

**Федеральное государственное бюджетное учреждение науки
Институт машиноведения им. А.А. Благонравова
Российской академии наук
(ИМАШ РАН)**

УТВЕРЖДАЮ
Заместитель директора по вопросам
подготовки кадров высшей
квалификации ИМАШ РАН
« 01 » _____ 2026 года
к.ф.-м.н. Жданов Э.Р.
М.П.



**Комплект оценочных материалов
по дисциплине «Наука о данных для медицины и биотехнологий»**

Направление подготовки:	02.04.03 Математическое обеспечение и администрирование информационных систем
Профиль подготовки:	Искусственный интеллект и интеллектуальная обработка данных
Квалификация:	Магистр
Форма обучения	Очная

Содержание

1. Общие положения.....	3
2. Результаты освоения дисциплины, подлежащие проверке.....	3
3. Контроль и оценка освоения учебной дисциплины.....	4
4. Оценочные материалы для проведения текущего контроля успеваемости.....	4
5. Оценочные материалы для проведения промежуточной аттестации.....	6

1. Общие положения

Комплект оценочных материалов (далее - ОМ) предназначен для контроля и оценки образовательных достижений обучающихся, освоивших программу дисциплины «Наука о данных для медицины и биотехнологий».

ОМ включают контрольные материалы для проведения текущего контроля успеваемости и промежуточной аттестации в форме зачета.

ОМ разработаны на основании:

- основной профессиональной образовательной программы по направлению подготовки 02.04.03 Математическое обеспечение и администрирование информационных систем;

- рабочей программы дисциплины «Наука о данных для медицины и биотехнологий».

2. Результаты освоения дисциплины, подлежащие проверке

В результате освоения дисциплины обучающийся должен обладать следующими компетенциями

ПК-1	Создание информационных технологий нового поколения
ПК-1.2	Управление разработкой продуктов, услуг и решений на основе больших данных
ПК-1.3	Разработка и внедрение новых методов и технологий исследования больших данных

Код индикатора компетенции	Планируемые результаты обучения по дисциплине
ПК-1.2	<i>Знать:</i> - жизненный цикл разработки продуктов на основе анализа медицинских данных; - методологии управления проектами в области Data Science (CRISP-DM); - требования к качеству и безопасности медицинских информационных систем. <i>Уметь:</i> - декомпозировать сложную бизнес-задачу медицинского учреждения на подзадачи анализа данных; - формировать техническое задание для команды разработчиков на сбор и обработку данных; - оценивать риски и ресурсы при реализации проектов с использованием больших данных. <i>Иметь практический опыт:</i> - участия в планировании проекта по созданию прототипа системы поддержки принятия врачебных решений; - управления требованиями к данным на всех этапах их жизненного цикла в рамках учебного проекта.
ПК-1.3	<i>Знать:</i> - современные методы машинного обучения и глубокого обучения, применяемые в биомедицине; - алгоритмы оптимизации гиперпараметров моделей; - подходы к интерпретации результатов работы сложных моделей ("Explainable AI"). <i>Уметь:</i> - адаптировать существующие алгоритмы машинного обучения под специфику биомедицинских задач; - разрабатывать пайплайны обработки данных от их загрузки до получения финального прогноза; - сравнивать эффективность различных моделей и выбирать оптимальную для конкретной задачи.

	<i>Иметь практический опыт:</i> - самостоятельной разработки и тестирования нового метода или модификации существующего алгоритма для решения прикладной задачи; - внедрения разработанного решения в виде программного модуля или скрипта и оценки его производительности на реальных или синтетических данных.
--	--

3. Контроль и оценка освоения учебной дисциплины

Элемент учебной дисциплины	Текущий контроль		Промежуточная аттестация	
	Форма контроля	Проверяемые идк	Форма контроля	Проверяемые идк
Тема 1. Введение в науку о данных в медицине и биотехнологиях	Отчет по практикуму по решению задач	ПК-1.2 ПК-1.3		
Тема 2. Математические основы и методы машинного обучения	Отчет по практикуму по решению задач	ПК-1.2 ПК-1.3		
Тема 3. Анализ изображений и сигналов в медицине	Отчет по практикуму по решению задач	ПК-1.2 ПК-1.3		
Тема 4. Биоинформатика и оптимизация биотехнологических процессов	Отчет по практикуму по решению задач	ПК-1.2 ПК-1.3		
			Зачет	ПК-1.2 ПК-1.3

4. Оценочные материалы для проведения текущего контроля успеваемости

Показатели и критерии оценивания компетенций на различных этапах их формирования, описание шкал оценивания

Практикум по решению задач
25-21 – работа и отчет выполнены в срок, самостоятельно, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, сделаны необходимые выводы, хорошо аргументированы, даны исчерпывающие ответы на все поставленные вопросы; 20-15 – работа и отчет выполнены в срок, самостоятельно, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, необходимые выводы сделаны частично, хорошо аргументированы, даны ответы на все поставленные вопросы; 14-9 – работа и отчет выполнены в срок, в основном самостоятельно, выводы сделаны частично, слабо аргументированы, даны ответы не на все вопросы; 8-1 – обучающийся подготовил работу и отчет несамоостоятельно и/или не завершил в срок, описание содержит ошибки, выводы и ответы на вопросы отсутствуют.

Порядок начисления баллов

Если студент набрал достаточное количество баллов ТКУ по дисциплине (более 70 баллов) в течение семестра, он может отказаться от сдачи зачета, и результат будет выставлен автоматически по его согласию.

Типовые контрольные задания и иные материалы, необходимые для процедуры оценивания знаний, умений, навыков и(или) опыта деятельности

Тема 1. Введение в науку о данных в медицине и биотехнологиях

Практикум 1. Первичный анализ и визуализация данных электронных медицинских

карт (ЭМК)

Задача:

Загрузить предоставленный датасет ЭМК (например, в формате CSV). Определить типы данных, количество записей и признаков. Вывести первые 10 строк и общую информацию о датасете.

Для каждого признака вычислить долю пропущенных значений. Построить тепловую карту (heatmap) для визуализации распределения пропусков по датасету.

Для количественных признаков (например, возраст, уровень глюкозы, артериальное давление) построить диаграммы размаха (boxplot) для выявления потенциальных выбросов.

Построить гистограмму распределения возраста пациентов.

Для категориальных признаков (например, пол, группа крови, диагноз) построить столбчатые диаграммы частот.

Проанализировать распределение целевого показателя (например, наличие/отсутствие заболевания).

Тема 2. Математические основы и методы машинного обучения

Практикум 2. Построение и оценка модели для прогнозирования риска заболевания

Задача:

На основе датасета из Практикума 1 сформировать матрицу признаков (X) и целевую переменную (y). Разделить данные на обучающую (70%) и тестовую (30%) выборки с сохранением распределения классов (stratified split).

Обучить модель логистической регрессии на обучающей выборке.

Используя тестовую выборку, рассчитать следующие метрики: Матрицу ошибок (Confusion Matrix).

Точность (Accuracy), полноту (Recall), точность (Precision).

F1-меру.

Площадь под ROC-кривой (ROC-AUC).

Обучить ту же модель на всей выборке и сравнить её метрики на обучающих и тестовых данных. Объяснить наблюдаемую разницу с точки зрения переобучения.

Применить метод кластеризации K-means к данным пациентов (без использования целевой переменной) для выделения групп пациентов со схожими характеристиками. Проанализировать полученные кластеры.

Тема 3. Анализ изображений и сигналов в медицине

Практикум 3. Детекция патологий на медицинских изображениях с помощью нейронных сетей

Задача:

Загрузить и изучить датасет, состоящий из рентгеновских снимков лёгких, разделённых на два класса: «Норма» и «Патология» (например, пневмония).

Привести все изображения к единому размеру, нормализовать значения пикселей.

Загрузить предобученную модель (например, ResNet50, VGG16), "заморозить" её сверточные слои и добавить поверх них новый классификатор для вашей задачи из двух классов.

Провести обучение нового классификатора на подготовленном датасете изображений.

Оценить точность модели на отложенной выборке. С помощью техники Grad-CAM или других методов визуализации внимания (attention maps) продемонстрировать, на какие области изображения модель обращала внимание при принятии решения.

Тема 4. Биоинформатика и оптимизация биотехнологических процессов

Практикум 4. Поиск биомаркеров по данным экспрессии генов и предсказание выхода продукта

Задача:

Загрузить датасет, содержащий данные об уровнях экспрессии тысяч генов для двух групп образцов: «Здоровые» и «Больные».

Провести нормализацию данных.

Использовать методы отбора признаков (feature selection) для выявления набора генов, наиболее сильно коррелирующих с заболеванием (потенциальных биомаркеров).

Построить классификатор (например, на основе логистической регрессии или SVM) для разделения образцов на группы и оценить его точность.

Моделирование биотехнологического процесса (задача регрессии). Загрузить датасет, описывающий параметры работы биореактора (температура, pH, концентрация субстрата) и итоговый выход целевого продукта (например, концентрация антибиотика).

Исследовать корреляции между параметрами процесса и выходом продукта.

Построить регрессионную модель для предсказания выхода продукта на основе параметров процесса.

Оценить качество модели с помощью метрик RMSE (среднеквадратичная ошибка) и R^2 (коэффициент детерминации).

5. Оценочные материалы для проведения промежуточной аттестации

Промежуточная аттестация по дисциплине проводится в форме зачета.

Зачет представляет собой выполнение обучающимся заданий билета, включающего в себя:

Задание №1 – теоретический вопрос на знание базовых понятий предметной области дисциплины, а также позволяющий оценить степень владения обучающегося принципами предметной области дисциплины, понимание их особенностей и взаимосвязи между ними.

Задание №2 – задание на анализ ситуации из предметной области дисциплины и выявление способности обучающегося выбирать и применять соответствующие принципы и методы решения практических проблем, близких к профессиональной деятельности.

Задания №3 – задания на проверку умений и навыков, полученных в результате освоения дисциплины.

Выполнение обучающимся заданий билета оценивается по следующей балльной шкале:

Задание 1: 0-30 баллов

Задание 2: 0-30 баллов

Задание 3: 0-40 баллов

«Зачтено»

– 90-100 – ответ правильный, логически выстроен, использована профессиональная терминология. Задания решены правильно. Обучающийся правильно интерпретирует полученный результат.

– 70 -89 – ответ в целом правильный, логически выстроен, использована профессиональная терминология. Ход решения заданий правильный, ответ неверный. Обучающийся в целом правильно интерпретирует полученный результат.

– 50 - 69 – ответ в основном правильный, логически выстроен, использована профессиональная терминология. Задание решено частично.

«Не зачтено»

– менее 50 – ответы на теоретическую часть неправильные или неполные. Задания не решены.

Типовые задания для проведения промежуточной аттестации обучающихся

Задания 1 типа

1. Перечислите основные источники данных в медицине и биотехнологиях (ЭМК, носимые устройства, геномные данные) и охарактеризуйте специфику каждого из них.

2. Что такое «проблема качества данных» в контексте медицины? Приведите примеры «шума» и «пропусков» в электронных медицинских картах (ЭМК).

3. Какие методы визуализации данных (графики, диаграммы) наиболее эффективны для анализа распределения показателей здоровья и почему?
4. Опишите алгоритм выявления аномалий (выбросов) в медицинских данных. Как отличить клинически значимую аномалию от ошибки ввода?
5. Каковы основные положения Федерального закона № 152-ФЗ «О персональных данных» в части обработки медицинской информации? Что такое «деперсонализация» данных?
6. В чём заключаются основные этические вызовы при сборе и анализе данных с носимых устройств (фитнес-трекеров, умных часов) для медицинских целей?
7. Объясните принцип работы логистической регрессии. Почему эта модель является базовой для задач бинарной классификации в медицине (например, болен/здоров)?
8. Дайте определения метрикам оценки качества классификации: чувствительность (Recall), специфичность, точность (Precision) и F1-мера. В каких клинических ситуациях приоритетной является высокая чувствительность, а в каких — высокая специфичность?
9. Что такое ROC-кривая и площадь под ней (ROC-AUC)? Как этот показатель помогает оценить диагностическую эффективность модели независимо от выбранного порога классификации?
10. Сформулируйте проблему переобучения (overfitting). Какие методы существуют для её предотвращения при построении прогностических моделей?
11. Опишите принцип работы алгоритма кластеризации K-means. Как можно применить этот метод для сегментации пациентов на группы риска без использования заранее известных диагнозов?
12. Объясните важность разделения данных на обучающую, валидационную и тестовую выборки при разработке модели прогнозирования риска заболевания.
13. Опишите архитектуру сверточной нейронной сети (CNN). Какие операции (свертка, пулинг) позволяют ей эффективно анализировать медицинские изображения (рентген, МРТ)?
14. Что такое трансферное обучение (transfer learning) и как оно применяется при разработке систем для детекции патологий на медицинских снимках с использованием предобученных нейронных сетей?
15. В чём заключается специфика анализа временных рядов в медицине (например, сигналов ЭКГ или ЭЭГ)? Какие задачи (классификация, сегментация, прогнозирование) можно решать с их помощью?
16. Что такое биомаркер? Опишите общую схему анализа данных экспрессии генов для поиска биомаркеров конкретного заболевания.
17. Как методы анализа данных применяются для оптимизации биотехнологических процессов (например, в ферментере)? Приведите пример задачи предсказания выхода целевого продукта.
18. Объясните разницу между задачами классификации и регрессии на примере моделирования кинетики роста клеточных культур.

Задания 2 типа

1. Вы получили доступ к анонимизированной базе данных электронных медицинских карт (ЭМК) региональной больницы для анализа. Первичный осмотр показал, что в поле «Уровень холестерина» у 30% пациентов старше 60 лет стоят пропуски, а в поле «Диагноз» встречаются опечатки (например, «Гепотит В» вместо «Гепатит В»). Опишите пошаговый план действий по подготовке этого датасета к анализу. Какие методы вы примените для обработки пропусков и очистки категориальных данных?
2. Фармацевтическая компания хочет провести исследование эффективности нового препарата, используя данные с носимых устройств (фитнес-браслетов) добровольцев. Данные включают пульс, уровень активности и качество сна. Какие основные этические и юридические риски (в контексте 152-ФЗ) вы видите в этом проекте? Какие технические меры необходимо предпринять для обеспечения анонимности и безопасности данных

участников?

3. При анализе данных геномного секвенирования для поиска мутаций, связанных с редким заболеванием, вы обнаружили, что объем данных слишком велик для обработки на одном компьютере, а качество прочтений (риды) в некоторых участках генома низкое

Каковы ваши действия? Опишите, как вы будете решать проблему больших данных и как — проблему низкого качества данных перед началом анализа.

4. Вам поставлена задача разработать модель, которая по данным анамнеза (возраст, ИМТ, курение, наследственность) будет предсказывать риск развития сахарного диабета 2-го типа в ближайшие 5 лет. Какую метрику качества модели вы выберете в качестве основной и почему? Как вы будете бороться с проблемой дисбаланса классов, если в вашей выборке здоровых людей значительно больше, чем заболевших?

5. Вы построили модель логистической регрессии для предсказания риска инфаркта. На обучающей выборке точность модели составила 98%, однако на тестовой выборке она упала до 70%. Как называется эта проблема? Объясните ее причину и предложите, как минимум два способа решения для улучшения обобщающей способности модели.

6. У вас есть данные о тысячах пациентов с онкологическим заболеванием, включая их генетический профиль и реакцию на различные виды терапии. Ваша цель — разделить всех пациентов на группы, чтобы для каждой группы подобрать наиболее эффективное лечение.

Какой класс задач машинного обучения здесь применим? Какой алгоритм вы бы выбрали в качестве базового и как бы вы оценивали качество полученной сегментации пациентов?

7. Больница закупила систему на основе нейросети для анализа рентгеновских снимков грудной клетки. Система показывает высокую точность ($AUC=0.95$), но врачи не понимают, почему она ставит тот или иной диагноз, что вызывает недоверие. Как называется эта проблема («черный ящик»)? Какие подходы и методы вы бы предложили для повышения интерпретируемости решений этой нейросети?

8. У вас есть небольшая база данных из 500 рентгеновских снимков легких с маркировкой «норма/пневмония». Для создания высокоточной модели этого недостаточно. Какую стратегию вы выберете для обучения модели? Опишите, как вы будете использовать предобученные нейронные сети и какие шаги предпримете для дообучения.

9. Необходимо разработать систему мониторинга состояния пациента по данным ЭКГ в реальном времени. Система должна не просто записывать сигнал, а автоматически выявлять аномалии (например, аритмию). В чем специфика анализа временных рядов, таких как ЭКГ, по сравнению с анализом табличных данных? Какой подход к моделированию (классификация, сегментация) вы бы использовали?

10. Биотехнологический стартап производит биотопливо с помощью микроорганизмов в биореакторе. Стоимость производства высока. У вас есть исторические данные о процессе: температура, pH, концентрация питательных веществ и итоговый выход продукта. Сформулируйте задачу анализа данных для этой ситуации. Какой тип модели (регрессия) вы бы построили и как бы использовали ее результаты для оптимизации процесса и снижения затрат?

11. Ученые провели анализ экспрессии генов у двух групп мышей: здоровой и с моделью нейродегенеративного заболевания. Они получили список из тысяч генов с измененной экспрессией. Какой следующий шаг анализа данных вы бы предложили? Какими статистическими методами можно сократить этот список до набора наиболее значимых генов-кандидатов в биомаркеры?

12. При анализе геномных данных необходимо найти участки, общие для большого числа штаммов бактерий, чтобы разработать универсальную вакцину. Какую задачу биоинформатики необходимо решить? Какие алгоритмы множественного выравнивания последовательностей (MSA) или другие методы анализа вы бы применили?

13. Фармацевтическая компания имеет данные о миллионах химических соединений и их активности против определенной биологической мишени. Цель — найти новые

перспективные соединения для создания лекарства. Опишите подход к решению этой задачи с помощью методов машинного обучения. Какую задачу вы будете решать (классификацию или регрессию) и какие особенности больших данных нужно учесть?

14. В больнице внедряется система поддержки принятия врачебных решений (СППВР). Она должна на основе текущих показателей пациента (из ЭМК) и его истории болезни выдавать прогноз риска осложнений после операции. Из каких этапов будет состоять ваш рабочий процесс по созданию такой системы? От сбора и подготовки данных до выбора модели и ее оценки.

15. Вы работаете с данными носимого устройства, которое собирает показания ЭКГ и акселерометра. Цель — отличить истинный сигнал сердца от артефактов движения во время физической активности пациента. Предложите методологию решения этой задачи. Какие признаки можно извлечь из сигналов и какой алгоритм классификации будет наиболее подходящим?

Задания 3 типа

1. Провести первичный анализ предоставленного датасета ЭМК.

Загрузить данные, определить долю пропущенных значений в каждом столбце. Найти и вывести 3 примера аномальных значений (например, возраст > 120 лет или отрицательный пульс) с помощью логических условий.

2. Проанализировать распределение одного из показателей здоровья.

Для столбца «Уровень глюкозы» построить гистограмму распределения. Описать форму распределения (нормальное, скошенное) и визуально оценить наличие выбросов.

3. Построить модель для бинарной классификации.

Разделить данные на признаки (X) и целевую переменную (y). Обучить модель логистической регрессии на 80% данных и вывести точность (ассурагу) на оставшихся 20%.

4. Оценить качество предсказаний модели.

Используя предсказания модели и истинные значения из Задания 3, построить матрицу ошибок (confusion matrix) и рассчитать Precision и Recall.

5. Сегментировать пациентов на группы по медицинским показателям.

Применить алгоритм K-means для разделения данных на 3 кластера. Вывести центроиды (средние значения) кластеров по двум признакам (например, «Возраст» и «Систолическое давление»).

6. Дообучить предобученную нейросеть для классификации снимков.

Используя предобученную модель (например, MobileNetV2), дообучить ее на небольшом наборе данных (например, «Норма/Пневмония») в течение 3–5 эпох. Вывести финальную точность на валидационной выборке.

7. Найти пики в сигнале ЭКГ.

Для предоставленного массива данных ЭКГ найти индексы пиков (R-зубцов) с помощью функции `find_peaks` из библиотеки `scipy.signal`.

Навыки: Обработка сигналов, использование научных библиотек Python.

8. Найти гены с разной экспрессией у больных и здоровых.

Даны два массива значений экспрессии генов. Применить t-тест для сравнения групп и вывести топ-5 генов с наименьшим p-value (наиболее статистически значимые).

9. Предсказать выход продукта в биореакторе.

По табличным данным с параметрами процесса построить модель линейной регрессии для предсказания выхода продукта. Вывести коэффициент детерминации (R^2) модели.

10. Объяснить предсказание модели для одного пациента.

Используя модель из Задания 3, применить метод SHAP для расчета вклада каждого признака в предсказание для одного конкретного пациента. Вывести график `summary plot` или `bar plot`, показывающий влияние признаков.