

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
СТАВРОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ**

УТВЕРЖДАЮ

Директор/Декан
факультета цифровых технологий
Шлаев Дмитрий Валерьевич

«__» _____ 20__ г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ (ОЦЕНОЧНЫХ МАТЕРИАЛОВ)

Б1.В.ДВ.01.02 Инжиниринг признаков и предобработка данных

09.03.02 Информационные системы и технологии

Инженерия систем искусственного интеллекта

бакалавр

очная

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

Процесс изучения дисциплины направлен на формирование следующих компетенций ОП ВО и овладение следующими результатами обучения по дисциплине:

Код и наименование компетенции	Код и наименование индикатора достижения	Перечень планируемых результатов обучения по дисциплине
ПК-2 Способен разрабатывать и внедрять интеллектуальные системы	ПК-2.4 Способен работать с инструментами и библиотеками для анализа данных и машинного обучения	знает основные этапы подготовки данных для решения задач анализа данных и машинного обучения; назначение и функциональные возможности современных инструментов и библиотек для обработки, анализа и визуализации данных; принципы работы с табличными, категориальными, числовыми, текстовыми и временными данными; методы очистки данных, обработки пропущенных значений, выбросов, дубликатов и некорректных записей; методы масштабирования, нормализации, кодирования, дискретизации, агрегации и трансформации признаков; основные подходы к отбору признаков, снижению размерности и оценке значимости признаков; принципы построения воспроизводимых пайплайнов предобработки данных и подготовки признакового пространства для моделей машинного обучения.
		умеет применять программные инструменты и библиотеки для загрузки, исследования, очистки и преобразования данных; выполнять первичный анализ структуры, качества и распределения данных; выявлять и обрабатывать пропущенные значения, выбросы, шумы, дубликаты и аномальные наблюдения; выполнять кодирование категориальных признаков, масштабирование числовых признаков и преобразование временных данных; конструировать новые признаки на основе исходных данных и особенностей предметной области; осуществлять отбор информативных признаков с использованием статистических, модельных и алгоритмических методов; формировать наборы данных, пригодные для обучения, проверки и тестирования моделей машинного обучения; оценивать влияние выбранных методов предобработки и инжиниринга признаков на качество моделей машинного обучения.

		владеет навыками навыками работы с библиотеками анализа данных и машинного обучения для решения прикладных задач; практическими приемами предобработки данных и формирования признаков описания объектов; методами построения и настройки пайплайнов подготовки данных для моделей машинного обучения; инструментами визуального и статистического анализа данных для выявления закономерностей и проблем качества данных; навыками применения методов отбора признаков и снижения размерности; приемами документирования и воспроизведения этапов подготовки данных; навыками интерпретации результатов предобработки данных и обоснования выбора методов инжиниринга признаков для конкретной прикладной задачи.
--	--	---

2. Перечень оценочных средств по дисциплине

№	Наименование раздела/темы	Семестр	Код индикаторов достижения компетенций	Оценочное средство проверки результатов достижения индикаторов компетенций
1.	1 раздел. Раздел			
1.1.	Основы анализа качества данных и базовая предобработка	5		Собеседование
1.2.	Инжиниринг признаков, отбор признаков и построение пайплайнов машинного обучения	5		Собеседование
	Промежуточная аттестация			За

3. Оценочные средства (оценочные материалы)

Примерный перечень оценочных средств для текущего контроля успеваемости и промежуточной аттестации

№ п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в фонде (Оценочные материалы)
Текущий контроль			
	Для оценки знаний		

1	Собеседование	Средство контроля, организованное как специальная беседа преподавателя с обучающимся на темы, связанные с изучаемой дисциплиной, и рассчитанное на выяснение объема знаний обучающегося по определенному разделу, теме, проблеме и т.п.	Вопросы по темам/разделам дисциплины
	Для оценки умений		
	Для оценки навыков		
	Промежуточная аттестация		
2	Зачет	Средство контроля усвоения учебного материала практических и семинарских занятий, успешного прохождения практик и выполнения в процессе этих практик всех учебных поручений в соответствии с утвержденной программой с выставлением оценки в виде «зачтено», «незачтено».	Перечень вопросов к зачету

4. Примерный фонд оценочных средств для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине (модулю) "Инжиниринг признаков и предобработка данных"

Примерные оценочные материалы для текущего контроля успеваемости

Примерные оценочные материалы для проведения промежуточной аттестации (зачет, экзамен) по итогам освоения дисциплины (модуля)

Модуль 1. Основы анализа качества данных и базовая предобработка

Раскройте роль предобработки данных в жизненном цикле разработки модели машинного обучения.

Объясните, как качество исходных данных влияет на точность, устойчивость и интерпретируемость модели машинного обучения.

Перечислите основные этапы подготовки данных для решения задач машинного обучения.

Охарактеризуйте основные типы данных, используемые в задачах анализа данных и машинного обучения.

Объясните различия между числовыми, категориальными, порядковыми, бинарными, временными и текстовыми признаками.

Опишите основные источники данных, применяемые при построении интеллектуальных информационных систем.

Раскройте понятие «качество данных» и перечислите его основные характеристики.

Объясните, что такое полнота, точность, согласованность, актуальность и непротиворечивость данных.

Опишите основные проблемы, возникающие при работе с реальными наборами данных.

Объясните, почему перед обучением модели необходимо проводить первичный анализ данных.

Разработайте алгоритм первичного исследования табличного набора данных.

Опишите, какие статистические показатели целесообразно использовать для анализа числовых признаков.

Объясните назначение визуализации данных на этапе предварительного анализа.

Приведите примеры графиков, используемых для выявления структуры и распределения данных.

Объясните, как выявить пропущенные значения в наборе данных.

Охарактеризуйте основные причины появления пропусков в данных.

Сравните подходы к обработке пропущенных значений: удаление, заполнение константой, заполнение средним, медианой, модой.

Объясните, в каких случаях удаление строк или столбцов с пропусками является допустимым.

Разработайте схему выбора метода обработки пропущенных значений в зависимости от типа признака.

На примере табличного набора данных опишите порядок заполнения пропущенных значений в числовых признаках.

На примере табличного набора данных опишите порядок заполнения пропущенных значений в категориальных признаках.

Объясните, что такое дубликаты в данных и почему они могут искажать результаты машинного обучения.

Опишите методы выявления и удаления полных и частичных дубликатов.

Объясните, как некорректные значения могут быть выявлены с использованием логических правил предметной области.

Приведите примеры ошибок в данных, связанных с форматом, единицами измерения и нарушением допустимых диапазонов.

Раскройте понятие «выброс» в данных.

Объясните различия между статистическим выбросом, ошибочным значением и редким, но корректным наблюдением.

Опишите методы выявления выбросов с использованием межквартильного размаха.

Опишите методы выявления выбросов с использованием z-оценки.

Объясните, как визуально обнаружить выбросы с помощью диаграммы размаха и гистограммы.

Сравните основные способы обработки выбросов: удаление, ограничение, замена, логарифмирование.

Разработайте порядок обработки выбросов для числового признака с асимметричным распределением.

Объясните назначение масштабирования числовых признаков.

Сравните нормализацию и стандартизацию признаков.

Опишите случаи, когда масштабирование признаков является обязательным этапом подготовки данных.

Объясните принцип работы Min-Max scaling.

Объясните принцип работы StandardScaler.

Охарактеризуйте устойчивое масштабирование признаков с использованием медианы и межквартильного размаха.

Объясните, как логарифмическое преобразование может использоваться для обработки скошенных распределений.

Раскройте назначение кодирования категориальных признаков.

Сравните Label Encoding, One-Hot Encoding и Ordinal Encoding.

Объясните, почему некорректное кодирование категориальных признаков может ухудшить качество модели.

Разработайте схему выбора метода кодирования категориального признака.

Опишите особенности предобработки бинарных признаков.

Выполните комплексное описание этапов базовой предобработки табличного набора данных перед обучением модели машинного обучения.

Модуль 2. Инжиниринг признаков, отбор признаков и построение пайплайнов машинного обучения

Раскройте сущность понятия «инжиниринг признаков».

Объясните, чем инжиниринг признаков отличается от базовой предобработки данных.

Охарактеризуйте роль предметной области при создании новых признаков.

Приведите примеры создания новых признаков на основе существующих числовых данных.

Приведите примеры создания новых признаков на основе категориальных данных.

Объясните, как агрегирование данных может использоваться при формировании признаков.

Разработайте пример признаков, полученных на основе группировки данных по категориям.

Объясните назначение признаков, отражающих отношения, доли, разности и темпы изменения.

Опишите подходы к созданию признаков на основе статистических характеристик группы объектов.

Раскройте особенности инжиниринга признаков для временных данных.

Объясните, как из даты и времени можно получить дополнительные признаки для модели машинного обучения.

Приведите примеры календарных признаков: день недели, месяц, квартал, сезон, праздничный день.

Объясните назначение лаговых признаков при работе с временными рядами.

Охарактеризуйте скользящие статистики и их применение при формировании признаков временных данных.

Разработайте пример набора признаков для задачи прогнозирования временного ряда.

Раскройте особенности предобработки текстовых данных.

Объясните назначение токенизации, приведения к нижнему регистру, удаления стоп-слов и лемматизации.

Сравните подходы Bag of Words и TF-IDF для представления текстовых данных.

Объясните, в каких задачах целесообразно использовать векторное представление текстов.

Опишите особенности работы с признаками высокой размерности.

Раскройте понятие «отбор признаков».

Объясните, зачем необходимо выполнять отбор признаков перед построением модели машинного обучения.

Сравните фильтрационные, оберточные и встроенные методы отбора признаков.

Опишите отбор признаков на основе корреляционного анализа.

Объясните, как мультиколлинеарность признаков может влиять на качество модели.

Опишите использование статистических критериев для отбора признаков.

Раскройте назначение метода SelectKBest.

Объясните, как модели машинного обучения могут использоваться для оценки важности признаков.

Опишите отбор признаков на основе коэффициентов линейной модели.

Опишите отбор признаков на основе важности признаков в деревьях решений и ансамблевых моделях.

Объясните, как регуляризация может использоваться для отбора признаков.

Раскройте понятие снижения размерности данных.

Сравните отбор признаков и извлечение признаков.

Объясните назначение метода главных компонент.

Опишите основные этапы применения PCA к набору данных.

Объясните, как выбрать количество компонент при снижении размерности.

Охарактеризуйте преимущества и ограничения методов снижения размерности.

Раскройте понятие пайплайна предобработки данных.

Объясните, зачем необходимо строить воспроизводимые пайплайны подготовки данных.

Опишите структуру пайплайна, включающего обработку пропусков, масштабирование и кодирование признаков.

Объясните, как избежать утечки данных при предобработке обучающей и тестовой выборок.

Разработайте алгоритм разделения данных на обучающую, валидационную и тестовую выборки с учетом последующей предобработки.

Объясните, как оценить влияние различных методов предобработки на качество модели машинного обучения.

Разработайте план экспериментального сравнения нескольких вариантов инжиниринга признаков для одной задачи машинного обучения.

Выполните описание полного цикла подготовки данных для прикладной задачи машинного обучения: анализ качества данных, очистка, обработка пропусков и выбросов, кодирование, масштабирование, создание новых признаков, отбор признаков и построение пайплайна.

Темы письменных работ (эссе, рефераты, курсовые работы и др.)