



**САМАРСКИЙ
ПОЛИТЕХ**
Опорный университет

МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Самарский государственный технический университет»
(ФГБОУ ВО «СамГТУ»)



УТВЕРЖДАЮ:

Первый проректор-проректор по
учебной работе

/ Д.Е. Овчинников

09 20 26 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.02.07 «Анализ данных и машинное обучение»

Код и направление подготовки (специальность)	02.03.03 Математическое обеспечение и администрирование информационных систем
Направленность (профиль)	Анализ данных в бизнес-аналитике
Квалификация	Бакалавр
Форма обучения	Очная
Год начала подготовки	2026
Институт / факультет	Институт инженерно-экономического и гуманитарного образования
Выпускающая кафедра	кафедра "Высшая математика"
Кафедра-разработчик	кафедра "Высшая математика"
Объем дисциплины, ч. / з.е.	288 / 8
Форма контроля (промежуточная аттестация)	Экзамен

Содержание

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы	4
2. Место дисциплины (модуля) в структуре образовательной программы	5
3. Объем дисциплины (модуля) в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся	5
4. Содержание дисциплины (модуля), структурированное по темам (разделам), с указанием отведенного на них количества академических часов и видов учебных занятий	6
4.1 Содержание лекционных занятий	6
4.2 Содержание лабораторных занятий	13
4.3 Содержание практических занятий	20
4.4. Содержание самостоятельной работы	20
5. Перечень учебной литературы и учебно-методического обеспечения по дисциплине (модулю)	22
6. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения	22
7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», профессиональных баз данных, информационно-справочных систем	23
8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)	23
9. Методические материалы	23
10. Фонд оценочных средств по дисциплине (модулю)	25

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

Наименование категории (группы) компетенций	Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Результаты обучения (знать, уметь, владеть, соотнесенные с индикаторами достижения компетенции)
Профессиональные компетенции			
Не предусмотрено	ПК-2 Способность проектировать и разрабатывать программные средства работы с базами данных современными инструментальными средствами и технологиями	ПК-2.1 Знать: средства СУБД для выявления проблем производительности при выполнении и повышении пропускной способности баз данных	Знать Структуры, форматы и типы данных и методы их обработки.
		ПК-2.2 Уметь: осуществлять поиск, хранение, обработку и анализ информации и баз данных, представление их в требуемом формате	Уметь Осуществлять конкатенацию и локализацию данных по булевой маске, применять средства визуализации.
		ПК-2.3 Владеть: проектированием баз данных, поддержанием схемы БД в соответствии с изменениями в требованиях и предметной области	Владеть Методами разведочного анализа, статистическими критериями для подтверждения рабочих гипотез.
Универсальные компетенции			
Системное и критическое мышление	УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Знать: основы системного и критического анализа информации	Знать Библиотеки Python для анализа данных.
		УК-1.2 Уметь: разрабатывать стратегию действий, принимать конкретные решения для ее реализации	Уметь Проводить предварительную обработку данных, выбирать библиотеки и методы в соответствии с решаемой задачей.
		УК-1.3 Владеть: методиками постановки цели, определения способов ее достижения, разработки стратегий действий	Владеть Методиками разведочного (EDA) и подтверждающего (CDA) анализа данных.

2. Место дисциплины (модуля) в структуре образовательной программы

Место дисциплины (модуля) в структуре образовательной программы: **часть, формируемая участниками образовательных отношений**

Код компетенции	Предшествующие дисциплины	Параллельно осваиваемые дисциплины	Последующие дисциплины
ПК-2	Визуализация данных; Основы работы с NoSQL-базами данных	Большие данные; Информационная бизнес-аналитика; Нейронные сети	Инженерия данных; Компьютерное зрение; Подготовка к процедуре защиты и защита выпускной квалификационной работы; Технологии интернета вещей
УК-1	Введение в цифровую экономику; Визуализация данных; Объектно-ориентированное программирование; Основы работы с NoSQL-базами данных; Основы экономического анализа; Производственная практика: научно-исследовательская работа; Учебная практика: научно-исследовательская работа (получение первичных навыков научно-исследовательской работы); Учебная практика: проектная практика; Финансовый менеджмент; Экономика предприятия; Языки программирования	Большие данные; Методология разработки DevOps; Нейронные сети; Производственная практика: научно-исследовательская работа; Развертывание и жизненный цикл программного обеспечения; Разработка и принятие управленческих решений; Стратегическое планирование; Цифровая трансформация бизнеса; Экономическое развитие и инновационная экономика	Инженерия данных; Компьютерное зрение; Моделирование бизнес-процессов; Подготовка к процедуре защиты и защита выпускной квалификационной работы; Производственная практика: преддипломная практика; Технологии интернета вещей

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

Вид учебной работы	Всего часов / часов в электронной форме	6 семестр часов / часов в электронной форме	7 семестр часов / часов в электронной форме
Аудиторная контактная работа (всего), в том числе:	128	64	64
Лабораторные работы	64	32	32
Лекции	64	32	32
Внеаудиторная контактная работа, КСР	8	4	4
Самостоятельная работа (всего), в том числе:	40	40	0
выполнение контрольных работ	10	10	0
подготовка к практическим занятиям	10	10	0
подготовка к экзамену	20	20	0

Контроль	72	36	36
Итого: час	288	144	144
Итого: з.е.	8	4	4

4. Содержание дисциплины (модуля), структурированное по темам (разделам), с указанием отведенного на них количества академических часов и видов учебных занятий

№ раздела	Наименование раздела дисциплины	Виды учебной нагрузки и их трудоемкость, часы				
		ЛЗ	ЛР	ПЗ	СРС	Всего часов
1	Основы анализа данных	16	16	0	20	52
2	Обучение с учителем	16	16	0	20	52
3	Обучение без учителя	16	16	0	20	52
4	Интеграция моделей в информационные системы	16	16	0	20	52
	КСР	0	0	0	0	8
	Контроль	0	0	0	0	72
	Итого	64	64	0	80	288

4.1 Содержание лекционных занятий

№ занятия	Наименование раздела	Тема лекции	Содержание лекции (перечень дидактических единиц: рассматриваемых подтем, вопросов)	Количество часов / часов в электронной форме
6 семестр				
1	Основы анализа данных	Экосистема Data Science	Введение в дисциплину. Обзор библиотек numpy, pandas, matplotlib, seaborn, sklearn. Набор данных titanic как модельный объект. Построение визуализаций. Формулирование гипотез. Визуальная верификация гипотез без применения формальных статистических процедур. Понятие Data Science как процесса выдвижения и проверки гипотез.	2
2	Основы анализа данных	Типы и структуры данных	Определение типов данных в объекте DataFrame. Обработка некорректно определенных типов. Сравнительная характеристика форматов хранения данных: CSV, JSON, Parquet. Списки как фундаментальная структура данных. Словари: определение, методы доступа, области применения. Множества: свойства, операции. Генераторы списков и словарей: синтаксис и типовые задачи.	2

3	Основы анализа данных	Векторные вычисления	Массивы библиотеки numpy. Сравнительный анализ производительности: цикл for на интерпретируемом языке и векторизованные операции. Принципы реализации векторизованных вычислений. Идентификация объектов: список языка Python, массив numpy, серия pandas — критерии различия. Ограничения применения цикла for при обработке строк объекта DataFrame. Сравнительная характеристика итераций по строкам и столбцам.	2
4	Основы анализа данных	Локализации, конкатенации и метод apply	Методы доступа к элементам DataFrame: iloc, loc. Конкатенация объектов DataFrame: функция concat, параметры и типовые сценарии. Генерация новых признаков: метод apply с параметром axis=1. Вычисление сводных статистик: метод apply с параметром axis=0.	2
5	Основы анализа данных	Группировки и агрегации	Группировка данных: метод groupby. Агрегация сгруппированных данных: метод agg. Синтаксический сахар при выполнении агрегаций. Трансформация сгруппированных данных: метод transform.	2
6	Основы анализа данных	Визуализация в matplotlib	Назначение визуализации в исследовательской деятельности. Разграничение подходов к визуализации: исследовательский и отчетный. Интерфейс pyplot: функции plot, hist, scatter. Объектно-ориентированный интерфейс: фигура и оси — fig, ax = plt.subplots(). Управление параметрами графических объектов.	2
7	Основы анализа данных	Визуализация в seaborn	Библиотека seaborn: архитектура и взаимосвязь с pandas и matplotlib. Построение графиков для разведочного анализа данных. Параметры context и style: управление внешним видом визуализаций.	2
8	Основы анализа данных	Разведочный анализ EDA+CDA	Фазы разведочного анализа данных. Статистические сводки. Идентификация пропусков. Методы обработки пропусков: заполнение, удаление. Выявление и обработка выбросов. Визуализация в рамках разведочного анализа. Генерация гипотез. Разведочный анализ как самостоятельный результат аналитической работы. Подтверждающий анализ данных: статистические критерии. Уровень значимости и p-value. Библиотека scipy.stats: импорт и применение. Комплексный пример EDA+CDA.	2

9	Обучение с учителем	Прогнозирующие модели	Понятие стохастического прогнозирования. Модель как объект. Обучение модели. Функция потерь: определение и назначение. Прогноз как результат применения обученной модели. Линейная регрессия как интуитивно понятная модель. Классификатор на основе метода ближайших соседей (kNN).	2
10	Обучение с учителем	Метрики прогнозирования	Переобучение модели: понятие и признаки. Метрики регрессионных моделей: средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE), коэффициент детерминации (R^2). Принцип разделения данных на обучающую и тестовую выборки. Вычисление метрик исключительно на тестовой выборке. Синтетический датасет как инструмент демонстрации. Матрица ошибок (confusion matrix): структура и интерпретация. Метрики классификации: accuracy, precision, recall, F1-мера. Примеры вычисления метрик классификации. Небинарная классификация: особенности оценки качества.	2
11	Обучение с учителем	Нормализация и кодирование данных	Необходимость приведения признаков к единому масштабу. Методы нормализации: нормализация по минимаксу (MinMaxScaler), стандартизация (StandardScaler), робастная нормализация (RobustScaler). Влияние нормализации на линейную регрессию: обоснование отсутствия эффекта. Влияние нормализации на метод ближайших соседей: обоснование чувствительности. Линейность и нелинейность моделей как фактор влияния нормализации. Кодирование категориальных данных: Label Encoding, One-Hot Encoding, Target Encoding. Примеры: регрессия с предварительным кодированием категориальных признаков; бинарная классификация с предварительным кодированием категориальных признаков и нормализацией.	2
12	Обучение с учителем	Регуляризация	Понятие регуляризации как метода борьбы с переобучением. L1-регуляризация (Lasso): принцип, механизм обнуления коэффициентов. L2-регуляризация (Ridge): принцип, штрафование больших коэффициентов. ElasticNet: комбинация L1 и L2 регуляризации. Взвешивание расстояний в методе kNN: весовые функции, влияние на прогноз. Параметр регуляризации: выбор и интерпретация.	2

13	Обучение с учителем	Регрессия на ближайших соседях и логистическая регрессия	Регрессия на основе метода k ближайших соседей (kNN Regressor): принцип, использование среднего значения соседей. Логистическая регрессия как модель бинарной классификации. Сигмоидная функция: математическое описание и интерпретация. Граница принятия решения. Функция потерь для логистической регрессии (логистическая потеря, кросс-энтропия). Сравнительный анализ kNN-регрессии и логистической регрессии: области применения, ограничения.	2
14	Обучение с учителем	Ансамблевые методы классификации	Деревья принятия решений (DecisionTreeClassifier): принцип построения, критерии расщепления, управление глубиной. Случайный лес (RandomForestClassifier): ансамблирование деревьев, бутстреп-агрегация (бэггинг), случайный подбор признаков. Градиентный бустинг (GradientBoostingClassifier): последовательное обучение слабых моделей, коррекция ошибок предшественников. Сравнительный анализ ансамблевых методов классификации: точность, склонность к переобучению, вычислительная сложность.	2
15	Обучение с учителем	Ансамблевые методы регрессии	Деревья принятия решений для регрессии (DecisionTreeRegressor): принцип, среднеквадратическая ошибка как критерий расщепления. Случайный лес для регрессии (RandomForestRegressor): усреднение прогнозов деревьев. Градиентный бустинг для регрессии (GradientBoostingRegressor): последовательная коррекция остатков. Сравнительный анализ ансамблевых методов регрессии: показатели точности, интерпретируемость, требования к данным.	2
16	Обучение с учителем	Подбор гиперпараметров модели и кросс-валидация	Понятие гиперпараметров: отличие от параметров модели. Методы подбора гиперпараметров: ручной перебор, поиск по сетке (GridSearchCV), случайный поиск (RandomizedSearchCV). Кросс-валидация: принцип, разделение данных на блоки. K-fold кросс-валидация: схема, выбор количества блоков. Стратифицированная кросс-валидация. Оценка качества модели на основе кросс-валидации. Предотвращение утечки данных при подборе гиперпараметров.	2
Итого за семестр:				32
7 семестр				

17	Обучение без учителя	Введение в обучение без учителя. Кластеризация: k-means	Понятие обучения без учителя: отличие от обучения с учителем (отсутствие целевой переменной). Задача кластеризации: разбиение объектов на группы по принципу максимального внутригруппового сходства и межгруппового различия. Алгоритм k-means: итеративное перемещение центроидов; визуализация процесса схождения. Выбор числа кластеров (k): метод локтя, силуэтный коэффициент (интуитивная интерпретация). Ограничения k-means: чувствительность к начальным центрам, масштабу признаков, выбросам.	2
18	Обучение без учителя	Кластеризация: DBSCAN и иерархическая кластеризация	Мотивация: неспособность k-means выделять кластеры произвольной формы и необходимость априорного задания k. DBSCAN: плотностная кластеризация; параметры eps и min_samples; идентификация шумовых точек. Иерархическая агломеративная кластеризация: дендрограмма как инструмент визуализации и выбора k. Сравнительный анализ методов: целесообразность применения в зависимости от структуры данных.	2
19	Обучение без учителя	Метрики качества кластеризации	Внутренние метрики (без эталонных меток): силуэтный коэффициент, индекс Калински-Харабаса, индекс Дэвиса-Болдина. Внешние метрики (при наличии истинных меток): Rand Index, Adjusted Rand Index, Normalized Mutual Information. Проблема «проклятия размерности»: несостоятельность метрик расстояния в пространствах высокой размерности; мотивация к снижению размерности.	2
20	Обучение без учителя	Сингулярное разложение	Напоминание: разложение матрицы на U , Σ , V^T . Геометрическая интерпретация SVD: поворот, масштабирование (сингулярные числа), повторный поворот. Сингулярные числа как мера «важности» направления. Усечённое SVD: наилучшее приближение матрицы ранга k (теорема Эккарта-Янга).	2
21	Обучение без учителя	PCA через SVD	Связь PCA и SVD: PCA как SVD центрированной матрицы данных. Алгоритм: центрирование, SVD, отбор k компонент, проекция. Доля объяснённой дисперсии (explained variance ratio). Выбор числа компонент. Визуализация: проекция многомерных данных на две главные компоненты.	2

22	Обучение без учителя	PCA: интерпретация, кластеризация и нелинейные обобщения	Интерпретация главных компонент через loadings: выявление признаков с наибольшим вкладом. Кластеризация в пространстве главных компонент как способ борьбы с «проклятием размерности». Нелинейные методы снижения размерности: t-SNE и UMAP (анонс, общая идея, отличие от PCA).	2
23	Обучение без учителя	Коллаборативная фильтрация на основе SVD	Постановка задачи рекомендательных систем: предсказание рейтинга пользователя для товара. Типы рекомендательных систем: коллаборативная фильтрация (поведение пользователей) и контентная фильтрация (свойства товаров). Метод SVD для разреженной матрицы рейтингов: разложение, заполнение пропусков, формирование рекомендаций. Демонстрация на малом примере: восстановление пропущенных значений.	2
24	Обучение без учителя	Рекомендательные системы	Продвинутые подходы: SVD++ (учёт implicit feedback), нейросетевые архитектуры (обзорно). Проблема холодного старта: новые пользователи и новые товары; гибридные подходы как решение. Этические аспекты: пузырь фильтров, поляризация, манипуляция потребительским поведением, проблема fairness (неравномерное качество рекомендаций для разных групп пользователей).	2
25	Интеграция моделей в информационные системы	Продвинутые подходы: SVD++ (учёт implicit feedback), нейросетевые архитектуры (обзорно). Проблема холодного старта: новые пользователи и новые товары; гибридные подходы как решение. Этические аспекты: пузырь фильтров, поляризация, манипуляция потребительским поведением, проблема fairness (неравномерное качество рекомендаций для разных групп пользователей).	Общая структура пайплайна данных. Место модели в корпоративной информационной системе. Понятия ETL-процессов, API-шлюзов, баз данных и хранилищ. Роль модели как вычислительного модуля, принимающего решения на основе входных данных. Классификация систем по степени интеграции ML-компонентов (batch-режим, real-time, streaming).	2

26	Интеграция моделей в информационные системы	Основы языка SQL для извлечения данных	Синтаксис запросов SELECT, JOIN, GROUP BY, HAVING, ORDER BY. Подзапросы и CTE. Фильтрация и агрегация данных. Типы данных в SQL. Практическая реализация запросов для извлечения признаков модели. Формирование датасета из реляционной базы данных.	2
27	Интеграция моделей в информационные системы	Библиотеки Python для взаимодействия с базами данных	Подключение к PostgreSQL через psycopg2. Выполнение SQL-запросов из Python-скрипта. Использование SQLAlchemy для ORM-абстракции. Преобразование результатов запроса в структуры данных pandas DataFrame. Обработка ошибок соединения и транзакций.	2
28	Интеграция моделей в информационные системы	Развертывание и администрирование PostgreSQL	Установка и настройка сервера PostgreSQL. Работа через Docker-контейнер. Создание базы данных, схем, пользователей. Проектирование простой реляционной схемы: таблицы для хранения исходных данных, признаков и предсказаний. Понятие миграций схемы.	2
29	Интеграция моделей в информационные системы	REST API: принципы и реализация на FastAPI	Архитектурный стиль REST. HTTP-методы (GET, POST, PUT, DELETE). Понятие эндпоинта, пути запроса, тела запроса и ответа. Разработка простого API на FastAPI: GET-эндпоинт для получения данных из базы, POST-эндпоинт для передачи признаков и получения предсказания. Валидация входных данных с помощью Pydantic.	2
30	Интеграция моделей в информационные системы	Контейнеризация приложений с использованием Docker	Понятие контейнера и образа. Dockerfile для сервиса на FastAPI. Docker Compose для оркестрации сервисов (API + база данных). Сборка и запуск multi-container приложения. Сетевые взаимодействия между контейнерами.	2
31	Интеграция моделей в информационные системы	Построение пайплайна: извлечение — предсказание — запись	Интеграция компонентов в единый конвейер. Скрипт-оркестратор: последовательность извлечения данных из БД, трансформации признаков, передачи в модель, получения предсказания, записи результата в таблицу predictions. Логирование этапов пайплайна. Обработка ошибок и fallback-логика.	2
32	Интеграция моделей в информационные системы	Эксплуатация ML-системы и принятие управленческих решений	Чтение предсказаний из БД внешними системами (дашборды, BI-инструменты). Интерпретация результатов модели для бизнес-пользователя. Понятие мониторинга дрейфа данных и качества предсказаний. Границы ответственности ML-системы: роль человека в принятии окончательного решения.	2
Итого за семестр:				32

Итого:	64
---------------	-----------

4.2 Содержание лабораторных занятий

№ занятия	Наименование раздела	Тема лабораторного занятия	Содержание лабораторного занятия (перечень дидактических единиц: рассматриваемых подтем, вопросов)	Количество часов / часов в электронной форме
6 семестр				
1	Основы анализа данных	Экосистема Data Science	Введение в дисциплину. Обзор библиотек numpy, pandas, matplotlib, seaborn, sklearn. Набор данных titanic как модельный объект. Построение визуализаций. Формулирование гипотез. Визуальная верификация гипотез без применения формальных статистических процедур. Понятие Data Science как процесса выдвижения и проверки гипотез.	2
2	Основы анализа данных	Типы и структуры данных	Определение типов данных в объекте DataFrame. Обработка некорректно определенных типов. Сравнительная характеристика форматов хранения данных: CSV, JSON, Parquet. Списки как фундаментальная структура данных. Словари: определение, методы доступа, области применения. Множества: свойства, операции. Генераторы списков и словарей: синтаксис и типовые задачи.	2
3	Основы анализа данных	Векторные вычисления	Массивы библиотеки numpy. Сравнительный анализ производительности: цикл for на интерпретируемом языке и векторизованные операции. Принципы реализации векторизованных вычислений. Идентификация объектов: список языка Python, массив numpy, серия pandas — критерии различия. Ограничения применения цикла for при обработке строк объекта DataFrame. Сравнительная характеристика итераций по строкам и столбцам.	2
4	Основы анализа данных	Локализации, конкатенации и метод apply	Методы доступа к элементам DataFrame: iloc, loc. Конкатенация объектов DataFrame: функция concat, параметры и типовые сценарии. Генерация новых признаков: метод apply с параметром axis=1. Вычисление сводных статистик: метод apply с параметром axis=0.	2
5	Основы анализа данных	Группировки и агрегации	Группировка данных: метод groupby. Агрегация сгруппированных данных: метод agg. Синтаксический сахар при выполнении агрегаций. Трансформация сгруппированных данных: метод transform.	2

6	Основы анализа данных	Визуализация в matplotlib	Назначение визуализации в исследовательской деятельности. Разграничение подходов к визуализации: исследовательский и отчетный. Интерфейс pyplot: функции plot, hist, scatter. Объектно-ориентированный интерфейс: фигура и оси — fig, ax = plt.subplots(). Управление параметрами графических объектов.	2
7	Основы анализа данных	Визуализация в seaborn	Библиотека seaborn: архитектура и взаимосвязь с pandas и matplotlib. Построение графиков для разведочного анализа данных. Параметры context и style: управление внешним видом	2
8	Основы анализа данных	Разведочный анализ EDA+CDA	Фазы разведочного анализа данных. Статистические сводки. Идентификация пропусков. Методы обработки пропусков: заполнение, удаление. Выявление и обработка выбросов. Визуализация в рамках разведочного анализа. Генерация гипотез. Разведочный анализ как самостоятельный результат аналитической работы. Подтверждающий анализ данных: статистические критерии. Уровень значимости и p-value. Библиотека scipy.stats: импорт и применение. Комплексный пример EDA+CDA. Общая структура пайплайна данных. Место модели в корпоративной информационной системе. Понятия ETL-процессов, API-шлюзов, баз данных и хранилищ. Роль модели как вычислительного модуля, принимающего решения на основе входных данных. Классификация систем по степени интеграции ML-компонентов (batch-режим, real-time, streaming).	2
9	Обучение с учителем	Прогнозирующие модели	Понятие стохастического прогнозирования. Модель как объект. Обучение модели. Функция потерь: определение и назначение. Прогноз как результат применения обученной модели. Линейная регрессия как интуитивно понятная модель. Классификатор на основе метода ближайших соседей (kNN).	2

10	Обучение с учителем	Метрики прогнозирования	Переобучение модели: понятие и признаки. Метрики регрессионных моделей: средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE), коэффициент детерминации (R^2). Принцип разделения данных на обучающую и тестовую выборки. Вычисление метрик исключительно на тестовой выборке. Синтетический датасет как инструмент демонстрации. Матрица ошибок (confusion matrix): структура и интерпретация. Метрики классификации: accuracy, precision, recall, F1-мера. Примеры вычисления метрик классификации. Небинарная классификация: особенности оценки качества.	2
11	Обучение с учителем	Нормализация и кодирование данных	Необходимость приведения признаков к единому масштабу. Методы нормализации: нормализация по минимаксу (MinMaxScaler), стандартизация (StandardScaler), робастная нормализация (RobustScaler). Влияние нормализации на линейную регрессию: обоснование отсутствия эффекта. Влияние нормализации на метод ближайших соседей: обоснование чувствительности. Линейность и нелинейность моделей как фактор влияния нормализации. Кодирование категориальных данных: Label Encoding, One-Hot Encoding, Target Encoding. Примеры: регрессия с предварительным кодированием категориальных признаков; бинарная классификация с предварительным кодированием категориальных признаков и нормализацией.	2
12	Обучение с учителем	Регуляризация	Понятие регуляризации как метода борьбы с переобучением. L1-регуляризация (Lasso): принцип, механизм обнуления коэффициентов. L2-регуляризация (Ridge): принцип, штрафование больших коэффициентов. ElasticNet: комбинация L1 и L2 регуляризации. Взвешивание расстояний в методе kNN: весовые функции, влияние на прогноз. Параметр регуляризации: выбор и интерпретация.	2

13	Обучение с учителем	Регрессия на ближайших соседях и логистическая регрессия	Регрессия на основе метода k ближайших соседей (kNN Regressor): принцип, использование среднего значения соседей. Логистическая регрессия как модель бинарной классификации. Сигмоидная функция: математическое описание и интерпретация. Граница принятия решения. Функция потерь для логистической регрессии (логистическая потеря, кросс-энтропия). Сравнительный анализ kNN-регрессии и логистической регрессии: области применения, ограничения.	2
14	Обучение с учителем	Ансамблевые методы классификации	Деревья принятия решений (DecisionTreeClassifier): принцип построения, критерии расщепления, управление глубиной. Случайный лес (RandomForestClassifier): ансамблирование деревьев, бутстреп-агрегация (бэггинг), случайный подбор признаков. Градиентный бустинг (GradientBoostingClassifier): последовательное обучение слабых моделей, коррекция ошибок предшественников. Сравнительный анализ ансамблевых методов классификации: точность, склонность к переобучению, вычислительная сложность.	2
15	Обучение с учителем	Ансамблевые методы регрессии	Деревья принятия решений для регрессии (DecisionTreeRegressor): принцип, среднеквадратическая ошибка как критерий расщепления. Случайный лес для регрессии (RandomForestRegressor): усреднение прогнозов деревьев. Градиентный бустинг для регрессии (GradientBoostingRegressor): последовательная коррекция остатков. Сравнительный анализ ансамблевых методов регрессии: показатели точности, интерпретируемость, требования к данным.	2
16	Обучение с учителем	Подбор гиперпараметров модели и кросс-валидация	Понятие гиперпараметров: отличие от параметров модели. Методы подбора гиперпараметров: ручной перебор, поиск по сетке (GridSearchCV), случайный поиск (RandomizedSearchCV). Кросс-валидация: принцип, разделение данных на блоки. K-fold кросс-валидация: схема, выбор количества блоков. Стратифицированная кросс-валидация. Оценка качества модели на основе кросс-валидации. Предотвращение утечки данных при подборе гиперпараметров.	2
Итого за семестр:				32
7 семестр				

17	Обучение без учителя	Введение в обучение без учителя. Кластеризация: k-means	Понятие обучения без учителя: отличие от обучения с учителем (отсутствие целевой переменной). Задача кластеризации: разбиение объектов на группы по принципу максимального внутригруппового сходства и межгруппового различия. Алгоритм k-means: итеративное перемещение центроидов; визуализация процесса схождения. Выбор числа кластеров (k): метод локтя, силуэтный коэффициент (интуитивная интерпретация). Ограничения k-means: чувствительность к начальным центрам, масштабу признаков, выбросам.	2
18	Обучение без учителя	Кластеризация: DBSCAN и иерархическая кластеризация	Мотивация: неспособность k-means выделять кластеры произвольной формы и необходимость априорного задания k. DBSCAN: плотностная кластеризация; параметры eps и min_samples; идентификация шумовых точек. Иерархическая агломеративная кластеризация: дендрограмма как инструмент визуализации и выбора k. Сравнительный анализ методов: целесообразность применения в зависимости от структуры данных.	2
19	Обучение без учителя	Метрики качества кластеризации	Внутренние метрики (без эталонных меток): силуэтный коэффициент, индекс Калински-Харабаса, индекс Дэвиса-Болдина. Внешние метрики (при наличии истинных меток): Rand Index, Adjusted Rand Index, Normalized Mutual Information. Проблема «проклятия размерности»: несостоятельность метрик расстояния в пространствах высокой размерности; мотивация к снижению размерности.	2
20	Обучение без учителя	Сингулярное разложение	Напоминание: разложение матрицы на U , Σ , V^T . Геометрическая интерпретация SVD: поворот, масштабирование (сингулярные числа), повторный поворот. Сингулярные числа как мера «важности» направления. Усечённое SVD: наилучшее приближение матрицы ранга k (теорема Эккарта-Янга).	2
21	Обучение без учителя	PCA через SVD	Связь PCA и SVD: PCA как SVD центрированной матрицы данных. Алгоритм: центрирование, SVD, отбор k компонент, проекция. Доля объяснённой дисперсии (explained variance ratio). Выбор числа компонент. Визуализация: проекция многомерных данных на две главные компоненты.	2

22	Обучение без учителя	PCA: интерпретация, кластеризация и нелинейные обобщения	Интерпретация главных компонент через loadings: выявление признаков с наибольшим вкладом. Кластеризация в пространстве главных компонент как способ борьбы с «проклятием размерности». Нелинейные методы снижения размерности: t-SNE и UMAP (анонс, общая идея, отличие от PCA).	2
23	Обучение без учителя	Коллаборативная фильтрация на основе SVD	Постановка задачи рекомендательных систем: предсказание рейтинга пользователя для товара. Типы рекомендательных систем: коллаборативная фильтрация (поведение пользователей) и контентная фильтрация (свойства товаров). Метод SVD для разреженной матрицы рейтингов: разложение, заполнение пропусков, формирование рекомендаций. Демонстрация на малом примере: восстановление пропущенных значений.	2
24	Обучение без учителя	Рекомендательные системы	Продвинутые подходы: SVD++ (учёт implicit feedback), нейросетевые архитектуры (обзорно). Проблема холодного старта: новые пользователи и новые товары; гибридные подходы как решение. Этические аспекты: пузырь фильтров, поляризация, манипуляция потребительским поведением, проблема fairness (неравномерное качество рекомендаций для разных групп пользователей).	2
25	Интеграция моделей в информационные системы	Продвинутые подходы: SVD++ (учёт implicit feedback), нейросетевые архитектуры (обзорно). Проблема холодного старта: новые пользователи и новые товары; гибридные подходы как решение. Этические аспекты: пузырь фильтров, поляризация, манипуляция потребительским поведением, проблема fairness (неравномерное качество рекомендаций для разных групп пользователей).	Общая структура пайплайна данных. Место модели в корпоративной информационной системе. Понятия ETL-процессов, API-шлюзов, баз данных и хранилищ. Роль модели как вычислительного модуля, принимающего решения на основе входных данных. Классификация систем по степени интеграции ML-компонентов (batch-режим, real-time, streaming).	2

26	Интеграция моделей в информационные системы	Основы языка SQL для извлечения данных	Синтаксис запросов SELECT, JOIN, GROUP BY, HAVING, ORDER BY. Подзапросы и CTE. Фильтрация и агрегация данных. Типы данных в SQL. Практическая реализация запросов для извлечения признаков модели. Формирование датасета из реляционной базы данных.	2
27	Интеграция моделей в информационные системы	Библиотеки Python для взаимодействия с базами данных	Подключение к PostgreSQL через psycopg2. Выполнение SQL-запросов из Python-скрипта. Использование SQLAlchemy для ORM-абстракции. Преобразование результатов запроса в структуры данных pandas DataFrame. Обработка ошибок соединения и транзакций.	2
28	Интеграция моделей в информационные системы	Развертывание и администрирование PostgreSQL	Установка и настройка сервера PostgreSQL. Работа через Docker-контейнер. Создание базы данных, схем, пользователей. Проектирование простой реляционной схемы: таблицы для хранения исходных данных, признаков и предсказаний. Понятие миграций схемы.	2
29	Интеграция моделей в информационные системы	REST API: принципы и реализация на FastAPI	Архитектурный стиль REST. HTTP-методы (GET, POST, PUT, DELETE). Понятие эндпоинта, пути запроса, тела запроса и ответа. Разработка простого API на FastAPI: GET-эндпоинт для получения данных из базы, POST-эндпоинт для передачи признаков и получения предсказания. Валидация входных данных с помощью Pydantic.	2
30	Интеграция моделей в информационные системы	Контейнеризация приложений с использованием Docker	Понятие контейнера и образа. Dockerfile для сервиса на FastAPI. Docker Compose для оркестрации сервисов (API + база данных). Сборка и запуск multi-container приложения. Сетевые взаимодействия между контейнерами.	2
31	Интеграция моделей в информационные системы	Построение пайплайна: извлечение — предсказание — запись	Интеграция компонентов в единый конвейер. Скрипт-оркестратор: последовательность извлечения данных из БД, трансформации признаков, передачи в модель, получения предсказания, записи результата в таблицу predictions. Логирование этапов пайплайна. Обработка ошибок и fallback-логика.	2
32	Интеграция моделей в информационные системы	Эксплуатация ML-системы и принятие управленческих решений	Чтение предсказаний из БД внешними системами (дашборды, BI-инструменты). Интерпретация результатов модели для бизнес-пользователя. Понятие мониторинга дрейфа данных и качества предсказаний. Границы ответственности ML-системы: роль человека в принятии окончательного решения.	2
Итого за семестр:				32

	Итого:	64
--	---------------	-----------

4.3 Содержание практических занятий

Учебные занятия не реализуются.

4.4. Содержание самостоятельной работы

Наименование раздела	Вид самостоятельной работы	Содержание самостоятельной работы (перечень дидактических единиц: рассматриваемых подтем, вопросов)	Количество часов
6 семестр			
Основы анализа данных	Подготовка к коллоквиуму	Экосистема Data Science. Типы и структуры данных. Векторные вычисления. Локализации, конкатенации и метод apply. Группировки и агрегации. Визуализация в matplotlib. Визуализация в seaborn. Визуализация в seaborn.	10
Основы анализа данных	Подготовка к контрольной работе.	Экосистема Data Science. Типы и структуры данных. Векторные вычисления. Локализации, конкатенации и метод apply. Группировки и агрегации. Визуализация в matplotlib. Визуализация в seaborn. Визуализация в seaborn.	10
Обучение с учителем	Подготовка к коллоквиуму	Прогнозирующие модели. Метрики прогнозирования. Нормализация и кодирование данных. Регуляризация. Регрессия на ближайших соседях и логистическая регрессия. Ансамблевые методы классификации. Ансамблевые методы регрессии. Подбор гиперпараметров модели и кросс-валидация	10
Обучение с учителем	Подготовка к контрольной работе.	Прогнозирующие модели. Метрики прогнозирования. Нормализация и кодирование данных. Регуляризация. Регрессия на ближайших соседях и логистическая регрессия. Ансамблевые методы классификации. Ансамблевые методы регрессии. Подбор гиперпараметров модели и кросс-валидация.	10
Итого за семестр:			40
7 семестр			

Обучение без учителя	Подготовка к коллоквиуму	Введение в обучение без учителя. Кластеризация: k-means. Кластеризация: DBSCAN и иерархическая кластеризация. Метрики качества кластеризации. Сингулярное разложение. PCA через SVD. PCA: интерпретация, кластеризация и нелинейные обобщения. Коллаборативная фильтрация на основе SVD. Рекомендательные системы.	10
Обучение без учителя	Подготовка к контрольной работе.	Введение в обучение без учителя. Кластеризация: k-means. Кластеризация: DBSCAN и иерархическая кластеризация. Метрики качества кластеризации. Сингулярное разложение. PCA через SVD. PCA: интерпретация, кластеризация и нелинейные обобщения. Коллаборативная фильтрация на основе SVD. Рекомендательные системы.	10
Интеграция моделей в информационные системы	Подготовка к коллоквиуму	Архитектура информационных систем с компонентами машинного обучения. Основы языка SQL для извлечения данных. Библиотеки Python для взаимодействия с базами данных. Развертывание и администрирование PostgreSQL. REST API: принципы и реализация на FastAPI. Контейнеризация приложений с использованием Docker. Построение пайплайна: извлечение — предсказание — запись. Эксплуатация ML-системы и принятие управленческих решений.	10
Интеграция моделей в информационные системы	Подготовка к контрольной работе.	Архитектура информационных систем с компонентами машинного обучения. Основы языка SQL для извлечения данных. Библиотеки Python для взаимодействия с базами данных. Развертывание и администрирование PostgreSQL. REST API: принципы и реализация на FastAPI. Контейнеризация приложений с использованием Docker. Построение пайплайна: извлечение — предсказание — запись. Эксплуатация ML-системы и принятие управленческих решений.	10
Итого за семестр:			40
Итого:			80

5. Перечень учебной литературы и учебно-методического обеспечения по

дисциплине (модулю)

№ п/п	Библиографическое описание	Ресурс НТБ СамГТУ (ЭБС СамГТУ, IPRbooks и т.д.)
Основная литература		
1	Алпайдин, Э. Машинное обучение: Новый искусственный интеллект : пер. с англ. / Э. Алпайдин.- М., Точка, 2017Альпина Паблишер.- 191 с.	Книжный фонд
2	Чубукова, И.А. Data Mining : Учеб.пособие / И. А. Чубукова .- 2-е изд.,испр.- М., Интернет-Ун-т Информ.Технологий, 2008М., БИНОМ.Лаб.знаний.- 382 с.	Книжный фонд
Дополнительная литература		
3	Python и анализ данных: практическое пособие / Маккинли У., Профобразование, пер. Слинкина А.: 2024.- Режим доступа: https://elib.samgtu.ru/getinfo?uid=els_samgtu iprbooks 145897	Электронный ресурс
4	Обработка данных в Python. Основы работы с библиотекой Pandas: учебно-методическое пособие / Титов А.Н., Тазиева Р.Ф., Издательство КНИТУ: 2022.- Режим доступа: https://elib.samgtu.ru/getinfo?uid=els_samgtu iprbooks 129244	Электронный ресурс
5	Тюгашев, А.А. Введение в машинное обучение и искусственный интеллект : учебное пособие / А. А. Тюгашев; Самарский государственный технический университет, Вычислительная техника.- Самара, 2025.- 208 с.- Режим доступа: https://elib.samgtu.ru/getinfo?uid=els_samgtu elib 6192	Электронный ресурс
Учебно-методическое обеспечение		
6	Батищева, О.М. Информационно-программное обеспечение анализа данных : учеб.пособие / О. М. Батищева; Самар.гос.техн.ун-т.- Самара, 2010.- 173 с.	Книжный фонд
7	Кудаева, Ф.Х. Python и анализ данных : учебное пособие / Ф. Х. Кудаева, А. Р. Бечелова, А. А. Кайгермазов; Кабардино-Балкарский государственный университет им. Х. М. Бербекова.- Нальчик, КБГУ, 2024.- 193 с.	Книжный фонд

Доступ обучающихся к ЭР НТБ СамГТУ (elib.samgtu.ru) осуществляется посредством электронной информационной образовательной среды университета и сайта НТБ СамГТУ по логину и паролю.

6. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения

При проведении лекционных занятий используется мультимедийное оборудование.

Организовано взаимодействие обучающегося и преподавателя с использованием электронной ин-формационной образовательной среды университета.

№ п/п	Наименование	Производитель	Способ распространения
1	Антиплагиат. Вуз	АО "Антиплагиат" (Отечественный)	Лицензионное

2	Python	Python Software Foundation (PSF) (Зарубежный)	Свободно распространяемое
3	Adobe Reader	Adobe Systems Incorporated (Зарубежный)	Свободно распространяемое

7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», профессиональных баз данных, информационно-справочных систем

№ п/п	Наименование	Краткое описание	Режим доступа
1	Научная электронная библиотека	http://elibrary.ru/	Ресурсы открытого доступа
2	Электронная библиотека изданий СамГТУ	http://irbis.samgtu.local/cgi-bin/irbis64r_01/cgiirbis_64.exe	Российские базы данных ограниченного доступа

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Лекционные занятия

Аудитории для лекционных занятий укомплектованы мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории (наборы демонстрационного оборудования (проектор, экран, компьютер/ноутбук), учебно-наглядные пособия, тематические иллюстрации).

Лабораторные занятия

Компьютерный класс с доступом в локальную сеть и Интернет, оснащенный компьютерами, редактором программного кода и средствами демонстрации учебных материалов.

Самостоятельная работа

Помещения для самостоятельной работы оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и доступом к электронной информационно-образовательной среде СамГТУ:

- читальный зал НТБ СамГТУ (ауд. 200 корпус № 8; ауд. 125 корпус № 1; ауд. 41, 31, 34, 35, Главный корпус библиотеки, ауд. 83а, 414, 416, 0209 АСА СамГТУ; ауд. 401 корпус №10);
- компьютерные классы (ауд. 208, 210 корпус № 8).

9. Методические материалы

Методические рекомендации при работе на лекции

До лекции студент должен просмотреть учебно-методическую и научную литературу по теме лекции с тем, чтобы иметь представление о проблемах, которые будут разбираться в лекции.

Перед началом лекции обучающимся сообщается тема лекции, план, вопросы, подлежащие рассмотрению, доводятся основные литературные источники. Весь учебный материал, сообщаемый преподавателем, должен не просто прослушиваться. Он должен быть активно воспринят, т.е. услышан, осмыслен, понят, зафиксирован на бумаге и закреплён в памяти. Приступая к слушанию нового учебного материала, полезно мысленно установить его связь с ранее изученным. Следя за техникой чтения лекции (акцент на существенном, повышение тона, изменение ритма, пауза и т.п.), необходимо вслед за преподавателем уметь выделять основные категории, законы и определять их содержание, проблемы, предполагать их возможные решения, доказательства и выводы. Осуществляя такую работу, можно значительно облегчить себе понимание учебного материала, его конспектирование и дальнейшее изучение.

Конспектирование лекции позволяет обработать, систематизировать и лучше сохранить полученную информацию с тем, чтобы в будущем можно было восстановить в памяти основные, содержательные моменты. Типичная ошибка, совершаемая обучающимся, дословное конспектирование речи преподавателя. Как правило, при записи «слово в слово» не остается времени на обдумывание, анализ и синтез информации. Отбирая нужную информацию, главные мысли, проблемы, решения и выводы, необходимо сокращать текст, строить его таким образом, чтобы потом можно было легко в нем разобраться. Желательно оставить в рабочих конспектах поля, на которых можно будет делать пометки из рекомендованной литературы, дополняющие материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений. С окончанием лекции работа над конспектом не может считаться завершённой. Нужно еще восстановить отдельные места, проверить, все ли понятно, уточнить что-то на консультации и т.п. с тем, чтобы конспект мог быть использован в процессе подготовки к практическим занятиям, зачету, экзамену. Конспект лекции – незаменимый учебный документ, необходимый для самостоятельной работы.

Методические рекомендации при работе на лабораторном занятии

Проведение лабораторной работы делится на две условные части: теоретическую и практическую.

Необходимыми структурными элементами занятия являются проведение лабораторной работы, проверка усвоенного материала, включающая обсуждение теоретических основ выполняемой работы.

Перед лабораторной работой, как правило, проводится технико-теоретический инструктаж по использованию необходимого оборудования. Преподаватель корректирует деятельность обучающегося в процессе выполнения работы (при необходимости). После завершения лабораторной работы подводятся итоги, обсуждаются результаты деятельности.

Возможны следующие формы организации лабораторных работ: фронтальная, групповая и индивидуальная. При фронтальной форме выполняется одна и та же работа (при этом возможны различные варианты заданий). При групповой форме работа выполняется группой (командой). При индивидуальной форме обучающимися выполняются индивидуальные работы.

По каждой лабораторной работе имеются методические указания по их выполнению, включающие необходимый теоретический и практический материал, содержащие элементы и последовательную инструкцию по проведению выбранной работы, индивидуальные варианты заданий, требования и форму отчетности по данной работе.

Методические рекомендации по выполнению самостоятельной работы

Организация самостоятельной работы обучающихся ориентируется на активные методы овладения знаниями, развитие творческих способностей, переход от поточного к индивидуализированному обучению с учетом потребностей и возможностей обучающегося.

Самостоятельная работа с учебниками, учебными пособиями, научной, справочной литературой, материалами периодических изданий и Интернета является наиболее эффективным методом получения дополнительных знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала. Все новые понятия по изучаемой теме необходимо выучить наизусть и внести в глоссарий, который целесообразно вести с самого начала изучения курса.

Самостоятельная работа реализуется:

- непосредственно в процессе аудиторных занятий;

- на лекциях, практических занятиях;
- в контакте с преподавателем вне рамок расписания;
- на консультациях по учебным вопросам, в ходе творческих контактов, при ликвидации задолженностей, при выполнении индивидуальных заданий и т.д.;
- в библиотеке, дома, на кафедре при выполнении обучающимся учебных и практических задач.

Эффективным средством осуществления обучающимся самостоятельной работы является электронная информационно-образовательная среда университета, которая обеспечивает доступ к учебным планам, рабочим программам дисциплин (модулей), практик, к изданиям электронных библиотечных систем.

10. Фонд оценочных средств по дисциплине (модулю)

Фонд оценочных средств представлен в приложении № 1.

Приложение 1 к рабочей программе дисциплины
Б1.В.02.07 «Анализ данных и машинное обучение»

**Фонд оценочных средств
по дисциплине
Б1.В.02.07 «Анализ данных и машинное обучение»**

Код и направление подготовки (специальность)	02.03.03 Математическое обеспечение и администрирование информационных систем
Направленность (профиль)	Анализ данных в бизнес-аналитике
Квалификация	Бакалавр
Форма обучения	Очная
Год начала подготовки	2026
Институт / факультет	Институт инженерно-экономического и гуманитарного образования
Выпускающая кафедра	кафедра "Высшая математика"
Кафедра-разработчик	кафедра "Высшая математика"
Объем дисциплины, ч. / з.е.	288 / 8
Форма контроля (промежуточная аттестация)	Экзамен

**Перечень планируемых результатов обучения по дисциплине (модулю),
соотнесенных с планируемыми результатами освоения образовательной
программы**

Наименование категории (группы) компетенций	Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Результаты обучения (знать, уметь, владеть, соотнесенные с индикаторами достижения компетенции)
Профессиональные компетенции			
Не предусмотрено	ПК-2 Способность проектировать и разрабатывать программные средства работы с базами данных современными инструментальными средствами и технологиями	ПК-2.1 Знать: средства СУБД для выявления проблем производительности при выполнении и повышении пропускной способности баз данных	Знать Структуры, форматы и типы данных и методы их обработки.
		ПК-2.2 Уметь: осуществлять поиск, хранение, обработку и анализ информации и баз данных, представление их в требуемом формате	Уметь Осуществлять конкатенацию и локализацию данных по булевой маске, применять средства визуализации.
		ПК-2.3 Владеть: проектированием баз данных, поддержанием схемы БД в соответствии с изменениями в требованиях и предметной области	Владеть Методами разведочного анализа, статистическими критериями для подтверждения рабочих гипотез.
Универсальные компетенции			
Системное и критическое мышление	УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Знать: основы системного и критического анализа информации	Знать Библиотеки Python для анализа данных.
		УК-1.2 Уметь: разрабатывать стратегию действий, принимать конкретные решения для ее реализации	Уметь Проводить предварительную обработку данных, выбирать библиотеки и методы в соответствии с решаемой задачей.
		УК-1.3 Владеть: методиками постановки цели, определения способов ее достижения, разработки стратегий действий	Владеть Методиками разведочного (EDA) и подтверждающего (CDA) анализа данных.

Матрица соответствия оценочных средств запланированным результатам обучения

Код индикатора достижения компетенции	Результаты обучения	Оценочные средства	Текущий контроль успеваемости	Промежуточная аттестация
Основы анализа данных				
ПК-2.1 Знать: средства СУБД для выявления проблем производительности при выполнении и повышении пропускной способности баз данных	Знать Структуры, форматы и типы данных и методы их обработки.	Экзамен	Да	Да
ПК-2.2 Уметь: осуществлять поиск, хранение, обработку и анализ информации и баз данных, представление их в требуемом формате	Уметь Осуществлять конкатенацию и локализацию данных по булевой маске, применять средства визуализации.	Экзамен	Да	Да
ПК-2.3 Владеть: проектированием баз данных, поддержанием схемы БД в соответствии с изменениями в требованиях и предметной области	Владеть Методами разведочного анализа, статистическими критериями для подтверждения рабочих гипотез.	Экзамен	Да	Да
УК-1.1 Знать: основы системного и критического анализа информации	Знать Библиотеки Python для анализа данных.	Экзамен	Да	Да
УК-1.2 Уметь: разрабатывать стратегию действий, принимать конкретные решения для ее реализации	Уметь Проводить предварительную обработку данных, выбирать библиотеки и методы в соответствии с решаемой задачей.	Экзамен	Да	Да
УК-1.3 Владеть: методиками постановки цели, определения способов ее достижения, разработки стратегий действий	Владеть Методиками разведочного (EDA) и подтверждающего (CDA) анализа данных.	Экзамен	Да	Да
Обучение с учителем				
ПК-2.1 Знать: средства СУБД для выявления проблем производительности при выполнении и повышении пропускной способности баз данных	Знать Структуры, форматы и типы данных и методы их обработки.	Экзамен	Да	Да
ПК-2.2 Уметь: осуществлять поиск, хранение, обработку и анализ информации и баз данных, представление их в требуемом формате	Уметь Осуществлять конкатенацию и локализацию данных по булевой маске, применять средства визуализации.	Экзамен	Да	Да

ПК-2.3 Владеть: проектированием баз данных, поддержанием схемы БД в соответствии с изменениями в требованиях и предметной области	Владеть Методами разведочного анализа, статистическими критериями для подтверждения рабочих гипотез.	Экзамен	Да	Да
УК-1.1 Знать: основы системного и критического анализа информации	Знать Библиотеки Python для анализа данных.	Экзамен	Да	Да
УК-1.2 Уметь: разрабатывать стратегию действий, принимать конкретные решения для ее реализации	Уметь Проводить предварительную обработку данных, выбирать библиотеки и методы в соответствии с решаемой задачей.	Экзамен	Да	Да
УК-1.3 Владеть: методиками постановки цели, определения способов ее достижения, разработки стратегий действий	Владеть Методиками разведочного (EDA) и подтверждающего (CDA) анализа данных.	Экзамен	Да	Да
Обучение без учителя				
ПК-2.1 Знать: средства СУБД для выявления проблем производительности при выполнении и повышении пропускной способности баз данных	Знать Структуры, форматы и типы данных и методы их обработки.	Экзамен	Да	Да
ПК-2.2 Уметь: осуществлять поиск, хранение, обработку и анализ информации и баз данных, представление их в требуемом формате	Уметь Осуществлять конкатенацию и локализацию данных по булевой маске, применять средства визуализации.	Экзамен	Да	Да
ПК-2.3 Владеть: проектированием баз данных, поддержанием схемы БД в соответствии с изменениями в требованиях и предметной области	Владеть Методами разведочного анализа, статистическими критериями для подтверждения рабочих гипотез.	Экзамен	Да	Да
УК-1.1 Знать: основы системного и критического анализа информации	Знать Библиотеки Python для анализа данных.	Экзамен	Да	Да
УК-1.2 Уметь: разрабатывать стратегию действий, принимать конкретные решения для ее реализации	Уметь Проводить предварительную обработку данных, выбирать библиотеки и методы в соответствии с решаемой задачей.	Экзамен	Да	Да

УК-1.3 Владеть: методиками постановки цели, определения способов ее достижения, разработки стратегий действий	Владеть Методиками разведочного (EDA) и подтверждающего (CDA) анализа данных.	Экзамен	Да	Да
Интеграция моделей в информационные системы				
ПК-2.1 Знать: средства СУБД для выявления проблем производительности при выполнении и повышении пропускной способности баз данных	Знать Структуры, форматы и типы данных и методы их обработки.	Экзамен	Да	Да
ПК-2.2 Уметь: осуществлять поиск, хранение, обработку и анализ информации и баз данных, представление их в требуемом формате	Уметь Осуществлять конкатенацию и локализацию данных по булевой маске, применять средства визуализации.	Экзамен	Да	Да
ПК-2.3 Владеть: проектированием баз данных, поддержанием схемы БД в соответствии с изменениями в требованиях и предметной области	Владеть Методами разведочного анализа, статистическими критериями для подтверждения рабочих гипотез.	Экзамен	Да	Да
УК-1.1 Знать: основы системного и критического анализа информации	Знать Библиотеки Python для анализа данных.	Экзамен	Да	Да
УК-1.2 Уметь: разрабатывать стратегию действий, принимать конкретные решения для ее реализации	Уметь Проводить предварительную обработку данных, выбирать библиотеки и методы в соответствии с решаемой задачей.	Экзамен	Да	Да
УК-1.3 Владеть: методиками постановки цели, определения способов ее достижения, разработки стратегий действий	Владеть Методиками разведочного (EDA) и подтверждающего (CDA) анализа данных.	Экзамен	Да	Да

**ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ ИЛИ ИНЫЕ МАТЕРИАЛЫ,
НЕОБХОДИМЫЕ ДЛЯ ОЦЕНКИ ЗНАНИЙ, УМЕНИЙ, НАВЫКОВ И (ИЛИ)
ОПЫТА ДЕЯТЕЛЬНОСТИ, ХАРАКТЕРИЗУЮЩИЕ ПРОЦЕСС ФОРМИРОВАНИЯ
КОМПЕТЕНЦИЙ В ХОДЕ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ**

ШЕСТОЙ СЕМЕСТР

Анализ данных. Экзаменационные материалы

Экзамен проходит письменно, в течение одной пары. Экзаменационный билет содержит два теоретических вопроса и одно практическое задание, состоящее из трех пунктов. Каждая позиция (теоретический вопросы и пункты практического задания) оцениваются в 1 балл. Итоговая оценка — от 0 до 5 баллов.

Теоретические вопросы

1. Перечислите ключевые библиотеки экосистемы Data Science и охарактеризуйте их функциональное назначение.
2. Сформулируйте определение Data Science как процесса. Какова роль выдвижения и проверки гипотез в данном процессе?
3. Какие типы данных могут быть определены в объекте DataFrame? Опишите алгоритм действий при некорректном определении типов.
4. Проведите сравнительный анализ форматов хранения данных CSV, JSON и Parquet. Укажите критерии выбора формата для конкретной задачи.
5. Дайте определение списка как структуры данных. В каких ситуациях применение списков является оптимальным решением?
6. Охарактеризуйте словари: синтаксис создания, методы доступа к элементам, типовые сценарии использования.
7. Какими свойствами обладает структура данных «множество»? Перечислите операции, допустимые над множествами.
8. Объясните синтаксис генераторов списков и словарей. Приведите пример задачи, решаемой с применением генератора.
9. Проведите сравнительный анализ производительности цикла for и векторизованных операций. Чем обусловлено различие в скорости выполнения?
10. Сформулируйте критерии, позволяющие различить список языка Python, массив numpy и серию pandas. В чём состоит принципиальное различие данных объектов?
11. Какие ограничения возникают при использовании цикла for для обработки строк объекта DataFrame? Предложите альтернативный подход.
12. Опишите методы доступа к элементам DataFrame: iloc и loc. В чём заключается различие между данными методами? Приведите примеры использования.
13. Дайте характеристику функции concat: её параметры и типовые сценарии применения при работе с объектами DataFrame.
14. Объясните назначение метода apply. Как изменяется его поведение в зависимости от значения параметра axis? Приведите примеры для axis=0 и axis=1.
15. Опишите механизм группировки данных с использованием метода groupby. Какие типовые операции выполняются над сгруппированными данными?

16. Проведите сравнительный анализ методов `agg` и `transform`. В каких ситуациях применяется каждый из них?
17. Сформулируйте различие между исследовательским и отчётным подходами к визуализации. Какие инструменты соответствуют каждому из подходов?
18. Опишите интерфейс `pyplot` библиотеки `matplotlib`. Какие функции используются для построения базовых графиков?
19. Дайте характеристику объектно-ориентированного интерфейса `matplotlib`. Что представляют собой фигура и оси в контексте данного подхода?
20. Какова роль библиотеки `seaborn` в экосистеме визуализации? Опишите её взаимосвязь с `pandas` и `matplotlib`.
21. Каково назначение параметров `context` и `style` в библиотеке `seaborn`? Приведите примеры их использования.
22. Перечислите фазы разведочного анализа данных (EDA). Какие методы применяются на каждой из фаз?
23. Какие подходы к обработке пропусков в данных вы знаете? В каких случаях целесообразно удаление пропусков, а в каких — заполнение?
24. Сформулируйте определение подтверждающего анализа данных (CDA). Что такое уровень значимости и `p-value`? Как интерпретируются результаты статистического теста?
25. Опишите структуру комплексного исследования, объединяющего EDA и CDA. Приведите пример исследовательского вопроса и укажите, какие методы EDA и CDA будут применены для ответа на него.
26. Дайте определение стохастического прогнозирования. Чем прогнозирующая модель отличается от аппроксимирующей?
27. Охарактеризуйте процесс обучения модели. Каково назначение функции потерь в данном процессе?
28. Поясните принцип работы линейной регрессии как прогнозирующей модели. В чём заключается её интерпретируемость?
29. Опишите алгоритм работы классификатора на основе метода `k` ближайших соседей (`kNN`). Каковы ключевые параметры данной модели?
30. Сформулируйте определение переобучения модели. Перечислите признаки, указывающие на наличие переобучения.
31. Назовите метрики качества регрессионных моделей: `MAE`, `MSE`, `R2`. Дайте их интерпретацию.
32. Обоснуйте необходимость разделения исходных данных на обучающую и тестовую выборки. На какой выборке вычисляются метрики качества и почему?
33. Опишите структуру матрицы ошибок (`confusion matrix`) для бинарной классификации. Какие элементы она содержит и как интерпретируются?
34. Перечислите метрики бинарной классификации: `accuracy`, `precision`, `recall`, `F1-мера`. Укажите, в каких ситуациях каждая из них является предпочтительной.
35. В чём заключаются особенности оценки качества моделей для задач небинарной (многоклассовой) классификации?
36. Обоснуйте необходимость приведения признаков к единому масштабу. Какие методы нормализации вам известны?
37. Поясните различие между нормализацией по минимаксу (`MinMaxScaler`), стандартизацией (`StandardScaler`) и робастной нормализацией (`RobustScaler`). В каких случаях применяется

каждый метод?

38. Как влияет нормализация на линейную регрессию? Обоснуйте отсутствие эффекта.
39. Как влияет нормализация на метод ближайших соседей (kNN)? Обоснуйте чувствительность данной модели к масштабу признаков.
40. Перечислите методы кодирования категориальных данных: Label Encoding, One-Hot Encoding, Target Encoding. Охарактеризуйте область применения каждого метода.
41. Сформулируйте понятие регуляризации. Какова её роль в борьбе с переобучением модели?
42. В чём заключается принципиальное различие между L1-регуляризацией (Lasso) и L2-регуляризацией (Ridge)? Охарактеризуйте механизм обнуления коэффициентов в Lasso.
43. Что представляет собой ElasticNet? В каких ситуациях комбинация L1 и L2 регуляризации является оптимальной?
44. Поясните принцип взвешивания расстояний в методе k ближайших соседей. Как выбор весовой функции влияет на прогноз?
45. Опишите принцип работы логистической регрессии как модели бинарной классификации. Какую роль выполняет сигмоидная функция?
46. Дайте определение функции потерь для логистической регрессии. Что такое логистическая потеря (кросс-энтропия)?
47. Охарактеризуйте принцип построения дерева принятия решений (DecisionTreeClassifier). Какие критерии расщепления используются?
48. Проведите сравнительный анализ ансамблевых методов классификации: случайного леса (RandomForestClassifier) и градиентного бустинга (GradientBoostingClassifier). Укажите преимущества и недостатки каждого.
49. Что такое гиперпараметры модели? Чем гиперпараметры отличаются от параметров модели, обучаемых в процессе тренировки?
50. Опишите метод k-fold кросс-валидации. Как выбор количества блоков (k) влияет на оценку качества модели? Что такое стратифицированная кросс-валидация и в каких случаях она применяется?

Примерный вариант практического задания

Прочитайте файл `bistro_bull.csv`.

Контекст

Сеть городских кафе быстрого питания «BistroBull» хочет понимать, почему одни заведения обладают высокой маржинальностью (больше 5%), а другие — низкий (меньше 5%).

Датасет

- `rating` — рейтинг заведения от одной до пяти звезд (object),
- `cuisine_type` — тип кухни заведения (4 значения: «русская», «итальянская», «азиатская», «фастфуд») (object),
- `restaurant_name` — название бренда (50-100 уникальных значений: «У Егора», «Пирожок на углу», «БургерХаус» и т.д. У одного бренда может быть несколько заведений) (object),
- `avg_check` — средний чек в заведении в рублях (int).

Таргет

- `high_margin` — маржинальность $> 5\%$ (1) или $< 5\%$ (0) (int).

Объем датасета: 5000 записей.

Задание 1. Визуализируйте распределения среднего чека в зависимости от рейтинга заведения.

Задание 2. Проведите Target Encoding всех категориальных признаков.

Задание 3. Обучите бинарный классификатор на ближайших соседях и вычислите его метрики прогнозирования.



Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования

«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра «Высшая математика»

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №1

по дисциплине «Анализ данных и машинное обучение»

Для направления: 02.03.03 Математическое обеспечение и администрирование информационных систем, профиль «Анализ данных в бизнес-аналитике»

6 семестр

1. Перечислите ключевые библиотеки экосистемы Data Science и охарактеризуйте их функциональное назначение.
2. В чём заключается принципиальное различие между L1-регуляризацией (Lasso) и L2-регуляризацией (Ridge)? Охарактеризуйте механизм обнуления коэффициентов в Lasso.
3. Прочитайте файл `bistro_bull.csv`.

Контекст

Сеть городских кафе быстрого питания «BistroBull» хочет понимать, почему одни заведения обладают высокой маржинальностью (больше 5%), а другие — низкий (меньше 5%).

Датасет

- `rating` — рейтинг заведения от одной до пяти звезд (object),
- `cuisine_type` — тип кухни заведения (4 значения: «русская», «итальянская», «азиатская», «фастфуд») (object),
- `restaurant_name` — название бренда (50-100 уникальных значений: «У Егора», «Пирожок на углу», «БургерХаус» и т.д. У одного бренда может быть несколько заведений) (object),
- `avg_check` — средний чек в заведении в рублях (int).

Таргет

- `high_margin` — маржинальность $> 5\%$ (1) или $< 5\%$ (0) (int).

Объем датасета: 5000 записей.

Задание 1. Визуализируйте распределения среднего чека в зависимости от рейтинга заведения.

Задание 2. Проведите Target Encoding всех категориальных признаков.

Задание 3. Обучите бинарный классификатор на ближайших соседях и вычислите его метрики прогнозирования.

Составитель:

«___» _____ 20__ года

Заведующий кафедрой

«___» _____ 20__ года

СЕДЬМОЙ СЕМЕСТР

Анализ данных. Машинное обучение.

Экзаменационные материалы.

Экзамен проходит письменно, в течение одной пары. Экзаменационный билет содержит два теоретических вопроса и одно практическое задание, состоящее из трех пунктов. Каждая позиция (теоретический вопросы и пункты практического задания) оцениваются в 1 балл. Итоговая оценка — от 0 до 5 баллов.

Теоретические вопросы

1. Сформулируйте задачу обучения без учителя. В чем ее отличие от обучения с учителем?
2. В чем состоит задача кластеризации? Какие виды кластеризации вам известны?
3. Опишите алгоритм k-means. Каков критерий останова итерационного процесса?
4. Какими методами осуществляется выбор числа кластеров k в алгоритме k-means? Поясните суть метода локтя.
5. Перечислите основные ограничения алгоритма k-means. Приведите примеры данных, на которых k-means дает неудовлетворительные результаты.
6. В чем состоит принцип работы алгоритма DBSCAN? Какова роль параметров ϵ и min_samples ?
7. Как в алгоритме DBSCAN идентифицируются шумовые точки?
8. Опишите процедуру агломеративной иерархической кластеризации. Что представляет собой дендрограмма и как она используется для выбора числа кластеров?
9. Проведите сравнительный анализ методов k-means, DBSCAN и иерархической кластеризации. В каких случаях предпочтителен каждый из них?
10. Перечислите внутренние метрики качества кластеризации. В чем их отличие от внешних метрик?
11. Поясните смысл силуэтного коэффициента. В каком диапазоне изменяется его значение и как оно интерпретируется?
12. Какие внешние метрики используются для оценки качества кластеризации при наличии истинных меток? Охарактеризуйте Rand Index и Adjusted Rand Index.
13. В чем заключается проблема «проклятия размерности» применительно к задаче кластеризации? Какое следствие она имеет для метрик расстояния?

14. Сформулируйте теорему Эккарта-Янга. Какое практическое значение она имеет для усеченного SVD?
15. Опишите геометрическую интерпретацию сингулярного разложения матрицы. Что представляют собой сингулярные числа?
16. Какова связь между PCA и SVD? Опишите алгоритм получения главных компонент через SVD.
17. Что такое доля объясненной дисперсии (explained variance ratio)? Как она используется для выбора числа главных компонент?
18. Каким образом осуществляется интерпретация главных компонент через loadings? Приведите пример.
19. С какой целью выполняется кластеризация в пространстве главных компонент?
20. В чем состоит принципиальное отличие t-SNE и UMAP от PCA? Для решения каких задач они применяются?
21. Сформулируйте задачу рекомендательной системы. Какие существуют основные типы рекомендательных систем?
22. Опишите метод коллаборативной фильтрации на основе SVD. Как с помощью SVD восстанавливаются пропущенные значения в матрице рейтингов?
23. В чем заключается проблема холодного старта в рекомендательных системах? Какие подходы используются для ее решения?
24. Какие этические проблемы возникают при использовании рекомендательных систем? Охарактеризуйте понятие «пузырь фильтров».
25. Что такое fairness в контексте рекомендательных систем? Приведите пример, когда модель может демонстрировать несправедливое поведение по отношению к определенным группам пользователей.
26. Какие основные компоненты входят в архитектуру пайплайна данных с ML-модулем?
27. В чем различие между batch-режимом, real-time и streaming при интеграции ML-модели в информационную систему?
28. Какие типы соединений реализуются с помощью оператора JOIN в SQL?
29. Для чего используются подзапросы и CTE в SQL-запросах? Приведите пример.
30. Как выполняется формирование датасета из реляционной базы данных для последующего обучения модели?
31. С помощью каких библиотек Python осуществляется подключение к PostgreSQL?
32. В чем различие между библиотеками psycorg2 и SQLAlchemy при работе с базами данных?
33. Каким образом результаты SQL-запроса преобразуются в структуру pandas DataFrame?
34. Какие основные шаги необходимы для установки и настройки сервера PostgreSQL?
35. Какие преимущества предоставляет использование Docker-контейнера для развертывания базы данных?
36. Какие таблицы должны присутствовать в реляционной схеме для хранения данных, признаков и предсказаний?
37. Что такое миграции схемы базы данных и для чего они применяются?
38. Какие HTTP-методы используются в архитектурном стиле REST?
39. Каким образом осуществляется передача признаков модели через POST-запрос к эндпоинту?
40. Для чего применяется валидация входных данных с помощью Pydantic в FastAPI?
41. В чем различие между GET и POST методами при разработке API для ML-модели?

42. Что такое Docker-образ и Docker-контейнер? В чем их различие?
43. Какие компоненты описываются в Dockerfile для сервиса на FastAPI?
44. Для чего предназначен Docker Compose при развертывании multi-container приложения?
45. Каким образом обеспечивается сетевое взаимодействие между контейнерами в Docker Compose?
46. Из каких этапов состоит пайплайн «извлечение — предсказание — запись»?
47. Каким образом осуществляется логирование этапов пайплайна и обработка ошибок?
48. Какие внешние системы могут считывать предсказания из базы данных для принятия решений?
49. Что такое дрейф данных и каким образом осуществляется мониторинг качества предсказаний?
50. Как распределяется ответственность между ML-системой и человеком при принятии окончательного решения?

Примерный вариант практического задания

Исходные данные:

В системе PostgreSQL развернута база данных ml_lab, содержащая схему public со следующими таблицами:

Таблица users — содержит данные пользователей:

- user_id (SERIAL PRIMARY KEY)
- age (INTEGER)
- gender (VARCHAR(10))
- city (VARCHAR(50))
- registration_date (DATE)
- total_purchases (INTEGER)
- avg_check (NUMERIC(10,2))

Таблица transactions — содержит историю транзакций:

- transaction_id (SERIAL PRIMARY KEY)
- user_id (INTEGER REFERENCES users(user_id))
- amount (NUMERIC(10,2))
- category (VARCHAR(50))
- timestamp (TIMESTAMP)

Таблица predictions — целевая таблица для записи результатов:

- prediction_id (SERIAL PRIMARY KEY)
- user_id (INTEGER)
- predicted_churn_probability (NUMERIC(5,4))
- prediction_date (TIMESTAMP DEFAULT NOW())

Задание 1. Напишите SQL-запрос, который для каждого пользователя вычисляет:

- общее количество транзакций,
- среднюю сумму транзакции,
- сумму транзакций за последние 30 дней.

Результат отсортируйте по убыванию суммы транзакций. Запрос должен корректно обрабатывать пользователей, у которых нет транзакций (выводить 0 или NULL).

Задание 2. Используя библиотеку `psycopg2`, напишите Python-скрипт (фрагмент кода), который:

1. подключается к базе данных `ml_lab` (хост: `localhost`, порт: `5432`, пользователь: `student`, пароль: `student_pass`),
2. выполняет SQL-запрос, полученный в вопросе 1,
3. преобразует результат запроса в `pandas DataFrame`,
4. сохраняет `DataFrame` в файл `features.csv`.

Укажите, как обрабатывается ошибка подключения к базе данных.

Задание 3. Опишите структуру POST-эндпоинта `/predict`, разработанного на FastAPI. Эндпоинт должен:

1. принимать на вход JSON-объект с полями: `user_id` (int), `features` (list of float),
2. выполнять валидацию входных данных с использованием `Pydantic`,
3. возвращать JSON-объект с полем `churn_probability` (float).

Приведите пример тела запроса и тела ответа.



Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования

«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра «Высшая математика»

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №1

по дисциплине «Анализ данных и машинное обучение»

Для направления: 02.03.03 Математическое обеспечение и администрирование информационных систем, профиль «Анализ данных в бизнес-аналитике»

7 семестр

1. Сформулируйте задачу обучения без учителя. В чем ее отличие от обучения с учителем?
2. Что такое Docker-образ и Docker-контейнер? В чем их различие?
3. Практическое задание
Исходные данные:

В системе PostgreSQL развернута база данных ml_lab, содержащая схему public со следующими таблицами:

Таблица users — содержит данные пользователей:

- user_id (SERIAL PRIMARY KEY)
- age (INTEGER)
- gender (VARCHAR(10))
- city (VARCHAR(50))
- registration_date (DATE)
- total_purchases (INTEGER)
- avg_check (NUMERIC(10,2))

Таблица transactions — содержит историю транзакций:

- transaction_id (SERIAL PRIMARY KEY)
- user_id (INTEGER REFERENCES users(user_id))
- amount (NUMERIC(10,2))
- category (VARCHAR(50))
- timestamp (TIMESTAMP)

Таблица predictions — целевая таблица для записи результатов:

- prediction_id (SERIAL PRIMARY KEY)
- user_id (INTEGER)
- predicted_churn_probability (NUMERIC(5,4))
- prediction_date (TIMESTAMP DEFAULT NOW())

Задание 1. Напишите SQL-запрос, который для каждого пользователя вычисляет:

- общее количество транзакций,
- среднюю сумму транзакции,
- сумму транзакций за последние 30 дней.

Результат отсортируйте по убыванию суммы транзакций. Запрос должен корректно обрабатывать пользователей, у которых нет транзакций (выводить 0 или NULL).

Задание 2. Используя библиотеку psycopg2, напишите Python-скрипт (фрагмент кода), который:

5. подключается к базе данных ml_lab (хост: localhost, порт: 5432, пользователь: student, пароль: student_pass),
6. выполняет SQL-запрос, полученный в вопросе 1,
7. преобразует результат запроса в pandas DataFrame,
8. сохраняет DataFrame в файл features.csv.

Укажите, как обрабатывается ошибка подключения к базе данных.

Задание 3. Опишите структуру POST-эндпоинта /predict, разработанного на FastAPI. Эндпоинт должен:

4. принимать на вход JSON-объект с полями: user_id (int), features (list of float),
5. выполнять валидацию входных данных с использованием Pydantic,
6. возвращать JSON-объект с полем churn_probability (float).

Приведите пример тела запроса и тела ответа.

Составитель:

«___» _____20__года

Заведующий кафедрой

«___» _____20__года

Тестовые задания для промежуточной аттестации
по дисциплине Б1.В.02.07 Анализ данных и машинное обучение

направление подготовки 02.03.03 Математическое обеспечение и
администрирование информационных систем
профиль "Анализ данных в бизнес-аналитике"
год приема на образовательную программу 2026

№ задания	Содержание задания	Ключ к ответу	Наименование темы РПД	Семестр
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач				
1.	Прочитайте текст вопроса и выберите правильный ответ. Какую модель лучше выбрать, если нужно предсказать цену квартиры по площади и количеству комнат? а) Логистическая регрессия; б) Линейная регрессия; в) KMeans; г) Random Forest Classifier.	б		6
2.	Прочитайте текст вопроса и выберите правильный ответ. В каком порядке правильно организовать аналитическую задачу? а) Сбор данных → постановка задачи → анализ → визуализация; б) Постановка задачи → сбор данных → обработка → моделирование → оценка; в) Обработка → сбор данных → моделирование → визуализация; г) Моделирование → сбор данных → оценка → публикация.	б		6
3.	Прочитайте текст вопроса и выберите правильный ответ. Какой метод лучше использовать для прогнозирования спроса на продукты по историческим данным? а) Классификация; б) Регрессия; в) Кластеризация; г) PCA.	б		6
4.	Прочитайте текст вопроса и выберите правильный ответ. Какой метод подходит для выявления скрытых сегментов клиентов? а) Линейная регрессия; б) Кластеризация; в) Логистическая регрессия; г) Decision Tree Classifier.	б		6

5.	Прочитайте текст вопроса и выберите правильный ответ. Какой тип задачи решается при прогнозировании температуры на следующую неделю? а) Классификация; б) Регрессия; в) Кластеризация; г) PCA.	б		7
6.	Прочитайте текст вопроса и выберите правильный ответ. Какой порядок действий при исследовании данных правильный? а) Сбор → анализ → моделирование → интерпретация → визуализация; б) Моделирование → сбор → визуализация → анализ; в) Визуализация → сбор → обработка → моделирование; г) Публикация → сбор → анализ → оценка.	а		7
7.	Прочитайте текст вопроса и выберите правильный ответ. Какой этап следует за выбором модели? а) Сбор данных; б) Оценка модели на тестовой выборке; в) Постановка задачи; г) Визуализация.	б		7
8.	Прочитайте текст вопроса и выберите правильный ответ. Какую модель лучше выбрать для прогнозирования вероятности того, что клиент откажется от подписки? а) Линейная регрессия; б) Логистическая регрессия; в) KMeans; г) PCA.	б		7
9.	Прочитайте текст вопроса и выберите правильный ответ. Если нужно предсказать категорию животного по изображениям, какую задачу решаем?	б		7
10.	Прочитайте текст вопроса и выберите правильный ответ. Какой аспект наиболее важен при объяснении темы «Деревья решений» с точки зрения информатики? а) умение решить задачу данным методом вручную; б) через этот метод логично осваивать библиотеку Matplotlib; в) важно объяснить принцип максимизации прироста информации или уменьшения энтропии; г) всё выше перечисленное.	в		7

Критерии и шкала оценивания результатов изучения дисциплины на промежуточной аттестации

Оценка «неудовлетворительно» ставится, если обучающийся

- не справился с ответом на теоретические вопросы билета, допустил существенные ошибки;
- при ответе обучающегося выявились существенные пробелы в знаниях основных положений материала;
- не смог ответить на дополнительные вопросы, предложенные преподавателем;
- не выполнил практические задания или при их выполнении допустил существенные ошибки.

Оценка «удовлетворительно» ставится, если обучающийся

- демонстрирует знания по предмету, теорию связывает с практикой, другими темами данного курса, других изучаемых предметов;
- правильно (фрагментарно) отвечает на теоретические вопросы;
- выполнил 2 письменных практических задания, допустив при этом арифметические ошибки.

Оценка «хорошо» ставится, если обучающийся

- демонстрирует систематизированные знания по предмету, владеет приемами рассуждения и сопоставляет материал из разных источников, теорию связывает с практикой, другими темами данного курса;
- правильно, аргументировано отвечает на все теоретические вопросы;
- выполнил без ошибок два из трех письменных практических заданий.

Оценка «отлично» ставится, если обучающийся

- демонстрирует глубокие систематизированные знания по предмету, владеет приемами рассуждения и сопоставляет материал из разных источников, теорию связывает с практикой, другими темами данного курса, других изучаемых предметов;
- правильно, аргументировано отвечает на два теоретических вопроса;
- выполнил без ошибок три письменных практических задания.