайт») и к электронной информационно-образовательной среде организации (portal.unn.ru), в системе электронного обучения ННГУ https://e-learning.unn.ru/.</p> <p>Данные электронно-библиотечные системы (электронные библиотеки) и электронная информационно-образовательная среда обеспечивают возможность доступа обучающегося из любой точки, в которой имеется доступ к информационно-телекоммуникационной сети «Интернет», как на территории организации (в библиотеке ИЭП ННГУ), так и вне ее.</p>