

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт дистанционного образования

УТВЕРЖДЕНО:

Директор ИДО - проректор по РДО
М.О. Шепель

Оценочные материалы по дисциплине

Разведывательный анализ данных

по направлению подготовки

09.04.03 Прикладная информатика

Направленность (профиль) подготовки:
«Науки о данных»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОПОП
Д.Д. Даммер

Председатель УМК
С.Б. Велединская

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 - способность самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;;

ОПК-3 - способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;;

ОПК-6 - способность исследовать современные проблемы и методы прикладной информатики и развития информационного общества;;

ПК-1 - способность управлять получением, хранением, передачей, обработкой больших данных;

ПК-2 - способность применять и разрабатывать новые методы и технологии исследования больших данных.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.2 Знает основы определения взаимосвязей, закономерностей, обобщения и абстрагирования фундаментальных моделей, законов и методик для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;

ИОПК-3.1 Знает основы сбора, обработки и анализа научно-технической информации, необходимой для решения профессиональных задач;

ИОПК-3.3 Знает основы формулирования результатов, полученных в ходе решения исследовательских задач, в виде аналитических обзоров с обоснованными выводами и рекомендациями;

ИОПК-6.1 Знает основы методов анализа прикладной области, информационных потребностей, технологий сбора, накопления, обработки, передачи и распространения информации;

ИПК-1.4 Знает основы разработки предложений по развитию и совершенствованию системы получения, хранения, передачи и обработки больших данных;

ИПК-1.5 Знает основы разработки моделей данных, адаптированных к технологиям больших данных;

ИПК-2.2 Знает основы проведения аналитических исследований по тематике информационных технологий и технологий больших данных.

2. Оценочные материалы текущего контроля и критерии оценивания

№	Этапы формирования компетенций (разделы дисциплины/модуля/практики)	Код и наименование результатов обучения	Вид оценочного средства (тесты, задания, кейсы, вопросы и др.)
	Тема 1. Введение в разведывательный анализ данных. Алгоритмы и методы EDA Жизненный цикл ML-модели и роль EDA в машинном обучении. Алгоритм и методы EDA. Знакомство с данными: винные обзоры. Проверка. БОНУС. EDA одной строкой кода	ИОПК-1.2; ИОПК-3.1; ИОПК-3.3; ИПК-1.4	Тест Контрольная работа Домашнее задание Проект

Тема 2. Математическая статистика в контексте EDA Описательная статистика. Меры центральной тенденции в Python. Корреляция. Типы корреляций. Корреляция Пирсона. Типы корреляций. Ранговые корреляции. Визуализация корреляций. Матрица корреляций. График рассеивания. Парные отношения в наборе данных.	ИОПК-1.2; ИОПК-3.1; ИОПК-3.3; ИОПК-6.1.	Тест Контрольная работа Домашнее задание Проект
Тема 3. Проектирование признаков Создание признаков. Создание признаков. Внешние источники данных. Создание признаков. Работа с форматом «дата-время». Кодирование признаков. Методы. Преобразование признаков. Нормализация. Стандартизация. Отбор признаков. Мультиколлинеарность.	ИОПК-3.1; ИПК-1.4; ИПК-1.5; ИПК-2.2.	Тест Контрольная работа Домашнее задание Проект
Тема 4. Статистические тесты Понятие статистической гипотезы. Статистическая значимость. Статистические тесты. Введение. Тесты на нормальность Параметрические тесты. Непараметрические тесты. Популярные статистические тесты в Python.	ИОПК-3.1; ИОПК-3.3; ИОПК-6.1; ИПК-2.2	Тест Контрольная работа Домашнее задание Проект
Тема 5. Статистические тесты Необходимость в статистических тестах. Понятие статистической гипотезы. Статистическая значимость. Статистические тесты: введение. Проверка на нормальность. Параметрические тесты. Непараметрические тесты. Статистические тесты для категориальных признаков. Операции с векторами. Практика. Статистические тесты в контексте EDA.	ИОПК-3.1; ИОПК-3.3; ИОПК-6.1; ИПК-2.2	Тест Контрольная работа Домашнее задание Проект
Тема 6. Проектирование экспериментов Введение в проектирование экспериментов. Логирование экспериментов. Знакомство с Comet.ml. Планирование эксперимента и	ИОПК-3.1; ИОПК-3.3; ИОПК-6.1; ИПК-2.2	Тест Контрольная работа Домашнее задание Проект

	логирование изменений в Comet.ml. Практика.		
	Тема 7. А/В-тестирование Суть и сферы применения А/В-тестирования. Алгоритм, принципы и параметры А/В-тестирования. Анализ результатов А/В-тестирования: кумулятивные метрики. Анализ результатов А/В-тестирования: статистические тесты. Анализ результатов А/В-тестирования: доверительные интервалы. Закрепление знаний	ИОПК-3.1; ИОПК-3.3; ИОПК-6.1; ИПК-2.2	Тест Контрольная работа Домашнее задание Проект
	Тема 8. Знакомство с Kaggle Основы Kaggle. Работа с ноутбуками в Kaggle. Участие в соревновании.	ИОПК-1.2; ИОПК-3.1; ИПК-1.4; ИПК-2.2	Тест Контрольная работа Домашнее задание Проект
	Тема 9. EDA + Feature Engineering. Соревнование на Kaggle Разбираемся с данными. Машинное обучение для самых маленьких. Лучший способ добиться успеха — следовать советам. Побеждаем в соревновании.	ИОПК-1.2; ИОПК-3.1; ИОПК-6.1; ИПК-1.4; ИПК-2.2	Тест Контрольная работа Домашнее задание Проект

Примеры элементов текущего контроля:

Тест

Задание 1

Какую роль играет разведывательный анализ данных (EDA) в жизненном цикле модели машинного обучения? (ИОПК-1.2)

- а) Определение взаимосвязей и закономерностей в данных
- б) Очистка данных
- в) Создание модели
- г) Оценка производительности модели

Задание 2

Какой тип корреляции используется для измерения линейной зависимости между двумя переменными? (ИОПК-6.1)

- а) Корреляция Пирсона
- б) Спирмен

в) Кендалл

г) Тау

Задание 3

Какой метод используется для приведения признаков к единому масштабу? (ИОПК-3.1)

а) Кодирование признаков

б) Нормализация

в) Отбор признаков

г) Создание признаков

Задание 4

Какой статистический тест используется для проверки нормальности распределения данных? (ИПК-2.2)

а) t-тест

б) ANOVA

в) Шапиро-Уилка

г) хи-квадрат

Идентификатор:

Задание 5.

Какой метод используется для оценки статистической значимости различий между двумя группами в A/B-тестировании? (ИОПК-3.3)

а) Регрессионный анализ

б) Корреляционный анализ

в) t-тест

г) Факторный анализ

Критерии оценивания: тест считается пройденным, если обучающий ответил правильно как минимум на половину вопросов.

Домашние задания

Задание 1 (ИОПК-1.2)

Проведите разведывательный анализ данных набора "Wine Reviews" с использованием Python. Ваша задача — ознакомиться с данными, выявить основные характеристики и закономерности.

Шаги выполнения:

1. Загрузите данные "Wine Reviews" из файла CSV.
2. Выполните первичный осмотр данных: количество строк и столбцов, типы данных, наличие пропусков.

3. Проведите анализ описательной статистики для числовых столбцов.
4. Постройте гистограммы и коробчатые диаграммы (box plots) для визуализации распределений числовых данных.
5. Определите и визуализируйте корреляции между числовыми признаками.
6. Сформулируйте выводы по проведенному анализу.

Задание 2 (ИОПК-3.1, ИПК-1.5)

Используя набор данных о продажах, проведите анализ описательной статистики, вычислите корреляции и выполните проектирование новых признаков.

Шаги выполнения:

1. Загрузите данные о продажах из файла CSV.
2. Проведите анализ описательной статистики для всех числовых признаков (среднее, медиана, стандартное отклонение и т.д.).
3. Постройте матрицу корреляций и визуализируйте ее с помощью heatmap.
4. Создайте новые признаки:
 - Признаки, основанные на датах (например, день недели, месяц, квартал).
 - Признаки, основанные на агрегировании данных (например, средние продажи по месяцам, максимальные продажи по дням недели).
 - Признаки, основанные на преобразованиях (например, логарифмирование или стандартизация).

Проект №1 (ИОПК-1.2, ИОПК-3.1, ИОПК-6.1, ИПК-1.4, ИПК-1.5, ИПК-2.2)

Цель проекта:

Провести полный цикл разведывательного анализа данных (EDA) и проектирования признаков, а также использовать статистические методы и эксперименты для прогнозирования рейтингов вин. Завершить проект участием в соревновании на Kaggle с применением полученных знаний.

Описание этапов проекта:

1. **Введение в разведывательный анализ данных (EDA)**
 - Выполните разведывательный анализ данных набора "Wine Reviews".
 - Определите жизненный цикл ML-модели и роль EDA в машинном обучении.
 - Ознакомьтесь с алгоритмами и методами EDA.
 - Реализуйте базовый EDA одной строкой кода для знакомства с данными.
2. **Математическая статистика в контексте EDA**
 - Проанализируйте описательную статистику для всех числовых признаков.
 - Вычислите и визуализируйте корреляции между числовыми признаками.
 - Постройте матрицу корреляций и графики рассеивания.
3. **Проектирование признаков**
 - Создайте новые признаки, используя внешние источники данных и работа с форматом «дата-время».
 - Выполните кодирование признаков, нормализацию и стандартизацию.
 - Осуществите отбор признаков с учетом мультиколлинеарности.
4. **Статистические тесты**
 - Проведите параметрические и непараметрические тесты для оценки значимости признаков.

- Проверьте данные на нормальность распределения.
 - Выполните популярные статистические тесты в Python и проанализируйте результаты.
5. **Проектирование экспериментов и A/B-тестирование**
- Запланируйте и проведите эксперимент по улучшению модели.
 - Логируйте результаты экспериментов, используя инструменты (например, Comet.ml).
 - Выполните A/B-тестирование для проверки гипотезы.
6. **Участие в соревновании на Kaggle**
- Загрузите и подготовьте данные для соревнования на Kaggle.
 - Реализуйте EDA и проектирование признаков для решения задачи на Kaggle.
 - Примите участие в соревновании, следуя рекомендациям и лучшим практикам.

Ожидаемые результаты:

1. Отчет по проведенному EDA с выводами и рекомендациями.
2. Набор новых признаков и их обоснование.
3. Результаты статистических тестов и выводы по значимости признаков.
4. Документ с планом эксперимента, логами и анализом результатов.
5. Итоговая модель, загруженная на Kaggle, с подробным описанием подхода и достигнутых результатов.

Контрольная работа (ИУК 1.1, ИПК-3.3)

Контрольная работа состоит из 4 теоретических вопросов и 6 задач.

Перечень теоретических вопросов:

Теоретические вопросы (4)

1. Опишите жизненный цикл ML-модели и укажите роль разведывательного анализа данных (EDA) на каждом этапе.
2. Что такое корреляция? Объясните различия между корреляцией Пирсона и ранговой корреляцией. Когда следует использовать каждый из этих методов?
3. Определите, что такое статистическая гипотеза и какова ее роль в статистических тестах. Приведите примеры параметрических и непараметрических тестов.
4. Объясните, что такое A/B-тестирование. Каковы его основные принципы и когда его целесообразно применять?

Примеры задач:

1. На основании предоставленного набора данных о винных обзорах выполните базовый разведывательный анализ данных (EDA) с использованием Python. Отобразите ключевые статистики и визуализации.
2. Используя набор данных о продажах, вычислите корреляцию между различными признаками (например, цена и количество продаж). Постройте матрицу корреляций и визуализируйте ее.
3. Проведите проверку нормальности распределения одного из признаков в наборе данных с использованием подходящих статистических тестов. Определите, является ли признак нормальным.
4. Создайте новый признак в наборе данных (например, год выпуска вина) и выполните его кодирование. Объясните, почему данный признак может быть полезен для анализа.

5. На основе выбранного набора данных проведите А/В-тест для оценки эффективности нового маркетингового подхода. Опишите шаги, которые вы предприняли, и интерпретируйте результаты.
6. Решите задачу отборов признаков на основе набора данных о здоровье, используя методы нормализации и стандартизации. Опишите, как это влияет на качество модели.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Промежуточная аттестация проводится в формате итогового индивидуального проекта. Индивидуальный проект представляет из себя письменный отчет (файл расширения docx/doc/pdf) на основе проведенного исследования, в рамках которого самостоятельно ставится исследовательский вопрос, выбираются данные, оценивается модель.

Отчет состоит из двух частей. Структура отчета следующая.

Часть 1

1. постановка задачи
2. ход исследования
3. обзор имеющегося опыта
4. формулировка ожидаемых результатов
5. данные и их подготовка
6. выбранная эконометрическая модель

Часть 2

1. описание данных, в т.ч. описательная статистика
2. оценка моделей, проведение тестов
3. интерпретация результатов
4. выводы

На основании оценок и комментариев преподавателя по первой части отчета, выполняется вторая часть отчета с учетом замечаний. Подробные требования к каждому пункту обозначаются преподавателем в презентации на первом занятии.

Зачет оценивается по результатам текущего контроля – выполнение промежуточных тестов по темам курса (11 тестов= 66 баллам) и по результатам индивидуального итогового индивидуального проекта (34 балла).

Критерии оценивания итогового проекта:

Часть 1 (19 баллов)

1. постановка задачи (3)
2. ход исследования (2)
3. обзор имеющегося опыта (4)
4. формулировка ожидаемых результатов (2)
5. данные и их подготовка (4)
6. выбранная эконометрическая модель (4)

Часть 2 (15 баллов)

1. описание данных, в т.ч. описательная статистика (4)
2. оценка моделей, проведение тестов (5)

3. интерпретация результатов (3)
4. выводы (3)

На основании оценок и комментариев преподавателя по первой части отчета, выполняется вторая часть отчета с учетом замечаний.

Итоговая оценка складывается из суммы баллов, набранных при прохождении тестовых заданий, предусмотренных для текущего контроля и баллов, набранных при выполнении итогового проекта:

Итоговая сумма набранных баллов	Оценка
≤ 40	неудовлетворительно
от 41 до 60	удовлетворительно
от 61 до 80	хорошо
от 81 до 100	отлично

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Тестовые вопросы

1. Какова роль EDA в жизненном цикле ML-модели? (ИОПК-1.2)
 - a) Определяет целевую переменную
 - b) Упрощает интерпретацию данных
 - c) Уменьшает количество данных
 - d) Обеспечивает более быструю обработку данных
2. Какой метод визуализации используется для оценки корреляции между двумя переменными? (ИОПК-3.1)
 - a) Гистограмма
 - b) График рассеивания
 - c) Круговая диаграмма
 - d) Линейный график
3. Что такое статистическая значимость? (ИПК-1.5)
 - a) Вероятность случайного результата
 - b) Степень уверенности в гипотезе
 - c) Разница между средними значениями
 - d) Количество проведенных тестов
4. Какой из нижеперечисленных методов относится к кодированию признаков? (ИПК-2.2)
 - a) Нормализация
 - b) Стандартизация
 - c) One-Hot Encoding
 - d) Отбор признаков

Задачи

Задача 1 (ИПК-3.3)

Вам дан набор данных с двумя признаками: возраст и доход. Какой метод корреляции вы бы использовали для анализа взаимосвязи между этими признаками, если оба являются количественными? Обоснуйте свой выбор.

Задача 2 (ИОПК-3.1)

Проведите EDA для набора данных о винных обзорах. Определите основные меры центральной тенденции и визуализируйте результаты. Какие выводы можно сделать на основе полученных данных?

Задача 3 (ИПК-1.4)

Сравните результаты A/B-тестирования для двух групп, одна из которых использовала новую рекламную стратегию. Какие статистические тесты можно применить для проверки значимости полученных результатов? Обоснуйте выбор.

Теоретические вопросы

1. Определите основные этапы разведывательного анализа данных (EDA) и их значение для машинного обучения. (ИОПК-1.2)
2. Объясните разницу между параметрическими и непараметрическими тестами. В каких случаях каждый из них следует использовать? (ИОПК-6.1)

Кейс (ИПК-1.5, ИОПК-2.2)

Описание кейса:

Представьте, что вы работаете в компании, которая разрабатывает новый продукт. Вам необходимо провести A/B-тестирование для оценки эффективности различных маркетинговых стратегий.

Задание:

1. Определите, какие метрики вы будете использовать для анализа.
2. Как вы будете формулировать гипотезу для A/B-тестирования?
3. Какие статистические тесты вы примените для анализа результатов и почему?
4. Ответ должен содержать формальную постановку задач, их решение и интерпретацию полученных выводов.

Информация о разработчиках

Салаева А.С., методист Skillfactory