ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«24» июня 2025 г.
Протокол № 2

Рабочая программа дисциплины (модуля)
«Machine Learning (Машинное обучение)»

Направление подготовки: *02.04.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Продуктовая аналитика*

Квалификация (степень) выпускника: *магистр*

Форма обучения: *очная*

Срок освоения программы: *2 года*

Год набора: *2025*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2025

Содержание

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Machine Learning (Машинное обучение)» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – магистратура по специальности 02.04.01 Математика и компьютерные науки, профиль Продуктовая аналитика, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 810 от 23.08.2017 года.

Изучение дисциплины (модуля) Machine Learning (Машинное обучение) позволяет сформировать системное понимание методов анализа данных и разработки интеллектуальных моделей, применяемых для решения прикладных и бизнес-задач, а также научиться применять современные алгоритмы и инструменты машинного обучения для повышения качества принимаемых решений и создания конкурентоспособных цифровых продуктов.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки магистратуры по направлению 02.04.01 Математика и компьютерные науки, профиль Продуктовая аналитика и входит в обязательную часть Блока 1.

Дисциплина (модуль) изучается на 1 курсе во 2 семестре, доступна для прохождения при условии успешного завершения дисциплины (модуля) «Основы Python».

Цель изучения дисциплины (модуля): формирование системного понимания методологии Data Science и практических навыков разработки, валидации и интерпретации моделей машинного обучения для решения прикладных и бизнес-задач.

Задачи изучения дисциплины (модуля):

- освоить этапы жизненного цикла ML-проекта и принципы работы ключевых алгоритмов машинного обучения для задач регрессии, классификации, кластеризации и uplift-моделирования;
- научиться проводить разведочный анализ данных, выполнять генерацию и отбор признаков, предотвращать утечки данных и готовить данные к моделированию;
- сформировать умение обучать, настраивать и валидировать модели с использованием кросс-валидации, пайплайнов, функций потерь и метрик качества;
- освоить методы борьбы с переобучением, регуляризации и интерпретации моделей с использованием SHAP, LIME и других инструментов объяснимого ИИ;
- развить навыки построения воспроизводимых ML-решений с применением современных библиотек Python и интерпретации результатов в бизнес-контексте.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- основы методологии Data Science, машинного обучения, включая этапы жизненного цикла ML-проектов;
- принципы работы ключевых алгоритмов: линейные модели, kNN, деревья решений, ансамбли (Random Forest, бустинг), кластеризация и снижение размерности;
- метрики качества, функции потерь и методы валидации моделей для задач регрессии, классификации и uplift-моделирования;
- концепции смещения и разброса, переобучения, регуляризации, а также подходы к интерпретации моделей (SHAP, LIME);

— типы признаков и методы работы с ними, включая генерацию, отбор и предотвращение утечек данных.

уметь:

— проводить разведочный анализ данных (EDA), формулировать гипотезы и готовить данные для моделирования;
— обучать, настраивать и оценивать модели машинного обучения с использованием кросс-валидации, пайплайнов и метрик;
— реализовывать задачи кластеризации, детекции аномалий и снижения размерности, интерпретируя результаты;
— строить и валидировать uplift-модели, а также генерировать признаки для сложных данных (временные, гео, финансы и др.);
— интерпретировать модели и визуализировать влияние признаков с помощью SHAP, LIME и других инструментов.

владеть:

— инструментами Python для анализа и моделирования: pandas, numpy, scikit-learn;
— продвинутыми библиотеками ML: XGBoost, CatBoost, lightgbm, а также SHAP, LIME для интерпретации;
— навыками построения воспроизводимых ML-пайплайнов и работы с разнотипными данными;
— методами предобработки, валидации и оптимизации моделей в реальных задачах;
— практиками применения и интерпретации моделей в бизнес-контексте, включая причинно-следственный анализ.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
ОПК-3.	Способен самостоятельно создавать прикладные программные средства на основе современных информационных технологий и сетевых ресурсов, в том числе отечественного производства	ОПК-3.1.	Знает основные принципы программирования, архитектуры программного обеспечения и современные языки программирования, а также особенности отечественных информационных технологий и сетевых ресурсов
		ОПК-3.2.	Умеет разрабатывать прикладные программные средства, используя современные инструменты и технологии, а также интегрировать их с сетевыми ресурсами для решения конкретных задач
		ОПК-3.3.	Имеет практический опыт разработки программных средств, используемых при построении математических моделей в естественных науках
ПК-3.	Способен решать задачи профессиональной деятельности в области продуктовой аналитики, формулировать результаты анализа и выявлять последствия полученных данных для принятия обоснованных решений и оптимизации продуктов	ПК-3.1.	Знает методы и инструменты продуктовой аналитики
		ПК-3.2.	Умеет применять аналитические инструменты и программное обеспечение для обработки и визуализации данных, а также формулировать выводы на основе проведенного анализа
		ПК-3.3.	Имеет опыт работы над реальными проектами в области продуктовой аналитики, включая анализ пользовательского поведения и оптимизацию продуктов на основе полученных данных
ПК-6.	Способен разрабатывать программное обеспечение для решения прикладных задач в сфере продуктовой аналитики	ПК-6.1.	Знает основные языки программирования, методы разработки программного обеспечения, а также принципы проектирования и архитектуры программных систем, применяемых в продуктовой аналитике
		ПК-6.2.	Умеет анализировать прикладные задачи, разрабатывать алгоритмы и реализовывать их в виде программного обеспечения, используя современные

			инструменты и технологии, а также проводить тестирование и отладку созданных решений
		ПК-6.3.	Имеет практический опыт разработки программного обеспечения в рамках реальных проектов, включая участие в командах, где были успешно реализованы решения для конкретных прикладных задач

3. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (Практические занятия)			
1	Основные термины машинного обучения и смежных областей. Оценка качества моделей.	12	6		20	Домашние задания Тесты
2	Линейные модели	8	4		14	Домашние задания Соревнование Тесты
3	Деревья решений и модели, основанные на них	20	10		31	Домашние задания Тесты
4	Обучение без учителя и работа с признаками	20	10		31	Домашние задания Проект Тесты
	Экзамен			***4***		
	Итого:	***60***	***30***	***4***	***96***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

4. Содержание дисциплины (модуля)

№ п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Основные термины машинного обучения и смежных областей. Оценка качества моделей.	Метрики и функционалы потерь в задачах регрессии и классификации Контроль качества и выбор модели Метрические алгоритмы
2	Линейные модели	Линейная регрессия Логистическая регрессия, SVM
3	Деревья решений и модели, основанные на них	Решающие деревья Сложность алгоритмов Ансамбли алгоритмов Случайные леса Градиентный бустинг
4	Обучение без учителя и работа с признаками	Аплифт-моделирование – принципы и методы. Кластеризация и методы понижения размерности Обнаружение аномалий Генерация и отбор признаков Интерпретация моделей и диагностика сдвига данных:

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Бринк, Х. Машинное обучение : практическое руководство / Х. Бринк, Д. Ричардс, М. Феверолф. - Санкт-Петербург : Питер, 2018. - 336 с. - (Серия «Библиотека программиста»). - ISBN 978-5-496-02989-6. - Текст : электронный. - URL: https://znanium.com/catalog/product/1766396.

2. Лакшманан, В. Машинное обучение. Паттерны проектирования : практическое пособие / В. Лакшманан, С. Робинсон, М. Мунн. - Санкт-Петербург : БХВ-Петербург, 2022. - 448 с. - ISBN 978-5-9775-6797-8. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2140204.

Дополнительная литература:

1. Коэльо, Л. Построение систем машинного обучения на языке Python : практическое руководство / Л. Коэльо, В. Ричарт ; пер. с англ. А. А. Слинкина. - 3-е изд. - Москва : ДМК Пресс, 2023. - 304 с. - ISBN 978-5-89818-331-8. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2102618.

2. Григорьев, А. Машинное обучение. Портфолио реальных проектов : практическое руководство / А. Григорьев. - Санкт-Петербург : Питер, 2023. - 496 с. - (Серия «Библиотека программиста»). - ISBN 978-5-4461-1978-3. - Текст : электронный. - URL: https://znanium.com/catalog/product/2123375.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1	Катастрофы, стихийные бедствия, аварии, эпидемии. Солнечная и геомагнитная активность. /ежедневный обзор	http://www.disasters.chat.ru
2	Каталог по безопасности жизнедеятельности	http://www.eun.chat.ru
3	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
4	База данных для IT-специалистов	https://habr.com
5	База данных ScienceDirect	https://www.sciencedirect.com
6	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
7	Федеральный портал «Российское образование"	https://www.edu.ru/
8	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
9	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
10	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		

КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Machine Learning (Машинное обучение)» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, соревнования, тесты, проект, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал, использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Соревнование – организованное мероприятие, в рамках которого участники соперничают друг с другом для достижения определенной цели, демонстрируя свои навыки, знания или способности в заданной области.

Тест – особая форма проверки знаний. Проводится после освоения одной или

нескольких тем и свидетельствует о качестве понимания основных понятий изучаемого материала. Тестовые задания составлены к ключевым понятиям, основным разделам, важным терминологическим категориям изучаемой дисциплины (модуля).

Для подготовки к тесту необходимо знать терминологический аппарат дисциплины (модуля), понимать смысл научных категорий и уметь их использовать в профессиональной лексике. Владение понятийным аппаратом, включённым в тестовые задания, позволяет преподавателю быстро проверить уровень понимания студентами важных методологических категорий.

Проект – исследовательская работа по дисциплине (модулю) и презентация результатов.

Для успешной подготовки к проекту рекомендуется: четко определить цели и задачи проекта; составить план работы, разбив проект на этапы с указанием сроков выполнения каждого из них; использовать разнообразные источники информации и инструменты для исследования темы; регулярно проверять прогресс и вносить коррективы в план, если это необходимо.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Machine Learning (Машинное обучение)».

Оценивание уровня учебных достижений обучающихся по дисциплине (модулю) осуществляется в виде текущего контроля успеваемости.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***экзамена***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом:

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину (модуль). Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

Дисциплина (модуль) «Machine Learning (Машинное обучение)» оценивается следующим образом:

Активность	Вес	Количество	Описание
Накопительная оценка			
Домашние задания	60%	8	Набор задач по темам недели
Соревнование		1	Kaggle-style соревнование с задачей на ML
Тесты		5	Письменная работа, состоящая из вопросов с вариантами ответов, на которые нужно ответить за ограниченное время
Проект		1	Исследовательская работа по дисциплине (модулю) и презентация результатов
Промежуточная аттестация			
Экзамен	40%	1	Письменная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Итоговая оценка рассчитывается по накопительной при условии, если средний балл студента составляет 4 и более баллов, по формуле: «0,35 × среднее за домашние задания + 0,25 × соревнование + 0,15 × среднее за тесты + 0,25 × проект».

Если студент не выполняет условие для получения оценки по накопительной системе, ему необходимо сдать экзамен. **В данном случае формула расчёта итоговой оценки по дисциплине (модулю) «Machine Learning (Машинное обучение)»:** «0,6 × накопительная оценка (0,35 × среднее за домашние задания + 0,25 × соревнование + 0,15 × среднее за тесты + 0,25 × проект) + 0,4 × экзамен».

В случае, если средний балл студента составляет 4 и более баллов, но он хочет улучшить оценку, итоговая оценка по дисциплине (модулю) выставляется по формуле: «0,7 × накопительная оценка (0,35 × среднее за домашние задания + 0,25 × соревнование + 0,15 × среднее за тесты + 0,25 × проект) + 0,3 × экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1

Выберите датасет для задачи регрессии или классификации, выполните разведочный анализ данных, разделите выборку на обучающую и тестовую, обучите не менее двух моделей (одну метрическую, например kNN), рассчитайте соответствующие метрики качества и функции потерь, выполните кросс-валидацию, сравните результаты моделей и обоснуйте выбор лучшей модели с точки зрения контроля качества и устойчивости.

Домашнее задание 2

Реализуйте линейную регрессию для задачи прогнозирования числового признака, оцените качество модели с использованием MAE, MSE и R^2, проанализируйте влияние регуляризации, затем решите задачу бинарной классификации с использованием логистической регрессии и SVM, сравните их по метрикам качества и визуализируйте границы принятия решений.

Домашнее задание 3

Обучите модель решающего дерева для задачи классификации или регрессии, проанализируйте глубину дерева и влияние гиперпараметров на переобучение, затем реализуйте случайный лес и градиентный бустинг (например, XGBoost или LightGBM), сравните модели по качеству и скорости обучения, выполните подбор гиперпараметров и сделайте выводы о преимуществах ансамблей по сравнению с одиночным деревом.

Домашнее задание 4

Выберите датасет и выполните кластеризацию (например, K-means или DBSCAN), примените метод снижения размерности (PCA или t-SNE) для визуализации, реализуйте алгоритм обнаружения аномалий, выполните генерацию и отбор признаков, проверьте наличие утечек данных, постройте uplift-модель для оценки инкрементального эффекта (при наличии подходящих данных), интерпретируйте результаты с помощью SHAP или аналогичного инструмента и проанализируйте возможный сдвиг данных между обучающей и тестовой выборками.

Примерные задания для тестов

1. Что нужно усреднить в задаче бинарной классификации, чтобы получить сбалансированную точность ?

   (a) точность и полноту
   (b) чувствительность и специфичность
   (c) полноту класса 1 и класса 0
   (d) точность класса 1 и класса 0

2. Требуется по описанию проекта предсказать его доходность (в рублях) через год после начала реализации. Предполагается, что имеется база подобных проектов с подробной статистикой их дальнейшей судьбы.

   (a) это задача классификации
   (b) это задача регрессии
   (c) это задача обучения с размеченными данными
   (d) это задача обучения без меток

3. Какой метод кластеризации актуален для иерархических структур?

   (a) K-Means
   (b) DBSCAN
   (c) Агломеративная кластеризация
   (d) PCA

4. Что представляет собой проблема "bias-variance trade-off"?

   (a) Компромисс между простотой и функциональной выразимостью модели
   (b) Выбор между различными алгоритмами машинного обучения
   (c) Определение оптимального порога для классификации данных

5. Требуется по описанию участников торгов выявить потенциальных мошенников. При этом есть истории предыдущих торгов, случаев доказанного мошенничества пока нет.

   (a) это задача классификации
   (b) это задача регрессии
   (c) это задача обучения с размеченными данными
   (d) это задача обучения без меток

6. Что происходит при увеличении k в методе kNN (отметьте все подходящие варианты)?

   (a) Улучшается качество
   (b) Алгоритм превращается в константный (для заданной обучающей выборки)
   (c) Увеличивается число гиперпараметров
   (d) Алгоритм "упрощается"(ответы становятся более стабильными при небольших изменениях обучающей выборки)

7. Какая из перечисленных тем не относится к методам кластеризации в машинном обучении?

   (a) K-Means
   (b) Support Vector Machines
   (c) DBSCAN
   (d) Hierarchical Clustering

8. Каким образом может оцениваться качество модели линейной классификации?

   (a) С помощью точности (accuracy)
   (b) С помощью матрицы ошибок (confusion matrix)
   (c) С помощью F1-меры (F1-score)

9. Почему наивная форма стекинга не применяется на практике (в отличие от других форм)?

   (a) из-за переобучения
   (b) из-за сложности
   (c) из-за того, что является частной формой бустинга

10. Как связано понятие BVD с терминами переобучения и недообучения?

    (a) Высокое смещение, низкий разброс – недообучение
    (b) Низкое смещение, высокий разброс – переобучение
    (c) Смещение и разброс не связаны с обучением
    (d) Высокий разброс, низкое смещение – недообучение

11. Выберите верные утверждения для гиперпараметров:

    (a) Гиперпараметров больше, чем параметров
    (b) Гиперпараметры настраиваются градиентным спуском
    (c) Гиперпараметры настраиваются в результате обучения (метод fit)
    (d) Значения гиперпараметров можно выбрать с помощью перекрестной проверки

12. Что представляет собой метод t-SNE в контексте уменьшения размерности?

    (a) Генерация новых признаков на основе комбинаций исходных признаков
    (b) Преобразование пространства признаков в пространство меньшей размерности с сохранением относительных расстояний
    (c) Классификация объектов на основе различия в их признаках
    (d) Оптимизация функции потерь для снижения ошибки модели

13. Для чего может использоваться L1-регуляризация?

    (a) Для уменьшения переобучения модели
    (b) Для нормировки коэффициентов
    (c) Для отбора признаков
    (d) Для подбора гиперпараметров

14. В косых деревьях (oblique decision trees)... :

    (a) Используются предикаты с линейными комбинациями признаков
    (b) Используются не все признаки
    (c) Используются ядра
    (d) Используется подрезка

15. В каких схемах контроля каждый объект гарантировано ровно 1 раз попадает в тестовую выборку?

    (a) K-fold CV
    (b) Стратифицированный K-fold CV
    (c) Бутстреп
    (d) LOO

16. Минимальное значение среднего внутрикластерного расстояния достигается...

    (a) На одном кластере
    (b) На числе кластеров, равном числу объектов
    (c) В точке перегиба

17. Какие два основных типа линейной регрессии существуют?

    (a) Простая линейная регрессия и множественная линейная регрессия
    (b) Логистическая регрессия и полиномиальная регрессия
    (c) Сигмоидная регрессия и эластичная регрессия

18. Что в стекинге называется метаалгоритмом? (выберите все правильные ответы)

    (a) базовый алгоритм
    (b) алгоритм, который учится на ответах базового
    (c) алгоритм, ответы которого будут итоговыми

19. Почему на практике минимизируют эмпирический риск?

    (a) теоретический риск невозможно минимизировать в общем случае

    (b) это делает модель проще

    (c) это позволяет уменьшить эффект переобучения / запоминания

20. На каких весах считается регуляризация в линейной регрессии и почему?

    (a) На всех, так как это позволяет избежать переобучения

    (b) На всех, так как это позволяет нормировать признаки

    (c) На всех кроме w_0, так как w_0 отвечает за «масштаб»

    (d) На всех кроме случайно выбранного, так как это позволяет избежать переобучения

21. Какие из функций ошибок / функционалов качества эквивалентны MSE (минимум MSE соответствует оптимуму функции)?

    (a) RMSE

    (b) Хьюбера

    (c) logcosh

    (d) R^2

22. Какая из компонент в BVD отвечает за чувствительность модели к конкретному набору данных?

    (a) Смещение

    (b) Разброс

    (c) Ошибка

    (d) Шум

23. Какие ансамбли являются параллельными? (выберите все правильные ответы)

    (a) бустинг

    (b) бэгинг

    (c) случайный лес

24. Что такое разброс (variance) модели?

    (a) Ошибка, связанная с неправильным представлением о целевой функции или недостаточной сложностью модели

    (b) Ошибка, связанная с шумом в данных

    (c) Ошибка, связанная с переобучением модели

25. Что пытается оптимизировать SNE (Stochastic Neighbor Embedding)?

    (a) Снижение размерности с сохранением глобальной структуры

    (b) Вероятностное приближение расстояний между точками в высокоразмерном и низкоразмерном пространствах

    (c) Увеличение плотности кластеров

    (d) Максимизация линейной зависимости между переменными

26. Как работает алгоритм градиентного бустинга?

    (a) Каждое следующее дерево итерационно настраивает на градиент ошибки

    (b) Каждое следующее дерево итерационно настраивается на антиградиент ошибки

    (c) Сумма деревьев совместно настраивается на градиент ошибки

    (d) Сумма деревьев совместно настраивается на антиградиент ошибки

27. Какой метод кластеризации дает возможность определять количество кластеров в данных автоматически?

    (a) K-Means

    (b) Hierarchical Clustering

    (c) DBSCAN

    (d) Agglomerative Clustering

28. Выберите все верные утверждения про ROC-AUC

    (a) Отражает качество ранжирования объектов
    (b) Не зависит от масштаба значений, которые выдает алгоритм
    (c) Если инвертировать порядок объектов в модели с ROC-AUC = 0, то получим ROC-AUC = 1
    (d) Можно считать, как долю верно отранжированных пар

29. Какой метод кластеризации позволяет размечать выбросы?

    (a) Mini-Batch K-Means
    (b) K-Means
    (c) DBSCAN
    (d) Иерархическая кластеризация

30. Какие из перечисленных алгоритмов ленивые?

    (a) 1NN
    (b) 3NN
    (c) ближайший центроид
    (d) линейная регрессия

Примерное задание и критерии оценивания к соревнованию

Задание:

На основе предоставленных данных необходимо:

1. Определить целевую переменную и тип задачи.
2. Разделить данные на обучающую и валидационную выборки.
3. Обучить и сравнить три модели:

- линейную модель (линейная или логистическая регрессия);
- SVM;
- k ближайших соседей (kNN).

4. Выбрать лучшую модель по валидационной метрике.
5. Сделать финальное предсказание для тестовой выборки.

Разрешается:

- использовать масштабирование признаков;
- подбирать гиперпараметры в разумных пределах;
- использовать кросс-валидацию (по возможности).

Ограничения:

- Используются только изученные модели: линейные, SVM, kNN.
- Нельзя использовать ансамбли и бустинг.
- Тестовая выборка используется только один раз для финальной проверки.

Критерии оценки:

Оценивание проводится по следующим критериям:

- Итоговое значение основной метрики на тестовой выборке.
- Корректность выбора метрики в соответствии с типом задачи.
- Наличие сравнения трёх обязательных моделей.
- Корректность разбиения данных и процедуры валидации.
- Обоснование выбора финальной модели.
- Использование корректной терминологии машинного обучения.

Примерное задание и критерии оценивания к проекту

Описание проекта:

Проект направлен на решение практической задачи машинного обучения (регрессии, классификации или аплифт-моделирования) с применением нескольких классов моделей, сравнением их качества и обоснованием выбора наилучшего решения.

Студент должен продемонстрировать владение основными понятиями машинного обучения, умение выбирать метрики и функционалы потерь, проводить контроль качества модели, работать с признаками и интерпретировать результаты.

В проекте необходимо реализовать и сравнить не менее трёх различных подходов:

- линейные модели;
- деревья решений и ансамбли;
- метрические алгоритмы;
- минимум один метод обучения без учителя (при обосновании применимости).

Цели проекта:

1. Закрепление понимания ключевых терминов и концепций машинного обучения.
2. Освоение методов оценки качества моделей и выбора метрик.
3. Практическое применение линейных моделей, деревьев решений, ансамблей и метрических алгоритмов.
4. Развитие навыков работы с признаками и интерпретации моделей.
5. Формирование навыков выбора и обоснования финальной модели.

Задачи проекта:

1. Постановка задачи

- Определить тип задачи и целевую переменную.
- Обосновать выбор метрик качества.
- Описать данные и их особенности.

2. Анализ и подготовка данных

- Провести исследовательский анализ данных.
- Обработать пропуски, выбросы и категориальные признаки.
- Выполнить генерацию и/или отбор признаков.
- При необходимости применить методы понижения размерности.

3. Обучение и сравнение моделей

Реализовать и сравнить:

- Линейные модели (линейная или логистическая регрессия, SVM).
- Деревья решений и ансамбли (решающее дерево, случайный лес, градиентный бустинг).
- Метрические алгоритмы (kNN).
- Минимум один метод обучения без учителя (кластеризация, обнаружение аномалий и др.).

4. Оценка качества и выбор модели

- Разделить данные на обучающую и тестовую выборки.
- Использовать кросс-валидацию.
- Проанализировать переобучение.
- Сравнить модели по выбранным метрикам.
- Обосновать выбор финальной модели.

5. Интерпретация и диагностика

- Проанализировать важность признаков.
- Провести анализ ошибок.
- Дать интерпретацию полученных результатов.
- Оценить устойчивость модели и возможный сдвиг данных (при наличии).

Критерии оценивания:

Оценка проекта осуществляется по следующим критериям:

- Корректность постановки задачи и использование терминологии машинного обучения.
- Обоснованность выбора метрик качества и функционалов потерь.
- Полнота и глубина анализа данных.
- Качество предобработки и работы с признаками.
- Реализация требуемых классов моделей.

- Корректность процедуры обучения и валидации.
- Сравнительный анализ моделей.
- Обоснованность выбора итоговой модели.
- Анализ переобучения и устойчивости модели.
- Интерпретация результатов и качество выводов.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Как называется множество всех возможных объектов в задаче машинного обучения? **Ответ запиши в виде словосочетания на русском языке.**	Пространство объектов/ пространство объектов	ПК-3
2.	Как называется средняя ошибка модели на обучающей выборке? **Ответ запиши в виде словосочетания на русском языке.**	Эмпирический риск/ эмпирический риск	ПК-6
3.	Верно ли то, что L1-регуляризация (LASSO) склонен к отбору признаков? **Ответ запиши в виде Да/Нет.**	да / Да / верно / Верно	ПК-3
4.	Как называется метод оценки качества модели через генерацию подвыборок с возвращением? **Ответ запиши в виде одного слова на английском или русском языках.**	Бутстреп/бутстреп/ bootstrap/Bootstrap	ПК-3
5.	Средняя …….. ошибка - это функция ошибки, вычисляемая как средний модуль отклонений **Ответ запиши в виде одного слова на русском языке.**	абсолютная/Абсолютная	ПК-6
6.	Среднеквадратичная _____ - это квадратный корень из MSE. Какое слово пропущено? **Ответ запиши в виде одного слова на русском языке.**	ошибка/Ошибка	ПК-6
7.	Какая компания создала наиболее популярную библиотеку градиентного бустинга Categorical Boosting (CatBoost)? **Ответ запиши в виде одного слова на русском языке.**	Яндекс/яндекс	ПК-6
8.	Верно ли то, что метод главных компонент - это метод, который проецирует данные на второстепенные компоненты? **Ответ запиши в виде Да/Нет.**	нет / Нет / не верно / Не верно	ПК-6
9.	Что такое "метка" в обучении с учителем? А. Значение, которое модель предсказывает Б. Признак объекта В. Функция ошибки Г. Гиперпараметр модели	А	ОПК-3
10.	Как называется вариант kNN, где ближайшие соседи влияют сильнее? А. kNN Б. Взвешенный kNN В. Soft-Margin SVM Г. Бэггинг	Б	ОПК-3
11.	Что решает проблему вырожденности матрицы в линейной регрессии? А. Увеличение обучающей выборки	Б	ОПК-3

	Б. Регуляризация (например, Ridge) В. Уменьшение числа признаков Г. Использование kNN Д. Стекинг Е. PCA		
12.	Как в случайном лесе вычисляется важность признака? А. Через корреляцию с целевой переменной Б. Через уменьшение ошибки после расщепления по признаку В. Через веса в линейной модели Г. Через расстояние между объектами	Б	ОПК-3