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

Механико-математический факультет

УТВЕРЖДЕНО:
Декан
Л. В.Гензе

Оценочные материалы по дисциплине

Современные методы анализа и визуализации больших данных

по направлению подготовки

**01.04.01 Математика**

Направленность (профиль) подготовки:
**Моделирование и цифровые двойники**

Форма обучения
**Очная**

Квалификация
**Магистр**

Год приема
**2025**

СОГЛАСОВАНО:
Руководитель ОП
Е.И. Гурина

Председатель УМК
Е.А. Тарасов

Томск – 2025

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

Целью освоения дисциплины является формирование следующих компетенций:

ПК-2 Способен проводить тестирование, валидацию и анализ данных цифровых двойников для обеспечения их корректной работы, оптимизации процессов и принятия решений.

ПК-4 Способен документировать процессы разработки и эксплуатации цифровых двойников, работать в команде и взаимодействовать с заказчиками и специалистами для успешной реализации проектов..

УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК 2.2 Анализирует и интерпретирует данные, полученные от цифровых двойников, для принятия предиктивных решений и оптимизации процессов.

ИПК 4.3 Организует и координирует работу команды для достижения поставленных целей проекта.

ИУК 1.1 Выявляет проблемную ситуацию, на основе системного подхода осуществляет её многофакторный анализ и диагностику.

ИУК 1.2 Осуществляет поиск, отбор и систематизацию информации для определения альтернативных вариантов стратегических решений в проблемной ситуации.

ИУК 1.3 Предлагает и обосновывает стратегию действий с учетом ограничений, рисков и возможных последствий.

**2. Оценочные материалы текущего контроля и критерии оценивания**

Элементы текущего контроля:

– индивидуальные задания;

Индивидуальное задание (ИПК-2.2, ИПК 4.3, ИУК 1.2, ИУК 1.3)

В ходе разведки месторождений нефти специалисты производят пробные бурения скважин и осуществляют анализ получаемых в ходе этого технических, геологических и геофизических данных. Целью этого является обнаружение нефтенасыщенных пластов, то есть пластов, содержащих в себе нефть и способных ее отдавать.

Перед вами стоит задача разработать алгоритм интеллектуального анализа реальных данных, позволяющий наиболее качественно определять наличие или отсутствие нефтяных пластов на тех или иных глубинах залегания скважин.

Метрикой качества выступает точность нахождения нефтенасыщенного пласта

$$Accuracy = \frac{samples\_true}{sampels\_all},$$

Где *samples_true* - количество правильных предсказаний наличия/отсутствия нефтяного пласта,

*sampels_all* - общее количество записей в таблице

**Формат ввода**

data_train.csv — файл с обучающими табличными данными

X_data_predict.csv — файл с данными, для которых необходимо предсказать целевую переменную

Файл с тренировочными табличными данными содержит информацию по 600 скважинам, для каждой из которых имеется различная техническая, геологическая и геофизическая информация в виде следующих полей:

- **MD** — относительная глубина скважины (относительно поверхности бурения), всегда является положительной величиной, используется для привязки глубин внутри скважины, но не может выступать в роли какого-то признака при прогнозе (по крайней мере с физической точки зрения).
- **TVDSS** — глубина скважины относительно уровня моря, всегда является положительной величиной, может отражать поверхность геологического пласта или уровень водонефтяного контакта.
- **Layer** — название пласта, геологическая принадлежность интервала, качественная характеристика, выдаваемая геологом на основе его понимания геометрических характеристик целевого пласта, служащая для сопоставления пластов из различных скважин между собой.
- **GK** — гамма-каротаж, измеряет естественную радиоактивность пород, различные минералы имеют разное содержание радиоактивных материалов, как правило, чем выше — тем больше глинистая составляющая и меньше песчанистая, может измеряться в единицах API или мкр/ч.
- **NNKT_big** — нейтронный каротаж, регистрирует относительное водородосодержание, что может говорить о количестве пор в горных породах (они не могут быть пустыми и всегда содержат какой-то флюид, который в значительном объеме содержит в себе водород). Меньшие значения отвечают за более высокое флюидосодержание.
- **PS** — каротаж естественной поляризации, последняя возникает при фильтрации флюида через породу, уменьшение значений говорит о наличии проницаемого интервала. Единица измерения — милливольты, может иметь совершенно разный масштаб в разных скважинах.
- **IK** — индукционный каротаж, отражает электрическую проводимость горных пород, величину, обратную сопротивлению. Поскольку нефть является диэлектриком, а вода проводником, высокие показания отражают водонасыщенные пласты, а низкие — интервалы, вмещающие нефть. С другой стороны, плотные породы, не содержащие в себе пор, также имеют высокое сопротивление, поскольку не имеют в себе флюида, который способен проводить ток.
- **BK** — боковой зонд, отражает сопротивление горной породы, интерпретируется схожим образом с кривой индукционного каротажа, но уже наоборот, повышенные значения — нефть или плотные породы, пониженные — вода или глина.
- **PZ** — потенциал-зонд, отражает сопротивление горной породы, интерпретируется схожим образом с кривой индукционного каротажа, но уже наоборот, повышенные значения — нефть или плотные породы, пониженные — вода или глина. Схож с боковым зондом (BK), но имеет другую глубинность исследования.
- **Grad_zond** — другая группа зондов, отвечающих за сопротивление горных пород, в зависимости от числа в названии определяется глубинность метода. При бурении буровой раствор попадает в пласт и может изменить содержание того или иного флюида, поэтому, в теории, пониженые сопротивления в затронутой части пласта и повышенные в глубинной могут быть признаком наличия углеводородов.
- **target_collector** — бинарная характеристика, выдаваемая специалистом по интерпретации каротажных данных, отвечающая за то, является ли тот или иной интервал коллекторским пластом, то есть пластом, способным принимать и отдавать флюид.
- **target_oil** — бинарная характеристика, выдаваемая специалистом по интерпретации каротажных данных, отвечающая за то, является ли тот или иной интервал коллекторским нефтенасыщенным пластом.
- **Well** — номер скважины.

В качестве целевой переменной выступает **target_oil**, которая при значении 1 говорит о наличии нефтенасыщенного пласта, а при значении 0 — о его отсутствии.

**Формат вывода**

В файл submission.csv необходимо записать одну колонку, в которой для каждой скважины из тестовой выборки стоит классифицирующая ее метка.

Критерии оценивания:

Результаты индивидуальной работы определяются оценками «зачтено», «незачтено». При оценке выполнения индивидуальных заданий учитывается правильность, оригинальность и сроки выполнения.

По темам 1, 2, 3 и 4 каждый студент получает индивидуальное задание. Оно включает в себя некий результирующий итог по освоению материала соответствующей темы курса. Работа оформляется в виде отчёта, который студенту необходимо защитить: рассказать о ходе выполнения работы и ответить на дополнительные вопросы по теории.

Оценка «зачтено» выставляется, если содержание отчета и ответ на вопросы по теме практических заданий является содержательным, четко, ясно, кратко изложенным. Студент корректно использует изученный инструмент. В полной мере понимает как именно работают используемые им методы и/или функции, и как именно задействованы и за что отвечают основные параметры.

Оценка «незачтено» выставляется, если содержание отчета и ответ на вопросы по теме практических заданий является неполным, изложен недостаточно четко и ясно. Студент корректно использует изученный инструмент. Слабо понимает как именно работают используемые им методы и/или функции, и как именно задействованы и за что отвечают основные параметры.

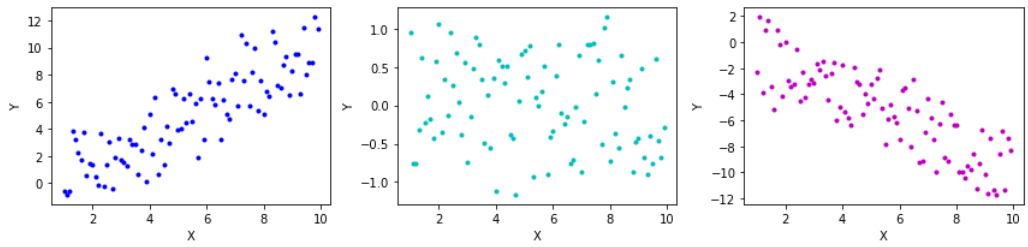
**3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания**

**Примерный перечень теоретических вопросов**

```
In [2]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
```

1. Какие описательные статистики есть и какую полезную информацию с их помощью можно узнать о данных?

2. Что такое корреляции? Какие бывают корреляции? Какую полезную информацию содержит в себе коэффициент корреляции?




3. Что такое доверительный интервал и что мы с его помощью оцениваем?

4. Что такое разведывательный анализ данных? Перечислите его основные этапы.

5. Какие проблемы в данных могут обнаружиться при их исследовании?

6. Какие способы предобработки данных вам известны? Продемонстрируйте на примере (указать не менее 3 шт.)
7. Дилемма: смещение/разброс. Переобучение, недобоучение.

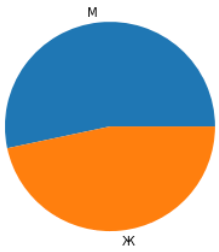
8. Для сгенерированного dataset'a построить несколько графиков

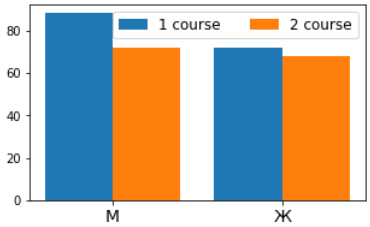
```
data = pd.DataFrame()
data['ID'] = np.array(np.floor(np.random.random(300)*10000),
dtype='int32')
data['sex'] = np.random.randint(0,2,300)
data['course'] = np.random.randint(1,3,300)
data['dis1'] = np.random.randint(30,100,300)
data['dis2'] = np.random.randint(40,100,300)
data['dis3'] = np.random.randint(35,95,300)
data.head()
```

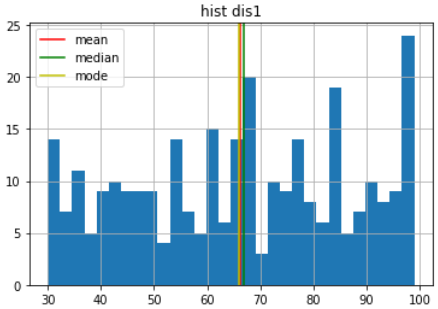
| | ID | sex | course | dis1 | dis2 | dis3 |
|---|---|---|---|---|---|---|
| **0** | 8744 | 0 | 2 | 92 | 70 | 55 |
| **1** | 97 | 0 | 2 | 47 | 54 | 90 |
| **2** | 4795 | 1 | 1 | 76 | 81 | 47 |
| **3** | 5203 | 0 | 1 | 55 | 40 | 49 |
| **4** | 1804 | 0 | 2 | 58 | 90 | 45 |

ID - идентификатор ученика
sex - пол ученика 0 - *м*, 1 - *ж*
course - курс, на котором ученик учится (1 или 2)
dis1, dis2, dis3 - результат аттестации ученика по дисциплинам

bar, pie, hist и др. (*примеры графиков ниже*)




? Какая дисциплина лучше всего даётся мальчикам, а какая девочкам на разных курсах

? Есть ли взаимосвязь между успеваемостью м\у дисциплин у мальчиков и

девочек на разных курсах

9. Задача регрессии - что это? Какие типы регрессий вы знаете? В чем их основное отличие? По каким метрикам можно оценить решение задачи регрессии?

10. Задача кластеризации - что это? В чём различие основных подходов в реализации методов кластеризации? Как мы можем оценить работу метода кластеризации?

Критерии оценивания:

К зачёту оценкой допускаются только те студенты, у которых зачтены все индивидуальные задания.

Результаты зачёта оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» выставляется, если ответ на теоретические вопросы является содержательным, четко, ясно, кратко изложенным. Студент корректно использует изученный инструмент. В полной мере понимает как именно работают используемые им методы и/или функции, и как именно задействованы и за что отвечают основные параметры.

Оценка «хорошо» выставляется, если ответ на теоретические вопросы является содержательным, однако изложен недостаточно четко, ясно. Студент корректно использует изученный инструмент. Не до конца понимает как именно работают используемые им методы и/или функции, и как именно задействованы и за что отвечают основные параметры.

Оценка «удовлетворительно» выставляется, если ответ на теоретические вопросы является неполным, изложен недостаточно четко и ясно. Студент корректно использует изученный инструмент. Слабо понимает как именно работают используемые им методы и/или функции, и как именно задействованы и за что отвечают основные параметры.

Оценка «неудовлетворительно» выставляется, если ответ на теоретические вопросы является поверхностным, изложен нечетко и неясно. Студент некорректно использует изученный инструмент. Плохо понимает как именно работают используемые им методы и/или функции, и как именно задействованы и за что отвечают основные параметры.

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

Тест (ИПК-2.2, ИУК 1.3)

1. Какое утверждение лучше всего описывает разницу между описательной и инференциальной статистикой?

A) Описательная статистика предсказывает будущие данные, инференциальная — обобщает текущие.

**B) Описательная статистика суммирует данные, инференциальная — делает выводы о популяции на основе выборки.**

C) Обе используются только для качественных данных.

D) Инференциальная статистика включает расчет среднего и медианы.

2. Коэффициент корреляции Пирсона равен -0.89. Что это означает?

A) Сильная прямая связь
B) Слабая обратная связь
**C) Сильная обратная связь**
D) Нет значимой связи

3. При увеличении размера выборки в 4 раза, ширина 95% доверительного интервала для среднего:

A) **Уменьшится**
B) Увеличится
C) Не изменится

4. Доверительный интервал для корреляции
Для r = 0.6 (n=50) 95% ДИ: [0.35, 0.77]. Как это интерпретировать?

A) 95% данных лежат в этом диапазоне
**B) С вероятностью 95% истинный коэффициент корреляции популяции в интервале**
C) Ошибка расчета
D) Интервал показывает возможные значения для выборки

5. Какая метрика НЕ используется для регрессии?

A) MSE (Mean Squared Error)
B) MAE (Mean Absolute Error)
**C) Accuracy**
D) $R^2$ (R-квадрат)

6. Модель с высоким смещением (bias) обычно:

A) Слишком сложная
**B) Недообученная (слишком простая)**
C) Идеально описывает данные
D) Имеет низкую ошибку на тесте

7. Чем отличается Ridge (L2) от Lasso (L1) регрессии?

A) Ridge уменьшает неважные коэффициенты до нуля, Lasso — нет
**B) Lasso уменьшает неважные коэффициенты до нуля, Ridge — приближает к нулю**
C) Оба метода работают одинаково
D) Ridge используется только для классификации

8. Если $R^2$=0.85 это означает, что:

A) 85% данных лежат внутри доверительного интервала
B) Все коэффициенты значимы на уровне 85%
C) Ошибка предсказания равна 15%
**D) 85% дисперсии Y объясняется моделью**

9. Что является главной задачей кластеризации?

A) Прогнозирование числовых значений
**B) Группировка объектов по схожести**
C) Классификация с учителем
*D) Удаление шумов из данных*

*Ответ D тоже возможет, потому что некоторые алгоритмы помогают маркировать зашумлённые данные.*

10. Почему для K-means важно масштабировать данные?

**A) Алгоритм чувствителен к единицам измерения**
B) Уменьшает время вычислений
C) Устраняет мультиколлинеарность
D) Требуется для визуализации

11. Что такое обучение с учителем и обучение без учителя?
A) Обучение с учителем - это задача классификации, а обучение без учителя - задача регрессии
**B) Обучение с учителем - это задача регрессии, а обучение без учителя - задача кластеризации**
C) Обучение с учителем - это задача кластеризации, а обучение без учителя - задача классификации

12. Что такое задача регрессии в машинном обучении?
**A) Прогнозирование непрерывного значения**
B) Прогнозирование категориальной метки или класса
C) Анализ временных рядов

Теоретические вопросы:

1. Что такое большие данные. 3 постулата по работе с большими данными. Откуда можно брать данные?
2. Какие методы работы с большими данными существуют?
3. Принцип работы Hadoop и Spark: в чём схожесть в чём отличие?
4. С какими проблемами можно столкнуться в процессе обработки «сырых» данных?
5. Что такое data mining?
6. Какие средства предоставляет библиотека dask по работе с большими объёмами данных?
7. Какие описательные статистики есть и какую полезную информацию с их помощью можно узнать о данных?
8. 
9. Что такое корреляции? Какие бывают корреляции? Какую полезную информацию содержит в себе коэффициент корреляции?
10. 
11. Что такое доверительный интервал и что мы с его помощью оцениваем?
12. Что такое разведывательный анализ данных? Перечислите его основные этапы.
13. Какие проблемы в данных могут обнаружиться при их исследовании?

14. Какие способы предобработки данных вам известны? Продемонстрируйте на примере (указать не менее 3 шт.)
15. Дилемма: смещение/разброс. Переобучение, недобоучение.
16. Задача регрессии - что это? Какие типы регрессий вы знаете? В чем их основное отличие? По каким метрикам можно оценить решение задачи регрессии?
17. Задача кластеризации - что это? В чём различие основных подходов в реализации методов кластеризации? Как мы можем оценить работу метода кластеризации?

**Информация о разработчиках**

Стребкова Екатерина Александровна, ст. преподаватель кафедры вычислительной математики и компьютерного моделирования ММФ ТГУ