Приложение 4
к образовательной программе

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Б1.О.02.ДВ.05.01 Разведывательный анализ данных
(индекс, наименование дисциплины в соответствии с учебным планом)

09.03.03 Прикладная информатика
(код, наименование направления подготовки/специальности)

Прикладная информатика в управлении корпоративными информационными системами
(наименование образовательной программы)

Очная форма обучения
(форма обучения)

Год набора – 2026
Донецк

Автор(ы)-составитель(и) РПД:

Литвак Елена Геннадиевна, доцент кафедры информационных технологий

Заведующий кафедрой:

Брадул Наталья Валерьевна, канд. физ.-мат. наук, заведующий кафедрой информационных технологий

Рабочая программа дисциплины Б1.О.02.ДВ.05.01 Разведывательный анализ данных одобрена на заседании кафедры информационных технологий факультета государственной службы и управления Донецкого филиала РАНХиГС.

Протокол № 7 от «05 » марта 2026 г.

СОДЕРЖАНИЕ

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Дисциплина Б1.О.02.ДВ.05.01 Разведывательный анализ данных обеспечивает формирование у обучающихся следующих общепрофессиональных компетенций*:

ОТФ /ТФ и реквизиты ПС *(при наличии)* **	**Код компетенции ****	**Наименование Компетенции ****	**Код индикатора достижения компетенций ****	**Наименование индикатора достижения компетенций ****	**Образовательный результат ****
-	ОПК-2	Способен понимать принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности	ОПК-2.4	Понимает принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности	ОПК-2.4. З-1 **Знает** методы разведывательного анализа данных и визуализации. ОПК-2.4. У-1 **Умеет** выполнять анализ данных и визуализировать результаты.

** Дисциплина может формировать компетенцию полностью или частично.*
***Должно соответствовать Приложению 1 к образовательной программе*

2. Объем и место дисциплины в структуре образовательной программы

Общий объем дисциплины:

2,00 з.е., 72 ак.час

Контактная работа обучающихся с преподавателем по видам учебных занятий: 36 ак. час на контактную работу с преподавателем, из них 18 ак. час на лекции и 18 ак. час на практические занятия. 32 ак. час на самостоятельную работу обучающихся.

Б1.О.02.ДВ.05.01 Разведывательный анализ данных

реализуется на 4-м семестре 2-го курса после изучения дисциплин:

- Программирование на Python.

.

3. Содержание и структура дисциплины

3.1. Структура дисциплины

Очнаяформа обучения

№ п/п	Наименование тем и (или) разделов		Объем дисциплины, ак.час												Форма текущего контроля успеваемости, промежуточной аттестации
		ВСЕГО	Контактная работа обучающихся с преподавателем по видам учебных занятий									Самостоятельная работа			
			Период теоретического обучения						Период промежуточной аттестации (сессия)						
			Занятия лекционного типа		Занятия семинарского типа		ИК	КСР	КЭ	Каттэк	Контроль	СРкр	СРэк	СР	
			Л	ВЛ	ЛР	ПЗ									
РАЗДЕЛ 1. ФУНДАМЕНТ EDA: СТАТИСТИКА И АНАЛИЗ СВЯЗЕЙ															
Тема 1	Жизненный цикл ML-модели и роль EDA в машинном обучении	12	4	0	0	4	0	0	0	0		0	0	4	Контрольные вопросы, практические занятия, КТ1
Тема 2	Математическая статистика в контексте	20	4	0	0	4	0	0	0	0	0	0	0	12	Контрольные вопросы, практические занятия, КТ 1

	EDA														
РАЗДЕЛ 2. ПРИКЛАДНОЙ EDA: ПРИЗНАКИ, ПРЕОБРАЗОВАНИЯ И ПРОВЕРКА ГИПОТЕЗ															
Тема 3	**Проектирование признаков**	16	4	0	0	4	0	0	0	0	0	0	0	8	Контрольные вопросы, практические занятия, КТ 2
Тема 5	**Статистические гипотезы в контексте EDA**	20	6	0	0	6	0	0	0	0	0	0	0	8	Контрольные вопросы, практические занятия, КТ 2
Промежуточная аттестация		0	0	0	0	0	0	0		4	0		0	0	Зачет
Итого		**72**	**18**	**0**	**0**	**18**	**0**	**0**	**0**	**4**	**0**		**0**	**32**	

Используемые сокращения:

Л – лекции - занятия, предусматривающие преимущественную передачу учебной информации обучающимся педагогическими работниками организации и (или) лицами, привлекаемыми организацией к реализации образовательных программ на иных условиях,).

ВЛ – видео лекции.

ЛР – лабораторные работы.

ПЗ – практические занятия (за исключением лабораторных работ).

ИК – индивидуальные консультации.

КСР – контроль самостоятельной работы

КЭ – консультации перед экзаменом

Каттэк – контактная работа на аттестацию в период экзаменационных сессий

Контроль - контактная работа на аттестацию в период экзаменационных сессий для заочной формы обучения

СРкр – самостоятельная работа на подготовку курсовой работы/ курсового проекта.

СРэк – самостоятельная работа на подготовку к экзамену.

СР – самостоятельная работа в семестре на подготовку к учебным занятиям.

3.2. Содержание дисциплины

РАЗДЕЛ 1. ФУНДАМЕНТ EDA: СТАТИСТИКА И АНАЛИЗ СВЯЗЕЙ

Тема 1. Жизненный цикл ML-модели и роль EDA в машинном обучении. ОПК-2.4.

Понятие жизненного цикла ML (CRISP-DM, TDSP). Место EDA между сбором данных и моделированием. Алгоритм и методы EDA. Что такое проектирование, отбор, кодирование признаков (обзорно).

Тема 2. Математическая статистика в контексте EDA. ОПК- 2.4.

Описательная статистика (mean, median, mode, std, IQR) в Python. Корреляция: типы (линейная, нелинейная, ложная). Корреляция Пирсона. Ранговые корреляции (Спирмен, Кендалл). Визуализация корреляций (heatmap). График рассеивания (scatter plot). Парные отношения (pairplot).

РАЗДЕЛ 2. ПРИКЛАДНОЙ EDA: ПРИЗНАКИ, ПРЕОБРАЗОВАНИЯ И ПРОВЕРКА ГИПОТЕЗ

Тема 3. Проектирование признаков. ОПК-2.4.

Создание признаков (feature engineering). Внешние источники данных (обогащение). Работа с признаками типа «дата-время» (извлечение года, месяца, дня недели, разницы дат). Методы кодирования признаков (One-Hot, Label, Target Encoding). Преобразование признаков: нормализация (MinMax) и стандартизация (StandardScaler). Отбор признаков (фильтрация, оберточные методы, встроенные). Мультиколинеарность (VIF, корреляционная матрица).

Тема4. Статистические гипотезы в контексте EDA. ОПК-2.4

Необходимость статистических тестов. Критерии выбора теста (тип данных, распределение, зависимость выборок). Статистическая значимость (p-value, α-уровень). Статистическая гипотеза (H0, H1). Проверка на нормальность (Shapiro–Wilk, Колмогорова–Смирнова). Параметрические тесты (t-test, ANOVA). Непараметрические тесты (Mann–Whitney, Kruskal–Wallis). Статистические тесты для категориальных признаков (χ^2-квадрат, точный тест Фишера).

4. Типы оценочных материалов, показатели и критерии оценивания

Оценочные материалы по дисциплине Б1.О.02.ДВ.05.01

Разведывательный анализ данных входят в состав оценочных материалов по образовательной программе. Совокупность оценочных материалов по всем дисциплинам (модулям) образовательной программы составляет фонд оценочных средств (далее – ФОС). ФОС используется при проведении текущего контроля успеваемости и промежуточной аттестации обучающихся с целью оценивания достижения обучающимися планируемых результатов обучения.

4.1. ФОС разработан как комплекс проверочных заданий различного типа и уровня сложности, включает критерии и шкалы оценивания, а также «ключи» правильных ответов. ФОС формируется как отдельный документ и хранится в электронном виде, доступ к ФОС предоставлен ограниченному кругу лиц.

4.2. Для самостоятельной работы обучающихся при подготовке к текущему контролю успеваемости и промежуточной аттестации в рабочих программах дисциплин размещены типовые проверочные задания, которые можно условно разделить на задания закрытого, комбинированного и открытого типов.

Задания закрытого типа – это тестовые задания, в которых каждый вопрос сопровождается готовыми вариантами ответов, из которых необходимо выбрать один или несколько правильных.

Задания комбинированного типа – это тестовые задания, в которых каждый вопрос сопровождается готовыми вариантами ответов, из которых необходимо выбрать один или несколько правильных и обосновать свой выбор.

Задания открытого типа – это задания, в которых на каждый вопрос должен быть предложен развернутый обоснованный ответ.

В зависимости от типа задания рекомендованы определенная последовательность выполнения и система оценивания выполнения заданий.

4.3. Типы заданий, сценарии выполнения, критерии оценивания

ТИП ЗАДАНИЯ	ИНСТРУКЦИЯ	СЦЕНАРИИ ВЫПОЛНЕНИЯ	КРИТЕРИИ ОЦЕНИВАНИЯ
Задание закрытого типа с выбором одного правильного ответа из нескольких вариантов предложенных	Прочитайте текст, выберите правильный ответ	1. Внимательно прочитать текст задания и понять, что в качестве ответа ожидается только один из предложенных вариантов. 2. Внимательно прочитать предложенные вариант-ты ответа. 3. Выбрать один верный ответ. 4. Записать только номер (или букву) выбранного варианта ответа (например, 3 или В).	Ответ считается верным, если правильно указана цифра или буква
Задание закрытого типа на установление соответствия	Прочитайте текст и установите соответствие	1. Внимательно прочитать текст задания и понять, что в качестве ответа ожидаются пары элементов. 2. Внимательно прочитать оба списка: список 1 – вопросы, утверждения, факты, понятия и т.д.; список 2 – утверждения, свойства объектов и т.д. 3. Сопоставить элементы списка 1 с элементами списка 2, сформировать пары элементов. 4. Записать попарно буквы и цифры (в зависимости от задания) вариантов ответа (например, А1 или Б4).	Ответ считается верным, если правильно указаны цифры или буквы

Задание закрытого типа с выбором нескольких правильных ответов из нескольких вариантов предложенных	Прочитайте текст, выберите правильные ответы	1. Внимательно прочитать текст задания и понять, что в качестве ответа ожидается несколько правильных ответов из предложенных вариантов. 2. Внимательно прочитать предложенные вариант-ты ответа. 3. Выбрать несколько правильных ответов. 4. Записать только номера (или буквы) выбранного варианта ответа (например, 1 4 или А Г).	Ответ считается верным, если правильно установлены все соответствия (позиции из одного столбца верно сопоставлены с позициями другого)

Задание закрытого типа на установление последовательности	Прочитайте текст и установите последовательность	1. Внимательно прочитать текст задания и понять, что в качестве ответа ожидается последовательность элементов. 2. Внимательно прочитать предложенные варианты ответа. 3. Построить верную последовательность из предложенных элементов. 4. Записать буквы/цифры (в зависимости от задания) вариантов ответа в нужной последовательности (например, БВА или 135).	Ответ считается верным, если правильно указана вся последовательность цифр
Задание комбинированного типа с выбором одного правильного ответа из предложенных и обоснованием выбора	Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа	1. Внимательно прочитать текст задания и понять, что в качестве ответа ожидается только один из предложенных вариантов. 2. Внимательно прочитать предложенные варианты ответа. 3. Выбрать один верный ответ. 4. Записать только номер (или букву) выбранного варианта ответа. 5. Записать аргументы, обосновывающие выбор ответа (например, 4 текст обоснования).	Ответ считается верным, если правильно указана цифра или буква и приведены корректные аргументы, используемые при выборе ответа

Задание открытого типа с развернутым ответом	Прочитайте текст и запишите развернутый обоснованный ответ	1. Внимательно прочитать текст задания и понять суть вопроса. 2. Продумать логику и полноту ответа. 3. Записать ответ, используя четкие компактные формулировки. 4. В случае расчетной задачи, записать решение и ответ	Ответ считается верным: 1. Отсутствие фактических ошибок. 2. Раскрытие объема используемых понятий (полнота ответа). 3. Обоснованность ответа (наличие аргументов). 4. Логическая последовательность излагаемого материала.

4.4. Общая шкала оценивания результатов текущего контроля успеваемости и промежуточной аттестации обучающихся с применением БРС

Оценка по шкале ECTS	**Сумма баллов за все виды учебной деятельности**	**Оценка по государственной шкале**	**Определение**
A	90 – 100	«Отлично»	отличное выполнение с незначительным количеством неточностей
B	80 – 89	«Хорошо»	в целом правильно выполненная работа с незначительным количеством ошибок (до 10%)
C	75 – 79		в целом правильно выполненная работа с незначительным количеством ошибок (до 15%)
D	70 – 74	«Удовлетворительно»	неплохо, но со значительным количеством недостатков
E	60 – 69		выполнение удовлетворяет минимальные критерии
FX	35 – 59	«Не удовлетворительно»	с возможностью повторной сдачи
F	0 – 34		с обязательным повторным изучением дисциплины (выставляется комиссией)

Соотношение баллов за текущий контроль успеваемости и промежуточную аттестацию, а также повторную промежуточную аттестацию:

Максимальная сумма баллов за текущий контроль успеваемости	Максимальная сумма баллов за промежуточную аттестацию	Максимальная итоговая балльная оценка	Максимальная сумма баллов за повторную промежуточную аттестацию
100 баллов	100 баллов	100 баллов	100 баллов

5. Формы аттестации, типовые оценочные материалы для текущего контроля успеваемости обучающихся, критерии и шкалы оценивания по контрольным точкам

5.1. В ходе реализации дисциплины Б1.О.02.ДВ.05.01 Разведывательный анализ данных используются следующие формы текущего контроля успеваемости обучающихся (в том числе, задания к контрольным точкам):

Контрольные вопросы для проведения опроса, задания открытого типа на практических занятиях, контрольные задания.

Таблица 5.1.

Распределение баллов по видам учебной деятельности (БРС)

Раздел/Темы	Формы текущего контроля		КТ
	УО	ПЗ	
Р-1. / Т-1	10	10	10
Р-1. / Т-2	10	10	
Р-2. / Т-3	10	10	10
Р-2. / Т-4	10	10	
Итого: 100 б	40	40	20

УО – устный опрос;
ТЗ – тестовое задание;
КЗ – контрольные задания;
ПЗ – практическое занятие;
Д – доклад;
КТ – контрольные точки.

Критерии оценивания опроса:

Балы	Описание критерия
9-10	Обучающийся полно излагает материал (отвечает на вопрос), дает правильное определение основных понятий; обнаруживает понимание материала, может обосновать свои суждения, применить знания на практике, привести необходимые примеры не только из учебника, но и самостоятельно составленные; излагает материал последовательно и правильно с точки зрения норм литературного языка.
6-8	Обучающийся дает ответ, удовлетворяющий тем же требованиям, что и для оценки «отлично», но допускает 1–2 ошибки, которые сам же исправляет, и 1–2 недочета в последовательности и языковом оформлении излагаемого.
4-5	Обучающийся обнаруживает знание и понимание основных положений данной темы, но излагает материал неполно и допускает неточности в определении понятий или формулировке правил; не умеет достаточно глубоко и доказательно обосновать свои суждения и привести свои примеры; излагает материал непоследовательно и допускает ошибки в языковом оформлении излагаемого.
0-3	Обучающийся обнаруживает незнание вопроса, допускает ошибки в формулировке определений и правил, искажающие их смысл, беспорядочно и неуверенно излагает материал.

0* - в журнал академической группы не выставляется

Критерии оценивания практических занятий:

Балы	Описание критерия	
9-10	Свыше 90% правильных ответов.	Обучающийся демонстрирует глубокое познание в освоенном материале.
6-8	Свыше 70% правильных ответов.	Обучающимся материал освоен полностью, без существенных ошибок.

4-5	Реализовано более 50% поставленных задач	Обучающимся материал освоен не полностью, имеются значительные пробелы в знаниях.
0-3	Реализовано менее 30% поставленных задач.	Обучающимся материал не освоен, знания обучающегося ниже базового уровня.

0* - в журнал академической группы не выставляется

Критерии оценивания контрольных заданий:

Балы	Описание критерия
9-10	Обучающимся задание выполнено без ошибок и в полном объеме.
7-8	Обучающимся в целом задание выполнено, имеются отдельные неточности или недостаточно полные ответы, не содержащие ошибок.
5-6	Обучающимся допущены отдельные ошибки при выполнении задания
0-4	У обучающегося отсутствуют ответы на большинство вопросов задачи, задание не выполнено или выполнено не верно.

0* - в журнал академической группы не выставляется

5.2. Типовые оценочные материалы для текущего контроля успеваемости обучающихся (вне контрольных точек):

РАЗДЕЛ 1. ФУНДАМЕНТ EDA: СТАТИСТИКА И АНАЛИЗ СВЯЗЕЙ

Тема 1. Жизненный цикл ML-модели и роль EDA в машинном обучении. ОПК-2.4.

1. Дайте определение EDA (разведывательному анализу данных) в соответствии с материалом модуля.
2. Перечислите четыре основных компонента, которые включает в себя EDA согласно модулю.
3. Восстановите правильную последовательность этапов жизненного цикла машинного обучения, начиная с «Постановка проблемы».
4. Соотнесите профессию и её основную задачу (дайте определение каждой роли одной фразой):
5. Назовите три инструмента автоматической визуализации и команды для их установки (pip), упомянутые в модуле.
6. Между какими двумя этапами жизненного цикла ML находится этап EDA, и что с ним происходит на этих этапах?

Тема 2. Математическая статистика в контексте EDA. ОПК- 2.4.

1. Дайте определения следующим понятиям в соответствии с материалом

модуля: математическая статистика; Статистические данные

2. С какой целью применяется описательная статистика согласно модулю?

3. Заполните таблицу: для каждой меры центральной тенденции дайте краткое описание.

Мера	Описание
Среднее	
Медиана	
Мода	

4. Дайте характеристику трём видам корреляционной связи по знаку коэффициента корреляции:

Положительная

Отрицательная

Нулевая

5. Соотнесите метод корреляции с его описанием и параметром method:

Метод	Параметр method	Описание
Пирсона		
Спирмена		
Кендалла		

6. По коэффициенту корреляции определите силу связи (заполните пропуски):

Диапазон коэффициента корреляции	Сила связи
0 – ±0.3	
±0,3 – ±0,5	
±0,5 – ±0,7	
±0,7 – ±0,9	
±0,7 – ±1 (по тексту: ±0,7 – ±1)	

7. Какие три функции (метода) библиотеки Seaborn можно использовать для визуализации корреляций?

РАЗДЕЛ 2. ПРИКЛАДНОЙ EDA: ПРИЗНАКИ, ПРЕОБРАЗОВАНИЯ И ПРОВЕРКА ГИПОТЕЗ

Тема 3. Проектирование признаков. ОПК-2.4.

1. Дайте определение следующим понятиям согласно материалу модуля: Создание признаков, Внешние источники данных (обогащение).
2. Какие операции с признаками типа «дата-время» перечислены в модуле? Назовите не менее трех.
3. Перечислите три метода кодирования категориальных признаков, указанных в модуле, и дайте их краткую характеристику.
4. В чем разница между нормализацией (MinMax) и стандартизацией (StandardScaler)?
5. Назовите три типа методов отбора признаков, перечисленных в модуле, и приведите по одному примеру для каждого.
6. Что такое мультиколинеарность и какие два инструмента для её обнаружения упомянуты в модуле?
7. Расположите следующие этапы работы с признаками в логическом порядке, соответствующем типичному конвейеру EDA (от начала к концу):

нормализация / стандартизация;
создание новых признаков;
кодирование категориальных признаков;
отбор признаков;
обнаружение и устранение мультиколинеарности.

Тема4. Статистические гипотезы в контексте EDA. ОПК-2.4

1. С какой целью применяются статистические тесты в анализе данных? (Почему они необходимы?)
2. Назовите три критерия, которые влияют на выбор статистического теста согласно модулю.
3. Дайте определение понятиям «статистическая значимость», «p-value» и «α-уровень». Как принимается решение о гипотезе на их основе?
4. Что такое H_0 (нулевая гипотеза) и H_1 (альтернативная гипотеза)? Приведите пример для сравнения средних двух групп.
5. Какие два метода проверки распределения на нормальность упомянуты в модуле? Для чего используется эта проверка?
6. Заполните таблицу: для каждого теста укажите, к какому типу он относится (параметрический / непараметрический) и для каких данных / задач применяется.

Тест	Тип	Для каких данных / задач

Тест	Тип	Для каких данных / задач
t-test		
ANOVA		
Mann–Whitney		
Kruskal–Wallis		

7. Какие два статистических теста для категориальных признаков перечислены в модуле? В каких ситуациях они применяются?

8. Расположите следующие шаги проверки статистической гипотезы в правильном порядке:

Вычислить p-value

Сформулировать H_0 и H_1

Выбрать статистический тест

Задать α-уровень

Сравнить p-value с α и принять решение

5.3. Один или несколько тематических блоков дисциплины завершаются контрольной точкой по разделу (далее – КТ). Текущий контроль успеваемости по дисциплине предусматривает не менее 2 (двух) и не более 10 (десяти) КТ в течение периода освоения дисциплины.

Максимальное количество баллов за любой тип работ в рамках КТ составляет 100 (сто) баллов.

Распределение весовых коэффициентов по КТ в рамках текущего контроля успеваемости по дисциплине и формулы расчета:

Наименование контрольной работы	Максимальное количество баллов за работу в рамках КР, которое может набрать студент	Коэффициент веса контрольной работы	Результат контрольной работы, участвующий в формировании итоговой балльной оценки по дисциплине
КТ 1	100	0,1	10
КТ 2	100	0,1	10
Итого:	x	0,2	20

Формула расчета результата контрольной работы:

Результат контрольной работы = Количество баллов за точку в рамках КТ X Коэффициент веса контрольной точки.

5.4. Формы текущего контроля успеваемости обучающихся в рамках КТ и типовые оценочные материалы:

КТ 1

Задание 1.

Скачайте датасет про винные обзоры: https://disk.yandex.ru/d/fUtd74Y-HVT2WA

Описание датасета:

country — страна-производитель вина.
description — подробное описание.
designation — название виноградника, где выращивают виноград для вина.
points — баллы, которыми *WineEnthusiast* оценил вино по шкале от 1 до 100.
price — стоимость бутылки вина.
province — провинция или штат.

region_1 — винодельческий район в провинции или штате (например Напа).
region_2 — конкретный регион. Иногда в пределах винодельческой зоны указываются более конкретные регионы (например Резерфорд в долине Напа), но это значение может быть пустым.
taster_name — имя сомелье.
taster_twitter_handle — твиттер сомелье.
title — название вина, которое часто содержит год и другую подробную информацию.
variety — сорт винограда, из которого изготовлено вино (например Пино Нуар).
winery — винодельня, которая производила вино.

Прочитайте файл с винными обзорами:

```
data = pd.read_csv('wine.csv')
```

1. Сколько всего дегустаторов приняло участие в винных обзорах? (Ответ: 19)
2. Какова максимальная цена за бутылку в этом наборе данных?
3. Проанализируйте представленный датасет и перечислите все числовые признаки через запятую.
4. Проанализируйте датасет на наличие дублирующихся винных обзоров. Если дублирующиеся записи есть, удалите их. Ответьте, сколько дублирующихся записей вам удалось обнаружить.
5. Проверьте датасет на наличие пропусков в данных. В каких из представленных признаках были обнаружены пропуски?
6. Обработайте пропущенные значения в наборе данных любым известным вам способом, который вы изучили в модуле PYTHON-14. Очистка данных.

Воспользуйтесь правилами:

Если какой-то из признаков имеет более 30-40 % пропусков, лучше избавьтесь от него: его заполнение может привести к сильному искажению общего распределения, а удаление записей — к большой утрате данных.

Заполняйте данные с умом! Если перед вами количественный признак, то использование нецелого числа в качестве константы является как минимум нелогичным.

Вы можете оставить пропуски как есть, просто заменив их на какой-то специальный символ. Например, для числовых неотрицательных признаков можно использовать число -1, а для категориальных — строку 'unknown'.

Задание 2.

Скачайте датасет: https://disk.yandex.ru/d/q4ig4c-_heS95A

Набор данных содержит ~600 записей о девушках и восемь признаков:
BMI — индекс массы тела (ИМТ)
year — год размещения модели в журнале
month — месяц размещения

waist — обхват талии модели
hips — обхват бёдер модели
height — рост модели
weight — вес модели
waist/hip — соотношение обхвата талии и бёдер

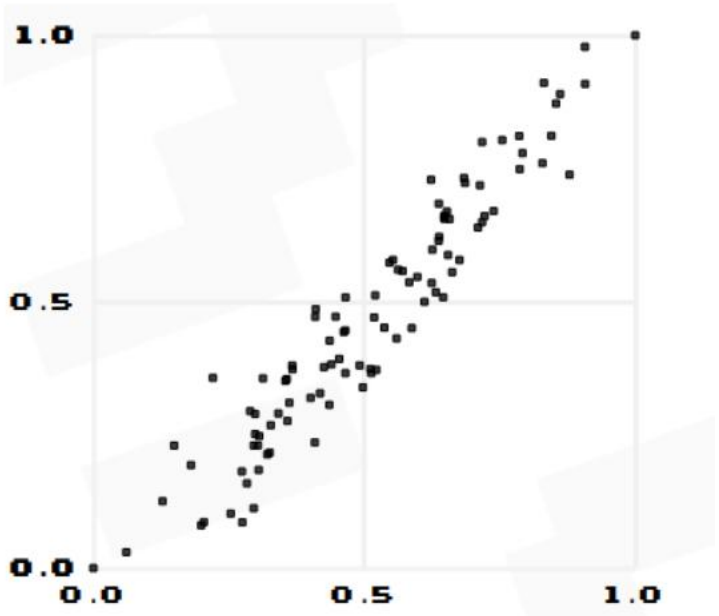
Прочитаем датасет и приступим к изучению способов визуализации.

```
data = pd.read_csv('model.csv')
```

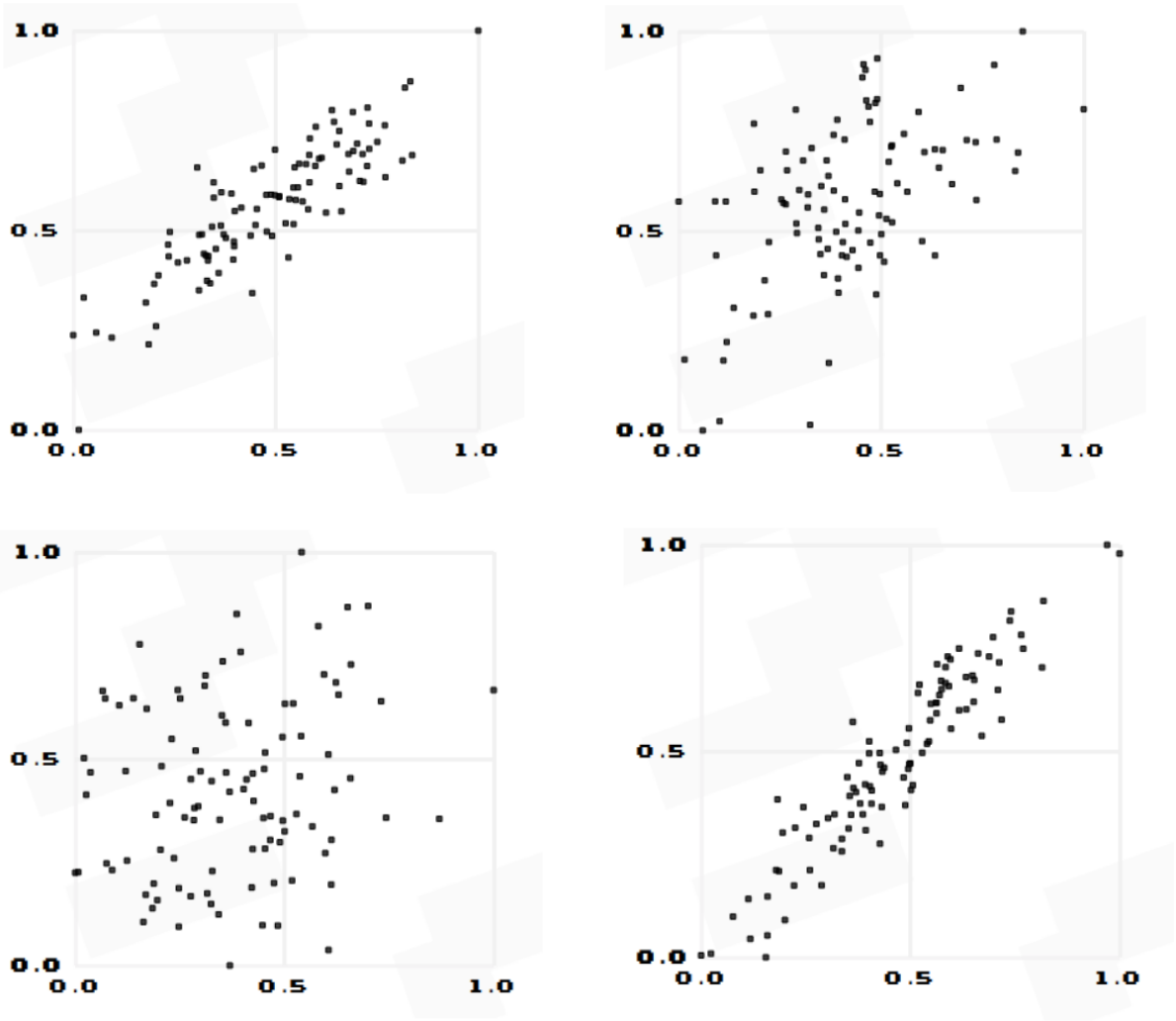
1. Постройте матрицу корреляций для датасета. В ответ впишите самый высокий коэффициент корреляции. Ответ округлите до сотых. (Ответ:0.74)
2. Постройте тепловую матрицу корреляций
3. Определите, к какому типу относятся признаки *weight* и *height*?
4. Рассчитайте среднее значение признаков при помощи библиотеки *statistics* для признаков *weight* и *height*. Ответ округлите до сотых.
5. Постройте матрицы визуализации корреляций, диаграммы рассеивания, проанализируйте и выполните задание. Заполните пропуски.

С увеличением обхвата бедер модели ИМТ незначительно _____ .
Обхват бёдер имеет _____ связь с весом модели.
С увеличением обхвата бёдер _____ вес модели.
Год выпуска журнала и вес модели имеют _____ связь.

6. Проанализируйте график и определите величину коэффициента корреляции между признаками на оси X и признаком на оси У на данном графике. Определите тип связи между признаком на оси X и признаком на оси Y.



7. Проанализируйте графики и выберите график с наибольшим коэффициентом корреляции.



КТ 2

Задание 1

Скачайте датасет «heart» болезней сердца: https://disk.yandex.ru/d/Pp5IXGNej_YcOA

Датасет болезней сердца содержит информацию о пациентах и переменную предсказания *target* — наличие у пациента болезни сердца.

Датасет содержит следующие признаки:

- *age* — возраст
- *sex* — пол (1 - мужчина, 0 - женщина)
- *cp* — тип боли в груди (4 значения)
- *trestbps* — артериальное давление в покое
- *chol* — холестерин сыворотки в мг/дл
- *fbs* — уровень сахара в крови натощак > 120 мг/дл
- *restecg* — результаты электрокардиографии в покое (значения 0,1,2)
- *thalach* — достигнута максимальная частота сердечных сокращений
- *exang* — стенокардия, вызванная физической нагрузкой
- *oldpeak* — депрессия *ST*, вызванная физической нагрузкой, по сравнению с состоянием покоя
- *slope* — наклон пикового сегмента *ST* при нагрузке
- *ca* — количество крупных сосудов (0-3), окрашенных при флюроскопии

- *thal* — дефект, где 3 = нормальный; 6 = фиксированный дефект; 7 = обратимый дефект

1. Создайте новый признак *old*, где 1 — при возрасте пациента более 60 лет. В ответ введите результат выполнения кода `heart['old'].sum()`.

2. Создайте новый признак trestbps_mean, который будет обозначать норму давления в среднем для его возраста и пола. trestbps — систолическое артериальное давление в состоянии покоя. Информацию о среднем давлении для возраста и пола возьмите из этой таблицы. В таблице систолическое давление написано первым, перед дробной чертой.

Возраст (лет)	Мужчины	Женщины
до 20	123/76	116/72
21-30	126/79	120/75
31-40	129/81	127/80
41-50	135/83	137/84
51-60	142/85	144/85
61 и старше	142/80	159/85

В ответ напишите значение признака `trestbps_mean` для пациента под номером 300. (Ответ: 142)

3. Проанализируйте датасет и выберите категориальные признаки.

4. Вышеперечисленные категориальные признаки уже представлены в числовом виде. Проанализируйте их и назовите те, которые нуждаются в дополнительном кодировании значений (например, OneHotEncoding). Учтите, что дополнительные методы кодирования требуются только для категориальных признаков с более чем двумя категориями. Бинарные признаки (принимающие два значения, такие как 0 и 1) не нуждаются в дополнительном кодировании.

5. Закодируйте вышеперечисленные признаки методом OneHotEncoding, удалив исходные признаки. Сколько признаков получилось в датасете?

6. Нормализуйте все числовые признаки подходящим способом. В ответе напишите стандартное отклонение признака chol. Ответ округлите до шести знаков после запятой.

7. Проведите корреляционный анализ и отберите признаки для будущей модели. Выберите пары сильно скоррелированных признаков.

Задание 2

Скачайте датасет: https://disk.yandex.ru/d/CbCag9g5qJD-gQ

Описание датасета:

work_year	Год, в котором была выплачена зарплата.

experience_level	Опыт работы на этой должности в течение года со следующими возможными значениями: ○ *EN — Entry-level/Junior;* ○ *MI — Mid-level/Intermediate;* ○ *SE — Senior-level/Expert;* ○ *EX — Executive-level/Director*.
employment_type	Тип трудоустройства для этой роли: ○ *PT* — неполный рабочий день; ○ *FT* — полный рабочий день; ○ *CT* — контракт; ○ *FL* — фриланс.
job_title	Роль, в которой соискатель работал в течение года.
salary	Общая выплаченная валовая сумма заработной платы.
salary_currency	Валюта выплачиваемой заработной платы в виде кода валюты *ISO* 4217.
salary_in_usd	Зарплата в долларах США (валютный курс, делённый на среднее значение курса доллара США за соответствующий год через *fxdata.foorilla.com*).
employee_residence	Основная страна проживания сотрудника в течение рабочего года в виде кода страны *ISO* 3166.
remote_ratio	Общий объём работы, выполняемой удалённо. Возможные значения: ○ 0 — удалённой работы нет (менее 20 %); ○ 50 — частично удалённая работа; ○ 100 — полностью удалённая работа (более 80 %).
company_location	Страна главного офиса работодателя или филиала по контракту в виде кода страны *ISO* 3166.
company_size	Среднее количество людей, работавших в компании в течение года: ○ *S* — менее 50 сотрудников (небольшая компания); ○ *M* — от 50 до 250 сотрудников (средняя компания); ○ *L* — более 250 сотрудников (крупная компания).

Исследуйте данные и сделайте выводы по полученным результатам. Подкрепите свои рассуждения и выводы визуализациями и с помощью статистического тестирования проверьте, являются ли выводы статистически значимыми.

В процессе своего анализа вы должны:

Выяснить, какие факторы влияют на зарплату специалиста Data Scientist. А также ответить на ключевые вопросы HR-агентства:

Наблюдается ли ежегодный рост зарплат у специалистов Data Scientist?

Как соотносятся зарплаты Data Scientist и Data Engineer в 2022 году?

Как соотносятся зарплаты специалистов Data Scientist в компаниях различных размеров?

Есть ли связь между наличием должностей Data Scientist и Data Engineer и размером компании?

Если вы найдёте в данных интересные закономерности, также отметьте их в своём анализе.

Продемонстрируйте использование разных тестов для проверки статистической значимости сделанных выводов:

тесты для количественного признака:

для одной выборки;

для двух выборок;

для нескольких выборок;

тест для категориальных признаков.

6. Формы промежуточной аттестации, критерии и шкала оценивания, типовые оценочные материалы по дисциплине

6.1. Промежуточная аттестация проводится в форме *зачета* в четвертом семестре в письменной форме.

6.2. Типовые оценочные материалы промежуточной аттестации.

РАЗДЕЛ 1. ФУНДАМЕНТ EDA: СТАТИСТИКА И АНАЛИЗ СВЯЗЕЙ

Тема 1. Жизненный цикл ML-модели и роль EDA в машинном обучении

1. Дайте определение EDA (разведывательному анализу данных). Какие задачи он решает?
2. Перечислите и кратко охарактеризуйте этапы жизненного цикла ML-модели.
3. Какое место занимает EDA в жизненном цикле ML? Между какими этапами он находится?
4. Что такое проектирование признаков (feature engineering)? Чем оно отличается от отбора признаков?
5. Что такое кодирование признаков? Назовите основные методы (обзорно).
6. Какие роли в Data Science существуют согласно модулю? В чём разница между дата-аналитиком и дата-сайентистом?
7. Назовите 3 инструмента автоматической визуализации. Для чего они используются?

Тема 2. Математическая статистика в контексте EDA

1. Что такое математическая статистика и статистические данные?
2. С какой целью применяется описательная статистика?
3. Дайте определение и формулу (или правило вычисления) для среднего, медианы и моды.
4. Что такое корреляция? Поясните положительную, отрицательную и нулевую корреляцию.
5. В чём различие между корреляцией Пирсона, Спирмена и Кендалла? Для каких данных каждый метод предпочтителен?
6. Как интерпретировать величину коэффициента корреляции (от отсутствия связи до очень сильной)?

7. Какие функции библиотеки Seaborn используются для визуализации корреляций? Что каждая из них показывает?

РАЗДЕЛ 2. ПРИКЛАДНОЙ EDA: ПРИЗНАКИ, ПРЕОБРАЗОВАНИЯ И ПРОВЕРКА ГИПОТЕЗ

Тема 3. Статистические гипотезы в контексте EDA

1. Для чего необходимы статистические тесты в анализе данных?
2. Назовите три критерия выбора статистического теста (тип данных, распределение, зависимость выборок).
3. Что такое статистическая значимость? Поясните понятия p-value и α-уровень.
4. Как принимается решение о гипотезе на основе p-value и α-уровня? Приведите пример.
5. Что такое нулевая (H_0) и альтернативная (H_1) гипотезы? Приведите пример для сравнения двух групп.
6. Какие тесты используются для проверки нормальности распределения? (Назовите не менее двух.)
7. В чём различие между параметрическими и непараметрическими тестами? Приведите по одному примеру каждого.
8. Какие статистические тесты применяются для категориальных признаков? В чём разница между χ^2-квадрат и точным тестом Фишера?

Тема 4. Проектирование и отбор признаков

1. Что такое создание признаков (feature engineering) и обогащение данных из внешних источников?
2. Какие операции можно выполнять с признаками типа «дата-время»? Назовите 3–4 примера.
3. Перечислите и кратко охарактеризуйте методы кодирования категориальных признаков (One-Hot, Label, Target Encoding).
4. В чём разница между нормализацией (MinMax) и стандартизацией (StandardScaler)? Когда какой метод предпочтительнее?
5. Какие существуют методы отбора признаков? Перечислите три группы (фильтрация, оберточные, встроенные) и приведите примеры.
6. Что такое мультиколинеарность? Какие проблемы она создаёт для моделей?
7. Какие инструменты используются для обнаружения мультиколинеарности? Поясните, как работает VIF и корреляционная матрица.

6.3. Критерии и шкала оценивания на основе БРС.

Соответствие государственной шкалы оценивания академической успеваемости и шкалы ECTS при экзамене

Оценка по шкале ECTS	**Сумма баллов за все виды учебной деятельности**	**Оценка по государственной шкале**	**Определение**
A	90 – 100	«Отлично»	отличное выполнение с незначительным количеством неточностей
B	80 – 89	«Хорошо»	в целом правильно выполненная работа с незначительным количеством ошибок (до 10%)
C	75 – 79		в целом правильно выполненная работа с незначительным количеством ошибок (до 15%)
D	70 – 74	«Удовлетворительно»	неплохо, но со значительным количеством недостатков
E	60 – 69		выполнение удовлетворяет минимальные критерии
FX	35 – 59	«Не удовлетворительно»	с возможностью повторной сдачи
F	0 – 34		с обязательным повторным изучением дисциплины (выставляется комиссией)

6.4. Описание дополнительных материалов и оборудования, необходимых для выполнения проверочных заданий

Компьютер с операционной системой RedOS, на котором установлены VS Code, Jupyter Notebook, интерпретатор Python.

7. Методические материалы по освоению дисциплины

Получение углубленных знаний по изучаемой дисциплине достигается за счет дополнительных часов к аудиторной работе самостоятельной работы студентов. Выделяемые часы целесообразно использовать для знакомства с дополнительной научной литературой по проблематике дисциплины, анализа научных концепций и современных подходов к осмыслению рассматриваемых проблем. К самостоятельному виду работы студентов относится работа в библиотеках, в электронных поисковых системах и т.п. по сбору материалов, необходимых для проведения практических занятий или выполнения конкретных заданий преподавателя по изучаемым темам. Студенты могут установить диалог с преподавателем, получать консультации по выполнению заданий. В качестве оценочных средств на протяжении семестра используются практические задания.

Обучение по дисциплине «Разведывательный анализ данных» предполагает изучение курса на аудиторных занятиях (лекции, практические занятия) и самостоятельную работу студентов. Практические занятия дисциплины предполагают их проведение в различных формах с целью выявления полученных знаний, умений, навыков и компетенций с проведением контрольных мероприятий. С целью обеспечения успешного обучения студент должен готовиться к лекции, поскольку она является важнейшей формой организации учебного процесса, поскольку:

– знакомит с новым учебным материалом;
– разъясняет учебные элементы, трудные для понимания;
– систематизирует учебный материал;
– ориентирует в учебном процессе.

Работа обучающегося на лекции:

Слушание и запись лекций – сложный вид вузовской аудиторной работы. Внимательное слушание и конспектирование лекций предполагает интенсивную умственную деятельность обучающегося. Краткие записи лекций, их конспектирование помогает усвоить учебный материал. Конспект является полезным тогда, когда записано самое существенное, основное и сделано это самим обучающимся.

Подготовка к практическим занятиям:

Подготовку к каждому практическому занятию каждый обучающийся должен начать с ознакомления с планом, который отражает содержание предложенной темы. Тщательное продумывание и изучение вопросов плана основывается на проработке текущего материала лекции, а затем изучения обязательной и дополнительной литературы, рекомендованную к данной теме. Если программой дисциплины предусмотрено выполнение практического задания, то его необходимо выполнить с учетом предложенной инструкции. Все новые понятия по изучаемой теме необходимо внести в глоссарий, который целесообразно вести с самого начала изучения курса. Результат такой работы должен проявиться в способности обучающегося свободно ответить на теоретические вопросы практического занятия, его выступлении и участии в коллективном обсуждении вопросов изучаемой темы, правильном выполнении практических заданий и контрольных работ.

Структура практического занятия:

В зависимости от содержания и количества отведенного времени на изучение каждой темы может практическое занятие состоять из четырех-пяти частей:

1. Устный опрос.

2. Обсуждение теоретических вопросов, определенных программой дисциплины.

3. Выполнение практических заданий с последующим разбором полученных результатов или обсуждение практического задания, выполненного дома.

4. Подведение итогов занятия.

Работа с литературными источниками:

В процессе подготовки к практическим занятиям, обучающимся необходимо обратить особое внимание на самостоятельное изучение рекомендованной учебно-методической (а также научной и популярной) литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной и популярной литературой, материалами периодических изданий и Интернета, статистическими данными является наиболее эффективным методом получения знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у обучающихся свое отношение к конкретной проблеме. Более глубокому раскрытию вопросов способствует знакомство с дополнительной литературой, рекомендованной преподавателем, что позволяет обучающимся проявить свою индивидуальность в рамках выступления на занятиях, выявить широкий спектр мнений по изучаемой проблеме.

8. Учебная литература и ресурсы информационно-телекоммуникационной сети Интернет

8.1. Основная литература

1. Пальмов, С. В. Интеллектуальные информационные системы и технологии : учебное пособие / С. В. Пальмов. — Самара : ПГУТИ, 2023. — 387 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/411827 (дата обращения: 07.05.2026). — Режим доступа: для авториз. пользователей.

2. Ванг, К. Конструирование систем глубокого обучения : руководство / К. Ванг, Д. Сзето ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2023. — 462 с. — ISBN 978-5-93700-181-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/456644 (дата обращения: 07.05.2026). — Режим доступа: для авториз. пользователей.

3. Сидоров, И. Г. Методы и технологии обработки больших данных : учебно-методическое пособие / И. Г. Сидоров. — Москва : Московский Политех, 2024. — 282 с. — ISBN 978-5-2760-2843-9. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/482792 (дата обращения: 07.05.2026). — Режим доступа: для авториз. пользователей.

8.2. Дополнительная литература

4. Меджедович, Д. Алгоритмы и структуры для массивных наборов данных : руководство / Д. Меджедович, Э. Тахирович ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2024. — 340 с. — ISBN 978-5-93700-250-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/456665 (дата обращения: 07.05.2026). — Режим доступа: для авториз. пользователей.

5. Хайндман, Р. Прогнозирование: принципы и практика : учебник / Р. Хайндман, Д. Атанасопулос ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2023. — 458 с. — ISBN 978-5-93700-151-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/348068 (дата обращения: 07.05.2026). — Режим доступа: для авториз. пользователей.

8.3. Нормативные правовые документы и иная правовая информация

1. Конституция Российской Федерации. – Текст : электронный // Сайт Президента Российской Федерации. – URL: http://www.kremlin.ru/acts/constitution

8.4 Интернет-ресурсы

1. Информационно-правовой портал ГАРАНТ.РУ. – URL: https://www.garant.ru/
2. Научная электронная библиотека eLIBRARY.RU. – URL: https://elibrary.ru/
3. Научная электронная библиотека «КиберЛенинка». – URL: https://cyberleninka.ru
4. Электронно-библиотечная система «Лань». – URL: http://e.lanbook.com
5. База знаний по ОС RedOS – URL: https://redos.red-soft.ru/base/
6. Документация по Python – URL: **https://docs.python.org/3/**

9. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы

Материально-техническое обеспечение дисциплины включает в себя:

- лекционные аудитории, оборудованные видеопроекционным оборудованием для презентаций, средствами звуковоспроизведения, экраном;
- помещения для проведения практических занятий, оборудованные учебной мебелью.

Дисциплина поддержана соответствующими программными продуктами с открытой лицензией: RedOS, VS Code, Jupyter Notebook, интерпретатор Python.

Вуз обеспечивает каждого обучающегося рабочим местом в компьютерном классе в соответствии с объемом изучаемых дисциплин, обеспечивает выход в сеть Интернет.

Помещения для самостоятельной работы обучающихся включают следующую оснащенность: столы аудиторные, стулья, доски аудиторные, компьютеры с подключением к локальной сети института (для компьютерных аудиторий) и Интернет. Для изучения учебной дисциплины используются автоматизированная библиотечная информационная система и электронные библиотечные системы.