

Рабочая программа дисциплины

Методы интеллектуального анализа больших данных

<i>Направление подготовки</i>	Информационные системы и технологии
<i>Код</i>	09.03.02
<i>Направленность (профиль)</i>	Информационные технологии и большие данные
<i>Квалификация (степень) выпускника</i>	бакалавр

1. Перечень кодов компетенций, формируемых дисциплиной в процессе освоения образовательной программы

Группа компетенций	Категория компетенций	Код
Профессиональные	-	ПК-1
Профессиональные	-	ПК-2

2. Компетенции и индикаторы их достижения

Код компетенции	Формулировка компетенции	Индикаторы достижения компетенции
ПК-1	ПК-1 Способен использовать математический аппарат, методологию программирования и современные компьютерные технологии для решения практических задач получения, хранения, обработки и передачи информации.	ПК-1.1. Подбирает парадигму программирования под решение конкретной прикладной задачи; ПК-1.3. Комбинирует известные алгоритмы решения задач. ПК-1.5. Применяет знания теоретической информатики, фундаментальной и прикладной математики для анализа и синтеза информационных систем и процессов.
ПК-2	ПК-2. Способен использовать математический аппарат и современные компьютерные средства для выполнения научно-исследовательских работ по закрепленной тематике	ПК-2.1. Демонстрирует способность и готовность к построению и исследованию математических моделей различных физических, биологических, экономических и социальных систем, а также применению идей, принципов и методов математического моделирования при решении прикладных задач. ПК-2.2. Использует естественнонаучные и математические знания для ориентирования в современном информационном пространстве ПК-2.3. Способен использовать естественнонаучные и математические знания для ориентирования в современном информационном пространстве. ПК-2.4. Собирает и анализирует информацию по решаемой задаче, составляет ее математическое описание, обеспечивает накопление, анализ и систематизацию собранных данных с использованием современных достижений науки и информационных систем; ПК-2.5. Выявляет и формулирует актуальные научные проблемы; обосновывать актуальность,

		теоретическую и практическую значимость темы научного исследования, разрабатывать план и программу проведения научного исследования; ПК-2.6. Проводит научно-исследовательские работы в области математики и компьютерных наук.
--	--	---

3. Описание планируемых результатов обучения по дисциплине

3.1. Описание планируемых результатов обучения по дисциплине

Планируемые результаты обучения по дисциплине представлены дескрипторами (знания, умения, навыки).

Дескрипторы по дисциплине	Знать	Уметь	Владеть
Код компетенции	ПК-1		
	- математический аппарат, методологию программирования и современные компьютерные технологии для решения практических задач получения, хранения, обработки и передачи информации; - основные идеи, понятия и методы, определяющие стиль написания, отладки и сопровождения программ; - характеристики основных парадигм программирования;	- использовать математический аппарат, методологию программирования и современные компьютерные технологии для решения практических задач получения, хранения, обработки и передачи информации; - применять современные компьютерные технологии для решения практических задач; - делать обоснованный выбор инструментария для решения прикладных задач;	- навыками использования математического аппарата, методологии программирования и современных компьютерных технологий для решения практических задач получения, хранения, обработки и передачи информации; - математическим аппаратом для построения вычислительных моделей практических задач; - навыками использования стандартных алгоритмических моделей для решения задач хранения и обработки информации
Код компетенции	ПК-2		

	Основы операционных систем: Знать архитектуру и основные компоненты операционных систем (например, Windows, Linux). Знать принципы работы файловых систем и управления процессами. Сетевые протоколы: Знать основные сетевые протоколы (TCP/IP, HTTP, HTTPS, FTP) и их функции. Знать принципы работы сетевых устройств (маршрутизаторов, коммутаторов, брандмауэров). Инструменты конфигурирования: Знать инструменты и утилиты для конфигурирования операционных систем и сетевых устройств (например, командная строка, PowerShell, SSH). Безопасность: Знать основные принципы и методы обеспечения безопасности операционных систем и сетевых устройств.	Конфигурировать операционные системы: Уметь устанавливать и настраивать операционные системы для веб-разработки. Уметь управлять пользователями и правами доступа в операционных системах. Настраивать сетевые устройства: Уметь конфигурировать маршрутизаторы и коммутаторы для обеспечения сетевой связности. Уметь настраивать брандмауэры для защиты веб-приложений. Использовать инструменты командной строки: Уметь использовать командную строку для выполнения задач администрирования и конфигурирования. Уметь применять скрипты для автоматизации процессов конфигурирования. Обеспечивать безопасность систем: Уметь применять методы шифрования и аутентификации для защиты данных. Уметь проводить аудит безопасности и выявлять уязвимости в системах.	Практическими навыками конфигурирования: Владеть навыками установки и настройки операционных систем в средах веб-разработки. Владеть навыками конфигурирования сетевых устройств для оптимизации работы веб-приложений. Работой с документацией: Владеть умением читать и интерпретировать техническую документацию по операционным системам и сетевым устройствам. Решением проблем: Владеть навыками диагностики и устранения неполадок в операционных системах и сетевых устройствах. Работой в команде: Владеть навыками взаимодействия с другими специалистами (разработчиками, системными администраторами и) для
--	--	---	--

			эффективного конфигурирования и поддержки веб-приложений.
--	--	--	---

4. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина «Методы интеллектуального анализа больших данных» относится к части, формируемой участниками образовательных отношений учебного плана ОПОП.

Данная дисциплина взаимосвязана с другими дисциплинами, такими как: «Математический анализ», «Линейная алгебра», «Алгоритмизация и методы программирования», «Математическая логика и дискретная математика», «Численные методы», «Системы искусственного интеллекта», «Базы данных», «Машинное обучение», «Технологии хранения и обработки больших данных».

В рамках освоения программы бакалавриата выпускники готовятся к решению задач профессиональной деятельности следующих типов:

Профиль (направленность) программы установлена путем ее ориентации на сферу профессиональной деятельности выпускников: Информационные технологии и большие данные

5. Объем дисциплины

<i>Виды учебной работы</i>	<i>Формы обучения</i>
	<i>Очно-заочное</i>
Общая трудоемкость: зачетные единицы/часы	6/216
Контактная работа:	
Занятия лекционного типа	12
Занятия семинарского типа	12
Промежуточная аттестация: экзамен, зачет	0,15
Самостоятельная работа (СРС)	191,85

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

6.1. Распределение часов по разделам/темам и видам работы

6.1.1. Очно-заочная форма обучения

№ п/п	Раздел/тема	Виды учебной работы (в часах)						
		Контактная работа						Самосто ятельная работа
		Занятия лекционного типа		Занятия семинарского типа				
		Лекции	Иные учебные занятия	Практи ческие занятия	Семи нары	Лабора торные работы	Иные	
1.	Введение в интеллектуальный анализ больших	2			2			38

	данных. Жизненный цикл аналитического проекта							
2.	Предобработка и разведочный анализ данных	2			2			38
3.	Классификация и регрессия. Контролируемое обучение	2			2			38
4.	Кластеризация и снижение размерности. Не контролируемое обучение	3			3			38
5.	Поиск ассоциативных правил и работа с большими объемами данных	3			3			39,85
	Промежуточная аттестация	0,15						
	Итого	12			12			191,85

6.2 Программа дисциплины, структурированная по темам / разделам

6.2.1 Содержание лекционного курса

№ п/п	Наименование раздела дисциплины	Содержание раздела дисциплины
1.	Введение в интеллектуальный анализ больших данных. Жизненный цикл аналитического проекта	- Определение Big Data и Data Mining. «Четыре V» больших данных: Volume (объем), Velocity (скорость), Variety (разнообразие), Veracity (достоверность). Дополнительные характеристики: Value (ценность), Variability (изменчивость). - Понятие CRISP-DM (Cross-Industry Standard Process for Data Mining) – кросс-отраслевой стандарт процессов анализа данных: понимание бизнес-целей и требований к проекту; сбор, описание, разведка и проверка качества данных; очистка, преобразование, создание признаков. - выбор и применение методов моделирования. - оценка результатов на соответствие бизнес-целям. - внедрение, мониторинг и сопровождение модели.
2.	Предобработка и разведочный анализ данных	-Основные проблемы качества данных: пропущенные значения, выбросы (выявление методами z-оценки, межквартального размаха, визуальными методами). Стратегии обработки (удаление, трансформация).

		- Дубликаты: поиск и устранение полных и частичных дубликатов. - Шум и некорректные форматы данных. - Нормализация и стандартизация данных: приведение к диапазону [0,1], центрирование и масштабирование к единичной дисперсии. - Кодирование категориальных признаков: Label Encoding (порядковые категории), One-Hot Encoding (номинальные категории), Target Encoding. - Визуализация в EDA: гистограммы, ящики с усами (boxplot), диаграммы рассеяния, тепловые карты корреляций. - Балансировка классов для задач классификации (undersampling, oversampling, SMOTE).
3.	Классификация и регрессия. Контролируемое обучение	- Общая схема контролируемого обучения: обучающая выборка, целевая переменная, признаки, функция потерь, обучение, предсказание. - Метрики качества: • Для классификации: точность, полнота, F1-мера, ROC-кривая и AUC, матрица ошибок. • Для регрессии: средняя абсолютная ошибка, среднеквадратичная ошибка, коэффициент детерминации. • Методы классификации: логистическая регрессия (бинарная и многоклассовая), метод k-ближайших соседей, деревья решений. - Методы регрессии: линейная регрессия: от МНК до регуляризации, полиномиальная регрессия. - Объединенные методы: Бэггинг (усреднение множества деревьев), Бустинг (XGBoost, LightGBM, CatBoost). - Стекинг. - Понятие переобучения (overfitting) и недообучения (underfitting). Методы борьбы: кросс-валидация, регуляризация, early stopping.
4.	Кластеризация и снижение размерности. Не контролируемое обучение	- Задачи кластеризации: • Сегментация пользователей, группировка документов, обнаружение аномалий. • Метрики расстояния между объектами: • Евклидово расстояние, манхэттенское расстояние, косинусное сходство, расстояние Минковского. - Методы кластеризации: • K-means: принцип работы, выбор числа кластеров (метод локтя, силуэтный коэффициент), чувствительность к начальным центрам (K-means++), ограничения. • Иерархическая кластеризация: агломеративная (снизу вверх) и дивизивная (сверху вниз). Дендрограмма. Методы связывания кластеров.

		• DBSCAN (Density-Based Spatial Clustering of Applications with Noise): параметры eps и minPts, поиск кластеров произвольной формы, устойчивость к выбросам. - Оценка качества кластеризации: • Внутренние метрики (силуэт, индекс Калински-Харабаса, индекс Дэвиса-Болдина). • Внешние метрики (Rand index, adjusted mutual information – если известна истинная разметка). • Снижение размерности (Dimensionality Reduction): • Проблема «проклятия размерности». • PCA (Principal Component Analysis): поиск главных компонент, объясненная дисперсия, выбор числа компонент. • t-SNE (t-distributed Stochastic Neighbor Embedding): нелинейное снижение размерности для визуализации высокоразмерных данных. • Autoencoders (краткое введение как нейросетевой метод снижения размерности). • Отбор признаков (feature selection): фильтрующие (корреляция), оберточные (RFE) и встроенные (Lasso regularization) методы.
5.	Поиск ассоциативных правил и работа с большими объемами данных	- Поиск ассоциативных правил (Association Rule Mining): - Задача: поиск часто встречающихся наборов элементов (itemsets) в транзакционных данных. - Ключевые метрики: Support (поддержка), Confidence (достоверность), Lift (лифт). - Алгоритм Apriori : принцип Apriori (все подмножества частого набора также частоты), генерация кандидатов и проверка support. Преимущества и недостатки (многократные проходы по данным). - Алгоритм FP-Growth (Frequent Pattern Growth): построение дерева FP-Tree, двухпроходный алгоритм без генерации кандидатов. Эффективность для больших данных. - Применение: маркетинговые корзины (market basket analysis), кросс-продажи, рекомендательные системы. - Масштабирование алгоритмов для больших данных: - Парадигма MapReduce (Hadoop): map-функция (фильтрация/преобразование), shuffle/sort, reduce-функция (агрегация/свертка). - Пример реализации word count на MapReduce. - Ограничения MapReduce: дисковые операции, сложность итеративных алгоритмов. - Apache Spark как промышленный стандарт:

		- Базовые концепты: драйвер, исполнители (executors), RDD (Resilient Distributed Dataset), DataFrame, Dataset. - Трансформации (map, filter, flatmap, groupBy, join) и действия (count, collect, reduce, save). - Ленивые вычисления (lazy evaluation) и оптимизации Catalyst, Tungsten. - MLlib – библиотека машинного обучения Spark: - Возможности: классификация (логистическая регрессия, SVM, деревья), регрессия, кластеризация (K-means), снижение размерности (PCA), поиск частых наборов (FP-Growth). - Примеры архитектур big data: - Хранилище данных (HDFS, S3, Data Lake). - Обработка (Spark, Flink, Hive). - Визуализация и BI (Tableau, Power BI, Superset).
--	--	---

6.2.2 Содержание практических занятий

№ п/п	Наименование раздела дисциплины	Содержание практических занятий
1.	Введение в интеллектуальный анализ больших данных. Жизненный цикл аналитического проекта	- Определение Big Data и Data Mining. «Четыре V» больших данных: Volume (объем), Velocity (скорость), Variety (разнообразие), Veracity (достоверность). Дополнительные характеристики: Value (ценность), Variability (изменчивость). - Понятие CRISP-DM (Cross-Industry Standard Process for Data Mining) – кросс-отраслевой стандарт процессов анализа данных: понимание бизнес-целей и требований к проекту; сбор, описание, разведка и проверка качества данных; очистка, преобразование, создание признаков. - выбор и применение методов моделирования. - оценка результатов на соответствие бизнес-целям. - внедрение, мониторинг и сопровождение модели.
2.	Предобработка и разведочный анализ данных	- Основные проблемы качества данных: пропущенные значения, выбросы (выявление методами z-оценки, межквартального размаха, визуальными методами). Стратегии обработки (удаление, трансформация). - Дубликаты: поиск и устранение полных и частичных дубликатов. - Шум и некорректные форматы данных. - Нормализация и стандартизация данных: приведение к диапазону [0,1], центрирование и масштабирование к единичной дисперсии. - Кодирование категориальных признаков: Label Encoding (порядковые категории), One-Hot Encoding (номинальные категории), Target Encoding. - Визуализация в EDA: гистограммы, ящики с усами (boxplot), диаграммы рассеяния, тепловые карты

		корреляций. - Балансировка классов для задач классификации (undersampling, oversampling, SMOTE).
3.	Классификация и регрессия. Контролируемое обучение	- Общая схема контролируемого обучения: обучающая выборка, целевая переменная, признаки, функция потерь, обучение, предсказание. - Метрики качества: • Для классификации: точность, полнота, F1-мера, ROC-кривая и AUC, матрица ошибок. • Для регрессии: средняя абсолютная ошибка, среднеквадратичная ошибка, коэффициент детерминации. • Методы классификации: логистическая регрессия (бинарная и многоклассовая), метод k-ближайших соседей, деревья решений. - Методы регрессии: линейная регрессия: от МНК до регуляризации, полиномиальная регрессия. - Объединенные методы: Бэггинг (усреднение множества деревьев), Бустинг (XGBoost, LightGBM, CatBoost). - Стекинг. - Понятие переобучения (overfitting) и недообучения (underfitting). Методы борьбы: кросс-валидация, регуляризация, early stopping.
4.	Кластеризация и снижение размерности. Не контролируемое обучение	- Задачи кластеризации: • Сегментация пользователей, группировка документов, обнаружение аномалий. • Метрики расстояния между объектами: • Евклидово расстояние, манхэттенское расстояние, косинусное сходство, расстояние Минковского. - Методы кластеризации: • K-means: принцип работы, выбор числа кластеров (метод локтя, силуэтный коэффициент), чувствительность к начальным центрам (K-means++), ограничения. • Иерархическая кластеризация: агломеративная (снизу вверх) и дивизивная (сверху вниз). Дендрограмма. Методы связывания кластеров. • DBSCAN (Density-Based Spatial Clustering of Applications with Noise): параметры eps и minPts, поиск кластеров произвольной формы, устойчивость к выбросам. - Оценка качества кластеризации: • Внутренние метрики (силуэт, индекс Калински-Харабаса, индекс Дэвиса-Болдина). • Внешние метрики (Rand index, adjusted mutual information – если известна истинная разметка). • Снижение размерности (Dimensionality Reduction): • Проблема «проклятия размерности».

		• PCA (Principal Component Analysis): поиск главных компонент, объясненная дисперсия, выбор числа компонент. • t-SNE (t-distributed Stochastic Neighbor Embedding): нелинейное снижение размерности для визуализации высокоразмерных данных. • Autoencoders (краткое введение как нейросетевой метод снижения размерности). • Отбор признаков (feature selection): фильтрующие (корреляция), оберточные (RFE) и встроенные (Lasso regularization) методы.
5.	Поиск ассоциативных правил и работа с большими объемами данных	- Поиск ассоциативных правил (Association Rule Mining): - Задача: поиск часто встречающихся наборов элементов (itemsets) в транзакционных данных. - Ключевые метрики: Support (поддержка), Confidence (достоверность), Lift (лифт). - Алгоритм Apriori : принцип Apriori (все подмножества частого набора также частоты), генерация кандидатов и проверка support. Преимущества и недостатки (многократные проходы по данным). - Алгоритм FP-Growth (Frequent Pattern Growth): построение дерева FP-Tree, двухпроходный алгоритм без генерации кандидатов. Эффективность для больших данных. - Применение: маркетинговые корзины (market basket analysis), кросс-продажи, рекомендательные системы. - Масштабирование алгоритмов для больших данных: - Парадигма MapReduce (Hadoop): map-функция (фильтрация/преобразование), shuffle/sort, reduce-функция (агрегация/свертка). - Пример реализации word count на MapReduce. - Ограничения MapReduce: дисковые операции, сложность итеративных алгоритмов. - Apache Spark как промышленный стандарт: - Базовые концепты: драйвер, исполнители (executors), RDD (Resilient Distributed Dataset), DataFrame, Dataset. - Трансформации (map, filter, flatmap, groupBy, join) и действия (count, collect, reduce, save). - Ленивые вычисления (lazy evaluation) и оптимизации Catalyst, Tungsten. - MLlib – библиотека машинного обучения Spark: - Возможности: классификация (логистическая регрессия, SVM, деревья), регрессия, кластеризация (K-means), снижение размерности (PCA), поиск частых наборов (FP-Growth).

		- Примеры архитектур big data: - Хранилище данных (HDFS, S3, Data Lake). - Обработка (Spark, Flink, Hive). - Визуализация и BI (Tableau, Power BI, Superset).
--	--	--

6.2.3 Содержание самостоятельной работы

№ п/п	Наименование раздела дисциплины	Содержание самостоятельной работы
1.	Введение в интеллектуальный анализ больших данных. Жизненный цикл аналитического проекта	- Определение Big Data и Data Mining. «Четыре V» больших данных: Volume (объем), Velocity (скорость), Variety (разнообразие), Veracity (достоверность). Дополнительные характеристики: Value (ценность), Variability (изменчивость). - Понятие CRISP-DM (Cross-Industry Standard Process for Data Mining) – кросс-отраслевой стандарт процессов анализа данных: понимание бизнес-целей и требований к проекту; сбор, описание, разведка и проверка качества данных; очистка, преобразование, создание признаков. - выбор и применение методов моделирования. - оценка результатов на соответствие бизнес-целям. - внедрение, мониторинг и сопровождение модели.
2.	Предобработка и разведочный анализ данных	- Основные проблемы качества данных: пропущенные значения, выбросы (выявление методами z-оценки, межквартильного размаха, визуальными методами). Стратегии обработки (удаление, трансформация). - Дубликаты: поиск и устранение полных и частичных дубликатов. - Шум и некорректные форматы данных. - Нормализация и стандартизация данных: приведение к диапазону [0,1], центрирование и масштабирование к единичной дисперсии. - Кодирование категориальных признаков: Label Encoding (порядковые категории), One-Hot Encoding (номинальные категории), Target Encoding. - Визуализация в EDA: гистограммы, ящики с усами (boxplot), диаграммы рассеяния, тепловые карты корреляций. - Балансировка классов для задач классификации (undersampling, oversampling, SMOTE).
3.	Классификация и регрессия. Контролируемое обучение	- Общая схема контролируемого обучения: обучающая выборка, целевая переменная, признаки, функция потерь, обучение, предсказание. - Метрики качества: • Для классификации: точность, полнота, F1-мера, ROC-кривая и AUC, матрица ошибок. • Для регрессии: средняя абсолютная ошибка, среднеквадратичная ошибка, коэффициент

		детерминации. • Методы классификации: логистическая регрессия (бинарная и многоклассовая), метод k-ближайших соседей, деревья решений. - Методы регрессии: линейная регрессия: от МНК до регуляризации, полиномиальная регрессия. - Объединенные методы: Бэггинг (усреднение множества деревьев), Бустинг (XGBoost, LightGBM, CatBoost). - Стекинг. - Понятие переобучения (overfitting) и недообучения (underfitting). Методы борьбы: кросс-валидация, регуляризация, early stopping.
4.	Кластеризация и снижение размерности. Не контролируемое обучение	- Задачи кластеризации: • Сегментация пользователей, группировка документов, обнаружение аномалий. • Метрики расстояния между объектами: • Евклидово расстояние, манхэттенское расстояние, косинусное сходство, расстояние Минковского. - Методы кластеризации: • K-means: принцип работы, выбор числа кластеров (метод локтя, силуэтный коэффициент), чувствительность к начальным центрам (K-means++), ограничения. • Иерархическая кластеризация: агломеративная (снизу вверх) и дивизивная (сверху вниз). Дендрограмма. Методы связывания кластеров. • DBSCAN (Density-Based Spatial Clustering of Applications with Noise): параметры eps и minPts, поиск кластеров произвольной формы, устойчивость к выбросам. - Оценка качества кластеризации: • Внутренние метрики (силуэт, индекс Калински-Харабаса, индекс Дэвиса-Болдина). • Внешние метрики (Rand index, adjusted mutual information – если известна истинная разметка). • Снижение размерности (Dimensionality Reduction): • Проблема «проклятия размерности». • PCA (Principal Component Analysis): поиск главных компонент, объясненная дисперсия, выбор числа компонент. • t-SNE (t-distributed Stochastic Neighbor Embedding): нелинейное снижение размерности для визуализации высокоразмерных данных. • Autoencoders (краткое введение как нейросетевой метод снижения размерности). • Отбор признаков (feature selection): фильтрующие (корреляция), оберточные (RFE)

		и встроенные (Lasso regularization) методы.
5.	Поиск ассоциативных правил и работа с большими объемами данных	- Поиск ассоциативных правил (Association Rule Mining): - Задача: поиск часто встречающихся наборов элементов (itemsets) в транзакционных данных. - Ключевые метрики: Support (поддержка), Confidence (достоверность), Lift (лифт). - Алгоритм Apriori : принцип Apriori (все подмножества частого набора также частоты), генерация кандидатов и проверка support. Преимущества и недостатки (многократные проходы по данным). - Алгоритм FP-Growth (Frequent Pattern Growth): построение дерева FP-Tree, двухпроходный алгоритм без генерации кандидатов. Эффективность для больших данных. - Применение: маркетинговые корзины (market basket analysis), кросс-продажи, рекомендательные системы. - Масштабирование алгоритмов для больших данных: - Парадигма MapReduce (Hadoop): map-функция (фильтрация/преобразование), shuffle/sort, reduce-функция (агрегация/свертка). - Пример реализации word count на MapReduce. - Ограничения MapReduce: дисковые операции, сложность итеративных алгоритмов. - Apache Spark как промышленный стандарт: - Базовые концепты: драйвер, исполнители (executors), RDD (Resilient Distributed Dataset), DataFrame, Dataset. - Трансформации (map, filter, flatmap, groupBy, join) и действия (count, collect, reduce, save). - Ленивые вычисления (lazy evaluation) и оптимизации Catalyst, Tungsten. - MLlib – библиотека машинного обучения Spark: - Возможности: классификация (логистическая регрессия, SVM, деревья), регрессия, кластеризация (K-means), снижение размерности (PCA), поиск частых наборов (FP-Growth). - Примеры архитектур big data: - Хранилище данных (HDFS, S3, Data Lake). - Обработка (Spark, Flink, Hive). - Визуализация и BI (Tableau, Power BI, Superset).

7. Текущий контроль по дисциплине (модулю) в рамках учебных занятий

В рамках текущего контроля преподаватель самостоятельно может проводить следующие мероприятия:

№	Контролируемые разделы (темы)	Наименование оценочного средства
---	-------------------------------	----------------------------------

п/п		
1.	Введение в интеллектуальный анализ больших данных. Жизненный цикл аналитического проекта	Опрос, информационный проект.
2.	Предобработка и разведочный анализ данных	Опрос, информационный проект, тестирование.
3.	Классификация и регрессия. Контролируемое обучение	Опрос, информационный проект.
4.	Кластеризация и снижение размерности. Не контролируемое обучение	Опрос, информационный проект.
5.	Поиск ассоциативных правил и работа с большими объемами данных	Опрос, информационный проект, тестирование.

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

8.1. Основная литература:

1. Методы интеллектуального анализа данных. Задача классификации : учебное пособие / О. С. Чудинова, Р. М. Безбородникова, Е. Н. Корнейченко [и др.]. — Оренбург : Оренбургский государственный университет, ЭБС АСБ, 2024. — 167 с. — ISBN 978-5-7410-3284-8. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/153049.html>
2. Целых, А. Н. Принятие решений на основе методов машинного обучения : учебное пособие по курсам «Модели и методы инженерии знаний», «Методы анализа больших данных» / А. Н. Целых, Н. В. Драгныш, Э. М. Котов. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2022. — 113 с. — ISBN 978-5-9275-4246-8. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/131458.html>
3. Хохлов, С. С. Основные методы анализа экспериментальных данных : учебное пособие / С. С. Хохлов, А. Н. Дмитриева. — Москва : Национальный исследовательский ядерный университет «МИФИ», 2025. — 126 с. — ISBN 978-5-7262-3149-5. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/156787.htm>

8.2. Дополнительная литература:

1. Целых, А. Н. Применение временных рядов для анализа больших данных : учебное пособие по курсу «Математические методы анализа больших данных» / А. Н. Целых, В. С. Васильев, Э. М. Котов. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2021. — 84 с. — ISBN 978-5-9275-3983-3. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/121929.html>
2. Целых, А. Н. Современные методы прикладной информатики в задачах анализа данных : учебное пособие по курсу «Методы интеллектуального анализа данных» / А. Н. Целых, А. А. Целых, Э. М. Котов. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2021. — 130 с. — ISBN 978-5-9275-3783-9. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/117165.html>

8.3. Периодические издания

1. Вестник Московского государственного технического университета имени Н.Э. Баумана. Серия Естественные науки. ISSN 1812-3368. <https://www.iprbookshop.ru/23124.html>
2. Открытые Системы. СУБД. ISSN 1028-7493. <https://www.iprbookshop.ru/76383.html>
3. Информационные технологии моделирования и управления. ISSN 1813-9744. <https://www.iprbookshop.ru/43350.html>

9. Перечень ресурсов информационно-телекоммуникационной сети "Интернет" (далее - сеть "Интернет"), необходимых для освоения дисциплины (модуля)

1. Федеральный портал «Российское образование» <http://www.edu.ru>
2. Электронно-библиотечная система «Научная электронная библиотека eLIBRARY.RU» <https://www.elibrary.ru>
3. Электронно-библиотечная система IPR BOOKS <https://www.iprbookshop.ru>
4. Электронная библиотечная система <https://biblioclub.ru>

10. Методические указания для обучающихся по освоению дисциплины (модуля)

Успешное освоение данного курса базируется на рациональном сочетании нескольких видов учебной деятельности – лекций, семинарских занятий, самостоятельной работы. При этом самостоятельную работу следует рассматривать одним из главных звеньев полноценного высшего образования, на которую отводится значительная часть учебного времени.

Самостоятельная работа студентов складывается из следующих составляющих:

1. Работа с основной и дополнительной литературой, с материалами интернета и конспектами лекций;
2. Внеаудиторная подготовка к контрольным работам, выполнение докладов, рефератов и курсовых работ;
3. Выполнение самостоятельных практических работ;
4. Подготовка к экзаменам (зачетам) непосредственно перед ними.

Для правильной организации работы необходимо учитывать порядок изучения разделов курса, находящихся в строгой логической последовательности. Поэтому хорошее усвоение одной части дисциплины является предпосылкой для успешного перехода к следующей. Задания, проблемные вопросы, предложенные для изучения дисциплины, в том числе и для самостоятельного выполнения, носят междисциплинарный характер и базируются, прежде всего, на причинно-следственных связях между компонентами окружающего нас мира. Важным составляющим в изучении данного курса является решение ситуационных задач и работа над проблемно-аналитическими заданиями, что предполагает знание соответствующей научной терминологии и т.д.

Для лучшего запоминания материала целесообразно использовать индивидуальные особенности и разные виды памяти: зрительную, слуховую, ассоциативную. Успешному запоминанию также способствует приведение ярких свидетельств и наглядных примеров. Учебный материал должен постоянно повторяться и закрепляться.

При выполнении докладов, творческих, информационных, исследовательских проектов особое внимание следует обращать на подбор источников информации и методику работы с ними.

Для успешной сдачи экзамена (зачета) рекомендуется соблюдать следующие правила:

1. Подготовка к экзамену (зачету) должна проводиться систематически, в течение всего семестра.

2. Интенсивная подготовка должна начаться не позднее, чем за месяц до экзамена.

3. Время непосредственно перед экзаменом (зачетом) лучше использовать таким образом, чтобы оставить последний день свободным для повторения курса в целом, для систематизации материала и доработки отдельных вопросов.

На экзамене высокую оценку получают студенты, использующие данные, полученные в процессе выполнения самостоятельных работ, а также использующие собственные выводы на основе изученного материала.

Учитывая значительный объем теоретического материала, студентам рекомендуется регулярное посещение и подробное конспектирование лекций.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)

1. Microsoft Windows Server;

2. Семейство ОС Microsoft Windows;

3. Libre Office свободно распространяемый офисный пакет с открытым исходным кодом;

4. Информационно-справочная система: Система КонсультантПлюс (КонсультантПлюс);

5. Информационно-правовое обеспечение Гарант: Электронный периодический справочник «Система ГАРАНТ» (Система ГАРАНТ);

Перечень используемого программного обеспечения указан в п.12 данной рабочей программы дисциплины.

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

12.1. Учебная аудитория для проведения учебных занятий, предусмотренных образовательной программой, оснащенная оборудованием и техническими средствами обучения.

Специализированная мебель:

Комплект учебной мебели (стол, стул) по количеству обучающихся; комплект мебели для преподавателя; доска (маркерная).

Технические средства обучения:

Компьютер в сборе для преподавателя, проектор, экран, колонки

Перечень лицензионного программного обеспечения, в том числе отечественного производства:

Windows 10, КонсультантПлюс, Kaspersky Endpoint Security.

Перечень свободно распространяемого программного обеспечения:

Яндекс Браузер, LibreOffice, МТС Линк.

Подключение к сети «Интернет» и обеспечение доступа в электронную информационно-образовательную среду ММУ.

12.2. Помещение для самостоятельной работы обучающихся.

Специализированная мебель:

Комплект учебной мебели (стол, стул) по количеству обучающихся; комплект мебели для преподавателя; доска (маркерная).

Технические средства обучения:

Компьютер в сборе для преподавателя; компьютеры в сборе для обучающихся;

колонки; проектор, экран.

Перечень лицензионного программного обеспечения, в том числе отечественного производства:

Windows Server 2016, Windows 10, Microsoft Office, КонсультантПлюс, Kaspersky Endpoint Security.

Перечень свободно распространяемого программного обеспечения:

Яндекс Браузер, LibreOffice, МТС Линк, Gimp, Paint.net, AnyLogic, Inkscape.

Помещение для самостоятельной работы обучающихся оснащено компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду ММУ.

13. Образовательные технологии, используемые при освоении дисциплины

Для освоения дисциплины используются как традиционные формы занятий – лекции (типы лекций – установочная, вводная, текущая, заключительная, обзорная; виды лекций – проблемная, визуальная, лекция конференция, лекция консультация); и семинарские (практические) занятия, так и активные и интерактивные формы занятий – деловые и ролевые игры, решение ситуационных задач и разбор конкретных ситуаций.

На учебных занятиях используются технические средства обучения мультимедийной аудитории: компьютер, монитор, колонки, настенный экран, проектор, микрофон, пакет программ Microsoft Office для демонстрации презентаций и медиафайлов, видеопроектор для демонстрации слайдов, видеосюжетов и др. Тестирование обучаемых может осуществляться с использованием компьютерного оборудования университета.

13.1. В освоении учебной дисциплины используются следующие традиционные образовательные технологии:

- чтение проблемно-информационных лекций с использованием доски и видеоматериалов;
- семинарские занятия для обсуждения, дискуссий и обмена мнениями;
- контрольные опросы;
- консультации;
- самостоятельная работа студентов с учебной литературой и первоисточниками;
- подготовка и обсуждение рефератов (проектов), презентаций (научно-исследовательская работа);
- тестирование по основным темам дисциплины.

13.2. Активные и интерактивные методы и формы обучения

Из перечня видов: («мозговой штурм», анализ НПА, анализ проблемных ситуаций, анализ конкретных ситуаций, инциденты, имитация коллективной профессиональной деятельности, разыгрывание ролей, творческая работа, связанная с освоением дисциплины, ролевая игра, круглый стол, диспут, беседа, дискуссия, мини-конференция и др.) используются следующие:

- диспут;
- анализ проблемных, творческих заданий, ситуационных задач;
- ролевая игра;
- круглый стол;
- мини-конференция;
- дискуссия;
- беседа.

13.3. Особенности обучения инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ)

При организации обучения по дисциплине учитываются особенности организации взаимодействия с инвалидами и лицами с ограниченными возможностями здоровья (далее – инвалиды и лица с ОВЗ) с целью обеспечения их прав. При обучении учитываются особенности их психофизического развития, индивидуальные возможности и при необходимости обеспечивается коррекция нарушений развития и социальная адаптация указанных лиц.

Выбор методов обучения определяется содержанием обучения, уровнем методического и материально-технического обеспечения, особенностями восприятия учебной информации студентов с инвалидностью и студентов с ограниченными возможностями здоровья и т.д. В образовательном процессе используются социально-активные и рефлексивные методы обучения, технологии социокультурной реабилитации с целью оказания помощи в установлении полноценных межличностных отношений с другими студентами, создании комфортного психологического климата в студенческой группе.

При обучении лиц с ограниченными возможностями здоровья электронное обучение и дистанционные образовательные технологии предусматривают возможность приема-передачи информации в доступных для них формах.

Обучающиеся из числа лиц с ограниченными возможностями здоровья обеспечены печатными и электронными образовательными ресурсами в формах, адаптированных к ограничениям их здоровья.

Автономная некоммерческая организация высшего образования
«МОСКОВСКИЙ МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ»

**ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ
ПО ДИСЦИПЛИНЕ**

Методы интеллектуального анализа больших данных

<i>Направление подготовки</i>	Информационные системы и технологии
<i>Код</i>	09.03.02
<i>Направленность (профиль)</i>	Информационные технологии и большие данные
<i>Квалификация (степень) выпускника</i>	бакалавр

1. Перечень кодов компетенций, формируемых дисциплиной в процессе освоения образовательной программы

Группа компетенций	Категория компетенций	Код
Профессиональные	-	ПК-1
Профессиональные	-	ПК-2

2. Компетенции и индикаторы их достижения

Код компетенции	Формулировка компетенции	Индикаторы достижения компетенции
ПК-1	ПК-1 Способен использовать математический аппарат, методологию программирования и современные компьютерные технологии для решения практических задач получения, хранения, обработки и передачи информации.	ПК-1.1. Подбирает парадигму программирования под решение конкретной прикладной задачи; ПК-1.3. Комбинирует известные алгоритмы решения задач. ПК-1.5. Применяет знания теоретической информатики, фундаментальной и прикладной математики для анализа и синтеза информационных систем и процессов.
ПК-2	ПК-2. Способен использовать математический аппарат и современные компьютерные средства для выполнения научно-исследовательских работ по закреплённой тематике	ПК-2.1. Демонстрирует способность и готовность к построению и исследованию математических моделей различных физических, биологических, экономических и социальных систем, а также применению идей, принципов и методов математического моделирования при решении прикладных задач. ПК-2.2. Использует естественнонаучные и математические знания для ориентирования в современном информационном пространстве ПК-2.3. Способен использовать естественнонаучные и математические знания для ориентирования в современном информационном пространстве. ПК-2.4. Собирает и анализирует информацию по решаемой задаче, составляет ее математическое описание, обеспечивает накопление, анализ и систематизацию собранных данных с использованием современных достижений науки и информационных систем; ПК-2.5. Выявляет и формулирует

		актуальные научные проблемы; обосновывать актуальность, теоретическую и практическую значимость темы научного исследования, разрабатывать план и программу проведения научного исследования; ПК-2.6. Проводит научно- исследовательские работы в области математики и компьютерных наук.
--	--	--

3. Описание планируемых результатов обучения по дисциплине

3.1. Описание планируемых результатов обучения по дисциплине

Планируемые результаты обучения по дисциплине представлены дескрипторами (знания, умения, навыки).

Дескрипторы по дисциплине	Знать	Уметь	Владеть
Код компетенции	ПК-1		
	- математический аппарат, методологию программирования и современные компьютерные технологии для решения практических задач получения, хранения, обработки и передачи информации; - основные идеи, понятия и методы, определяющие стиль написания, отладки и сопровождения программ; - характеристики основных парадигм программирования;	- использовать математический аппарат, методологию программирования и современные компьютерные технологии для решения практических задач получения, хранения, обработки и передачи информации; - применять современные компьютерные технологии для решения практических задач; - делать обоснованный выбор инструментария для решения прикладных задач;	- навыками использования математического аппарата, методологии программирования и современных компьютерных технологий для решения практических задач получения, хранения, обработки и передачи информации; - математическим аппаратом для построения вычислительных моделей практических задач; - навыками использования стандартных алгоритмических моделей для решения задач хранения и обработки информации
Код	ПК-2		

компетенции	
	Основы операционных систем: Знать архитектуру и основные компоненты операционных систем (например, Windows, Linux). Знать принципы работы файловых систем и управления процессами. Сетевые протоколы: Знать основные сетевые протоколы (TCP/IP, HTTP, HTTPS, FTP) и их функции. Знать принципы работы сетевых устройств (маршрутизаторов, коммутаторов, брандмауэров). Инструменты конфигурирования: Знать инструменты и утилиты для конфигурирования операционных систем и сетевых устройств (например, командная строка, PowerShell, SSH). Безопасность: Знать основные принципы и методы обеспечения безопасности операционных систем и сетевых устройств. Конфигурировать операционные системы: Уметь устанавливать и настраивать операционные системы для веб-разработки. Уметь управлять пользователями и правами доступа в операционных системах. Настраивать сетевые устройства: Уметь конфигурировать маршрутизаторы и коммутаторы для обеспечения сетевой связности. Уметь настраивать брандмауэры для защиты веб-приложений. Использовать инструменты командной строки: Уметь использовать командную строку для выполнения задач администрирования и конфигурирования. Уметь применять скрипты для автоматизации процессов конфигурирования. Обеспечивать безопасность систем: Уметь применять методы шифрования и аутентификации для защиты данных. Уметь проводить аудит безопасности и выявлять уязвимости в системах. Практическими навыками конфигурирования: Владеть навыками установки и настройки операционных систем в средах веб-разработки. Владеть навыками конфигурирования сетевых устройств для оптимизации работы веб-приложений. Работой с документацией: Владеть умением читать и интерпретировать техническую документацию по операционным системам и сетевым устройствам. Решением проблем: Владеть навыками диагностики и устранения неполадок в операционных системах и сетевых устройствах. Работой в команде: Владеть навыками взаимодействия с другими специалистами (разработчиками, системными

			администраторам и) для эффективного конфигурирования и поддержки веб-приложений.
--	--	--	--

3.2. Критерии оценки знаний студентов

Шкала оценивания	Индикаторы достижения	Показатели оценивания результатов обучения
ОТЛИЧНО/Зачтено	Знает:	- студент глубоко и всесторонне усвоил материал, уверенно, логично, последовательно и грамотно его излагает, опираясь на знания основной и дополнительной литературы, - на основе системных научных знаний делает квалифицированные выводы и обобщения, свободно оперирует категориями и понятиями.
	Умеет:	- студент умеет самостоятельно и правильно решать учебно-профессиональные задачи или задания, уверенно, логично, последовательно и аргументировано излагать свое решение, используя научные понятия, ссылаясь на нормативную базу.
	Владеет:	- студент владеет рациональными методами (с использованием рациональных методик) решения сложных профессиональных задач, представленных деловыми играми, кейсами и т.д.; При решении продемонстрировал навыки - выделения главного, - связкой теоретических положений с требованиями руководящих документов, - изложения мыслей в логической последовательности, - самостоятельного анализа факты, событий, явлений, процессов в их взаимосвязи и диалектическом развитии.
ХОРОШО/Зачтено	Знает:	- студент твердо усвоил материал, достаточно грамотно его излагает, опираясь на знания основной и дополнительной литературы, - затрудняется в формулировании квалифицированных выводов и обобщений, оперирует категориями и понятиями, но не всегда правильно их верифицирует.
	Умеет:	- студент умеет самостоятельно и в основном правильно решать учебно-профессиональные задачи или задания, уверенно, логично, последовательно и аргументировано излагать свое решение, не в полной мере используя научные понятия и ссылки на нормативную базу.
	Владеет:	- студент в целом владеет рациональными методами решения сложных профессиональных задач, представленных деловыми играми, кейсами и т.д.; При решении смог продемонстрировать достаточность, но не глубинность навыков

		- выделения главного, - изложения мыслей в логической последовательности. - связки теоретических положений с требованиями руководящих документов, - самостоятельного анализа факты, событий, явлений, процессов в их взаимосвязи и диалектическом развитии.
УДОВЛЕТВОРИТЕЛЬНО/Зачтено	Знает:	- студент ориентируется в материале, однако затрудняется в его изложении; - показывает недостаточность знаний основной и дополнительной литературы; - слабо аргументирует научные положения; - практически не способен сформулировать выводы и обобщения; - частично владеет системой понятий.
	Умеет:	- студент в основном умеет решить учебно-профессиональную задачу или задание, но допускает ошибки, слабо аргументирует свое решение, недостаточно использует научные понятия и руководящие документы.
	Владеет:	- студент владеет некоторыми рациональными методами решения сложных профессиональных задач, представленных деловыми играми, кейсами и т.д.; При решении продемонстрировал недостаточность навыков - выделения главного, - изложения мыслей в логической последовательности. - связки теоретических положений с требованиями руководящих документов, - самостоятельного анализа факты, событий, явлений, процессов в их взаимосвязи и диалектическом развитии.
НЕУДОВЛЕТВОРИТЕЛЬНО/Не зачтено	Знает:	- студент не усвоил значительной части материала; - не может аргументировать научные положения; - не формулирует квалифицированных выводов и обобщений; - не владеет системой понятий.
	Умеет:	студент не показал умение решать учебно-профессиональную задачу или задание.
	Владеет:	не выполнены требования, предъявляемые к навыкам, оцениваемым “удовлетворительно”.

При ответе на вопросы в рамках прохождения промежуточной аттестации (зачет/зачет с оценкой/экзамен) допускается вольная формулировка ответа, по смыслу раскрывающая содержание ответа, указанного в фонде оценочных средств, в качестве верного ответа.

4. Типовые контрольные задания (закрытого, открытого и иного типа) для проведения промежуточной аттестации, необходимые для оценки достижения компетенции, соотнесенной с результатами обучения по дисциплине

ТЕСТИРОВАНИЕ

9 СЕМЕСТР

ПК-1

1. Какой этап CRISP-DM предшествует непосредственно моделированию?
A) Business Understanding
B) Data Understanding
C) Data Preparation
D) Data Preparation
(Пояснение: сначала понимание данных, затем подготовка, и только потом моделирование)
2. Что из перечисленного НЕ относится к характеристикам «четырех V» больших данных?
A) Volume (объем)
B) Velocity (скорость)
C) Validity (валидность)
D) Variety (разнообразие)
3. Какая проблема решается на этапе «Data Preparation»?
A) Определение KPI проекта
B) Очистка данных и заполнение пропусков
C) Сбор первичных данных
D) Развертывание модели в продуктивной среде
4. Заключительный этап CRISP-DM, на котором модель внедряется в бизнес-процессы:
A) Evaluation
B) Deployment
C) Modeling
D) Business Understanding
5. Что понимается под «проклятием размерности» (curse of dimensionality)?
A) Невозможность хранения многомерных массивов
B) Рост вычислительной сложности при увеличении количества объектов
C) Экспоненциальный рост объема пространства признаков, приводящий к разрежению данных
D) Проблема лицензирования многомерных баз данных
6. Какой из перечисленных этических аспектов наиболее критичен при работе с персональными данными в РФ?
A) Скорость обработки данных
B) Согласие субъекта на обработку и соблюдение требований 152-ФЗ
C) Использование открытых библиотек
D) Регулярное резервное копирование
7. Какой метод заполнения пропусков (missing values) является простейшим, но

может сильно исказить распределение?

- A) Интерполяция сплайнами
- B) Предсказание пропусков моделью
- C) Заполнение средним арифметическим**
- D) Удаление всех строк с пропусками

8. Какое преобразование нормализует данные к диапазону $[0, 1]$?

- A) Min-Max Normalization**
- B) Z-score Standardization
- C) Log Transformation
- D) Box-Cox Transformation

9. Какая метрика наилучшим образом визуализируется с помощью «ящика с усами» (boxplot)?

- A) Коэффициент корреляции
- B) Выбросы (outliers)**
- C) Нормальность распределения
- D) Кластерная структура

10. Для номинального категориального признака «цвет» (красный, синий, зеленый) без внутреннего порядка корректно применить:

- A) Label Encoding
- B) One-Hot Encoding**
- C) Target Encoding (без дополнительного анализа)
- D) Min-Max Normalization

11. Алгоритм SMOTE используется для:

- A) Снижения размерности
- B) Балансировки классов путем генерации синтетических объектов миноритарного класса**
- C) Поиска выбросов
- D) Кодирования категорий

12. На тепловой карте корреляций (correlation heatmap) значение коэффициента корреляции Пирсона $+0.95$ между двумя признаками означает:

- A) Отсутствие связи
- B) Сильную положительную линейную связь**
- C) Сильную отрицательную линейную связь
- D) Ошибку в вычислениях

13. Какая метрика показывает долю правильно классифицированных объектов среди всех объектов?

- A) Recall
- B) Accuracy**
- C) Precision
- D) F1-score

14. Матрица ошибок (confusion matrix) для бинарной классификации имеет следующий вид: TP=80, FN=20, FP=10, TN=90. Чему равна полнота (Recall)?

- A) 0.8
- B) 0.89
- C) 0.8**

D) 0.9

15. В методе k-ближайших соседей (k-NN) увеличение параметра k обычно приводит к:

- A) Уменьшению переобучения и сглаживанию границы решения
- B) Увеличению переобучения
- C) Ускорению работы алгоритма
- D) Необходимости нормализации данных

9 СЕМЕСТР

ПК-2

1. Какая функция потерь (loss function) чаще всего используется в задаче линейной регрессии с нормальным распределением ошибок?

- A) MAE
- B) **MSE (Mean Squared Error)**
- C) Hinge Loss
- D) Cross-Entropy

2. Значение $R^2 = 0.85$ в задаче регрессии означает:

- A) Модель предсказывает хуже, чем среднее значение
- B) **Модель объясняет 85% дисперсии целевой переменной**
- C) Средняя ошибка модели составляет 85%
- D) Модель переобучена

3. Какой алгоритм является ансамблевым и использует бэггинг (bagging) на основе деревьев решений?

- A) Gradient Boosting
- B) AdaBoost
- C) **Random Forest**
- D) Logistic Regression

4. Признак переобучения (overfitting) модели – это:

- A) Низкая ошибка на обучающей выборке и низкая на тестовой
- B) **Низкая ошибка на обучающей и высокая на тестовой выборке**
- C) Высокая ошибка на обучающей и низкая на тестовой
- D) Высокая ошибка на обеих выборках

5. Метод кросс-валидации (k-Fold Cross-Validation) используется для:

- A) **Более надежной оценки качества модели и борьбы с переобучением**
- B) Ускорения обучения моделей
- C) Очистки данных от выбросов
- D) Визуализации данных

6. Алгоритм K-means относится к классу задач:

- A) Регрессии
- B) Классификации
- C) **Кластеризации**
- D) Поиска ассоциативных правил

7. Какая метрика расстояния наименее чувствительна к выбросам?

- A) Евклидово расстояние
- B) **Манхэттенское расстояние (L1)**
- C) Косинусное сходство

- D) Расстояние Махаланобиса
8. Метод определения оптимального числа кластеров в K-means, основанный на анализе суммы квадратов внутрикластерных расстояний:
- A) Силуэтный коэффициент
 - B) Метод локтя (Elbow method)**
 - C) Индекс Rand
 - D) P-значение
9. Алгоритм DBSCAN способен находить кластеры:
- A) Только сферической формы
 - B) Произвольной формы и выделять шум**
 - C) Только заранее заданного количества
 - D) Только при нормализации данных
10. Какой метод снижения размерности является линейным и ищет направления максимальной дисперсии?
- A) t-SNE
 - B) Principal Component Analysis (PCA)**
 - C) Autoencoder
 - D) Isomap
11. Силуэтный коэффициент (silhouette score) принимает значения в диапазоне:
- A) [0, 1]
 - B) [-1, 1]**
 - C) $[-\infty, +\infty]$
 - D) $[0, +\infty]$
12. Метрика «достоверность» (confidence) правила $A \rightarrow B$ определяется как:
- A) $P(B|A)$ или $P(A \cap B)/P(A)$**
 - B) $P(A \cap B)$
 - C) $P(A \cap B) / (P(A) P(B))$
 - D) $P(A \cup B)$
13. Ключевое преимущество алгоритма FP-Growth перед Apriori:
- A) Отсутствие генерации кандидатов и двухпроходный обход данных**
 - B) Лучшая работа с категориальными данными
 - C) Возможность параллелизации на MapReduce
 - D) Встроенная визуализация правил
14. В экосистеме Apache Spark базовым абстрактным типом данных, обеспечивающим отказоустойчивость, является:
- A) DataFrame
 - B) RDD (Resilient Distributed Dataset)**
 - C) Dataset
 - D) DataLake
15. Операция в Apache Spark, которая запускает вычисление и возвращает результат в драйвер (например, count, collect), называется:
- A) Трансформацией
 - B) Действием (Action)**
 - C) Шардированием
 - D) Ленивой инициализацией