

**Федеральное государственное образовательное бюджетное
учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)**

**Департамент анализа данных и машинного обучения
Факультета информационных технологий и анализа больших данных**

УТВЕРЖДАЮ

Проректор по учебной и
методической работе

_____ Е.А. Каменева
29.12.2022 г.

Коротеев М.В.

Машинное обучение

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
09.03.03 - Прикладная информатика,
ОП «Инженерия данных»,
ОП «Прикладные информационные системы в экономике и финансах»

*Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших данных
(протокол №27 от 15.12.2022 г.)*

*Одобрено Советом учебно-научного
Департамента анализа данных и машинного обучения
(протокол №6 от 13.12.2022 г.)*

Москва 2022

Оглавление

1. Наименование дисциплины.....	2
2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине.....	2
3. Место дисциплины в структуре образовательных программ.....	3
4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся.....	3
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий.....	4
5.1. Содержание дисциплины.....	4
5.2. Учебно-тематический план.....	7
5.3. Содержание семинаров, практических занятий.....	10
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине.....	12
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы.....	12
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю.....	14
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине.....	18
8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины.....	25
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.....	25
10. Методические указания для обучающихся по освоению дисциплины	27
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем.....	30
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.....	30

1. Наименование дисциплины

«Машинное обучение».

2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКН-4	Способность проектировать и создавать интеллектуальные информационные системы, выбирать метод обучения в соответствии с анализом задачи	Демонстрирует знание основных понятий машинного обучения и интеллектуального анализа данных, понимание области и границ применимости, основные виды задач.	Знать: основные понятия машинного обучения и интеллектуальной обработки данных, их области применения и сравнительные характеристики Уметь: применять основные методы обучения с учителем и без учителя для решения типовых и практико-ориентированных задач машинного обучения и анализа данных
		Демонстрирует знание популярных инструментальных средств машинного обучения, собирает датасет, строит модели, проводит их анализ и диагностику, делает содержательные выводы.	Знать: основные инструментальные средства (языки программирования, библиотеки, фреймворки) для обработки данных и машинного обучения Уметь: применять современные и популярные инструментальные средства для решения практических задач машинного обучения
		Презентабельно демонстрирует результаты анализа данных и машинного обучения в форме, доступной непрофессионалу, структурирует отчет по	Знать: основные средства и приемы визуализации данных и представления результатов машинного обучения, включая основные метрики и нефункциональные требования к интеллектуальным системам Уметь: представлять результаты проведенного анализа в форме, понятной

		проведенному анализу.	неспециалисту; визуализировать данные и обнаруженные в них зависимости, проводить презентацию проведенного анализа
--	--	-----------------------	--

3. Место дисциплины в структуре образовательных программ

Дисциплина «Машинное обучение» относится к Общепрофессиональному циклу дисциплин по направлению подготовки 09.03.03 - Прикладная информатика, ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах».

Изучение дисциплины «Машинное обучение» основывается на сумме знаний, полученных при изучении дисциплины «Алгоритмы и структуры данных в языке Python», «Анализ данных» либо «Теория вероятностей и математическая статистика», «Иностранный язык», «Дискретная математика». Для изучения данной дисциплины студент должен обладать базовыми знаниями в области информационных технологий и программирования, навыками программирования на языке Python.

4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся

ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах»
Очная форма обучения

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 3 (в часах)	Семестр 4 (в часах)
Общая трудоёмкость дисциплины	8/288	144	144
Контактная работа-Аудиторные занятия	100	50	50
Лекции	32	16	16
Семинары, практические занятия	68	34	34
Самостоятельная работа	188	94	94
Вид текущего контроля	контрольная работа	контрольная работа	-
Вид промежуточной аттестации	зачет, экзамен	зачет	экзамен

ОП «Прикладные информационные системы в экономике и финансах»

Очно-заочная форма обучения

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 3 (в часах)	Семестр 4 (в часах)
Общая трудоёмкость дисциплины	8/288	144	144
Контактная работа-Аудиторные занятия	68	34	34
Лекции	32	16	16
Семинары, практические занятия	36	18	18
Самостоятельная работа	220	110	110
Вид текущего контроля	контрольная работа	контрольная работа	-
Вид промежуточной аттестации	зачет, экзамен	зачет	экзамен

Институт онлайн-образования, заочная форма обучения

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 4 (в часах)	Семестр 5 (в часах)
Общая трудоёмкость дисциплины	8/288	144	144
Контактная работа-Аудиторные занятия	32	16	16
Лекции	8	4	4
Семинары, практические занятия	24	12	12
Самостоятельная работа	256	92	164
Вид текущего контроля	контрольные работы	контрольные работы	контрольные работы
Вид промежуточной аттестации	зачет, экзамен	зачет	экзамен

5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий

5.1. Содержание дисциплины

Тема 1. Введение в машинное обучение

Основные понятия машинного обучения. Связь с другими дисциплинами. Контекст машинного обучения как дисциплины - анализ данных и искусственный интеллект. Сферы применения машинного

обучения. Типы задач машинного обучения - обучение с учителем и без учителя. Структура данных для машинного обучения. Инструментальные средства машинного обучения. Понятие модели машинного обучения.

Тема 2. Регрессия

Постановка задачи регрессии. Линейная регрессия с одной переменной - функция гипотезы, функция ошибки, метод градиентного спуска. Регрессия с несколькими переменными - множественная линейная регрессия, нормализация признаков, полиномиальная регрессия. Практическое построение регрессии - загрузка и представление данных, реализация метода градиентного спуска, оценка качества регрессии, подбор скорости обучения, знакомство с библиотекой `sklearn`.

Тема 3. Классификация

Постановка задачи классификации как задачи машинного обучения. Отличия от задачи регрессии. Структура данных для классификации. Логистическая регрессия - функция гипотезы, граница принятия решений, функция ошибки логистической регрессии, градиентный спуск для логистической регрессии, многоклассовая классификация, алгоритм “один против всех”.

Тема 4. Методы обучения с учителем

Универсальность методов обучения с учителем, общая постановка задачи. Линейные модели - линейная и логистическая регрессии - как единая модель. Полиномиальные модели. Метод опорных векторов, ядра - линейное, гауссово, другие. Перцептрон. Деревья решений. К ближайших соседей. Наивная байесовская модель. Достоинства и недостатки разных типов моделей, их сравнительная характеристика, применимость. Применение этих моделей для решения задач классификации и регрессии.

Тема 5. Диагностика систем машинного обучения

Метрики эффективности машинного обучения - сравнение с функциями ошибки. Типичные метрики эффективности для моделей регрессии - MAE, MSE, RMSE, MSLE, MAPE и другие. Метрики эффективности для моделей классификации - accuracy, precision, recall, F1, ROC, PR и другие. Недообучение и переобучение. Проблема bias-variance. Оценка сложности моделей. Обобщающая способность моделей, тестовый набор, кривые обучения. Методы борьбы с недо- и переобучением. Регуляризация. Задача выбора модели - кросс-валидация, гиперпараметры моделей, поиск по сетке, валидационный набор.

Тема 6. Предварительный анализ и обработка данных

Сбор данных для моделей обучения с учителем - реляционная форма данных, понятие чистых данных, оценка источников и объемов данных. Описательный (предварительный) анализ данных (EDA) - анализ репрезентативности, шкалы и типы, визуализация, проблема несбалансированности, обнаружение корреляций, аномалий в данных. Очистка и преобразование данных - удаление лишних признаков, удаление непоказательных объектов, заполнение отсутствующих значений, создание суррогатных признаков, преобразование шкал, воспроизводимость преобразования данных.

Тема 7. Задачи обучения без учителя

Задача кластеризации - постановка задачи, структура датасета, результат и интерпретация. Метод K средних - формализация, гиперпараметры, применимость. Другие методы кластеризации - DBSCAN, иерархическая, агломеративная кластеризация. Задача обнаружения аномалий. Задача понижения размерности - метод главных компонент, метод независимых компонент. Обучение с подкреплением.

Тема 8. Практическое использование моделей машинного обучения

Стохастический и пакетный градиентный спуск. Отбор признаков. Частичное обучение с учителем. Ансамблирование моделей - беггинг, бустинг, стакинг, случайный лес, XGBoost, CatBoost. Конвейеризация моделей машинного обучения. Основные этапы проекта по машинному обучению. Визуализация, интерпретация, представление и анализ результатов машинного обучения. Работа с разными типами данных - преобразование графической информации, методы векторизации текста.

5.2. Учебно-тематический план

ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах»
Очная форма обучения

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваем ости
		Всего	Контактная работа- Аудиторная работа			Самост оательн ая работа	
			Общая, в т.ч.:	Лекции	Семинары, практические занятия		
1	Введение в машинное обучение	36	6	2	4	30	опрос, проверка лаборатор ных работ
2	Регрессия	36	10	2	8	26	опрос, проверка лаборатор ных работ
3	Классификация	36	6	2	4	30	опрос, проверка лаборатор ных работ
4	Методы обучения с учителем	36	16	6	10	20	опрос, проверка лаборатор ных работ
5	Диагностика систем машинного	36	20	8	12	16	опрос, проверка

	обучения						лабораторных работ
6	Предварительный анализ и обработка данных	36	18	6	12	18	опрос, проверка лабораторных работ
7	Задачи обучения без учителя	36	12	4	8	24	опрос, проверка лабораторных работ
8	Практическое использование моделей машинного обучения	36	12	2	10	24	опрос, проверка лабораторных работ
	В целом по дисциплине	288	100	32	68	188	Согласно учебному плану: контрольная работа
	Итого в %		35	32	68	65	

ОП «Прикладные информационные системы в экономике и финансах»
Очно-заочная форма обучения

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваем ости
		Всего	Контактная работа- Аудиторная работа			Самост оательн ая работа	
			Общая, в т.ч.:	Лекции	Семинары, практические занятия		
1	Введение в машинное обучение	36	4	2	2	32	опрос, проверка лаборатор ных работ
2	Регрессия	36	6	2	4	30	опрос, проверка лаборатор ных работ
3	Классификация	36	4	2	2	32	опрос, проверка лаборатор ных работ
4	Методы обучения с	36	12	6	6	24	опрос,

	учителем						проверка лабораторных работ
5	Диагностика систем машинного обучения	36	16	8	8	20	опрос, проверка лабораторных работ
6	Предварительный анализ и обработка данных	36	14	6	8	22	опрос, проверка лабораторных работ
7	Задачи обучения без учителя	36	6	4	2	30	опрос, проверка лабораторных работ
8	Практическое использование моделей машинного обучения	36	6	2	4	30	опрос, проверка лабораторных работ
	В целом по дисциплине	288	68	32	36	220	Согласно учебному плану: контрольная работа
	Итого в %		24	47	53	76	

Институт онлайн-образования, заочная форма обучения

№ п/п	Наименование темы (раздела) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваем ости
		Всего	Контактная работа- Аудиторная работа			Самост оательн ая работа	
			Общая, в т.ч.:	Лекции	Семинары, практические занятия		
1	Введение в машинное обучение	36	3	1	2	33	опрос, проверка лаборатор ных работ
2	Регрессия	36	3	1	2	33	опрос, проверка лаборатор ных работ
3	Классификация	36	3	1	2	33	опрос, проверка

							лабораторных работ
4	Методы обучения с учителем	36	5	1	4	31	опрос, проверка лабораторных работ
5	Диагностика систем машинного обучения	36	5	1	4	31	опрос, проверка лабораторных работ
6	Предварительный анализ и обработка данных	36	5	1	4	31	опрос, проверка лабораторных работ
7	Задачи обучения без учителя	36	3	1	2	33	опрос, проверка лабораторных работ
8	Практическое использование моделей машинного обучения	36	5	1	4	31	опрос, проверка лабораторных работ
	В целом по дисциплине	288	32	8	24	256	Согласно учебному плану: контрольная работа
	Итого в %		11	25	75	89	

5.3. Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарских, практических занятиях, рекомендуемые источники из разделов 8,9 (указывается раздел и порядковый номер источника)	Формы проведения занятий
Введение в машинное обучение	Входной контроль. Изучение технологического стека анализа данных, построенного на базе языка программирования Python. Знакомство с библиотеками numpy, pandas, matplotlib. Выполнение задания по простому статистическому анализу данных инструментальными средствами,	Решение и обсуждение задач

	включающему продвинутое владение соответствующими инструментами. [1-3]	
Регрессия	Построение модели регрессии методом градиентного спуска своими руками. Знакомство с основами обучения параметров модели. Построение модели множественной регрессии своими руками. Построение модели библиотечными средствами. Сравнение результатов. Реализация модели регрессии на примере, близком к реальному (датасет boston). Интерпретация результатов моделирования. [1-3]	Решение и обсуждение задач
Классификация	Реализация модели классификации (логистической регрессии) своими руками наподобие линейной регрессии. Использование библиотечных функций. [1-3]	Решение и обсуждение задач
Методы обучения с учителем	Построение модели классификации на простых данных. Построение нескольких моделей, сравнение их эффективности. Построение модели на сложных данных с нелинейными и слабыми зависимостями. Оценка эффективности, выбор модели. Построение модели регрессии на сложных данных. Построение разных типов регрессоров. Оценка эффективности, выбор модели. Выполнение контрольной работы. [1-3]	Решение и обсуждение задач
Диагностика систем машинного обучения	Оценка эффективности модели регрессии с помощью разных метрик. Оценка эффективности классификации с помощью разных метрик. Построение диагностических кривых (PR, ROC) для бинарной классификации. Диагностика недо-и переобучения в модели классификации. Построение кривых обучения, их интерпретация. Тестовый набор данных Демонстрация необходимости кросс-валидации. Использование кросс-валидированных оценок качества модели. Демонстрация оптимизации гиперпараметров модели, использование валидационного набора для выбора модели. Поиск модели от простых к сложным. [1-3]	Решение и обсуждение задач
Предварительный анализ и обработка данных	Сбор и интеграция данных из разных источников. Основные численные характеристики датасета, анализ на чистоту. Нормализация признаков и заполнение	Решение и обсуждение задач

	<p>отсутствующих значений на искусственном примере.</p> <p>Преобразование категориальных признаков в численные.</p> <p>Краткий алгоритм анализа данных на реальном примере. Интерпретация результатов.</p> <p>Отбор и инжиниринг признаков на примере, приближенном к реальному (датасет titanic).</p> <p>[1-3]</p>	
Задачи обучения без учителя	<p>Решение задачи кластеризации на искусственных данных. Метод К-средних. Метод локтя.</p> <p>Обнаружение аномалий на искусственных данных.</p> <p>Обнаружение аномалий как задача предварительного анализа данных.</p> <p>Понижение размерности на искусственных данных.</p> <p>Интерпретация результата. Понижение размерности для визуализации данных.</p> <p>[1-3]</p>	Решение и обсуждение задач
Практическое использование моделей машинного обучения	<p>Использование простых ансамблей моделей.</p> <p>Оптимизация гиперпараметров моделей.</p> <p>Визуализация, интерпретация и представление результатов машинного обучения исходя из бизнес-задачи моделирования.</p> <p>Построение конвейера машинного обучения.</p> <p>Построение модели классификации текстов.</p> <p>Знакомство с разными методами векторизации текстов.</p> <p>Построение модели классификации изображений.</p> <p>Представление изображений в виде вектора.</p> <p>[1-3]</p>	Решение и обсуждение задач

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
Введение в машинное	Изучение истории машинного обучения как дисциплины, связи с другими областями знания.	Работа с учебной литературой. Разбор вопросов по теме

обучение	Изучение базовых библиотек анализа данных на языке программирования Python - numpy, pandas, matplotlib по официальной документации (при необходимости).	занятия. Выполнение домашних заданий к каждому занятию.
Регрессия	Знакомство с нормированием данных, адаптивной скоростью обучения в ручной реализации метода градиентного спуска. Векторизация вычислений в модели множественной регрессии. Использование множественной регрессии для моделирования полиномиальных признаков и нелинейных зависимостей (корень, экспонента, синус, логарифм, обратная пропорциональность). Визуализация исходных данных и результатов моделирования в реальном случае. Анализ и представление результатов моделирования и совершение предсказания. Статистический анализ достоверности предсказания.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Классификация	Реализация алгоритма “один против всех”. Оценка эффективности классификации. Сравнение двух реализаций.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Методы обучения с учителем	Визуализация результатов классификации. Построение и анализ отчета от классификации. Оценка скорости обучения и работы разных моделей машинного обучения на большом объеме данных. Визуализация результатов классификации, интерпретация полученных результатов. Совершение предсказания и оценка его надежности.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Диагностика систем машинного обучения	Автоматизация построения модели и подсчета выбранных метрик. Построение сводного отчета. Обоснование выбора метрик эффективности исходя из поставленной задачи. Визуализация диагностических кривых для множественной классификации. Использование регуляризованных моделей - ridge, lasso, elastic net. Демонстрация позитивного влияния регуляризации на точность модели. Построение и робастная оценка эффективности модели на данных, близких к реальным. Использование кросс-валидации.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.

	Построение модели и ее последующая оптимизация. Построение ряда простых моделей, диагностика и выбор класса модели, оптимизация и оценка качества.	
Предварительный анализ и обработка данных	Анализ распределения каждого атрибута и связь с целевой переменной. Визуализация. Оценка влияния нормализации признаков на качество обучения. Заполнение отсутствующих значений на реальном примере. Группировка данных и продвинутые алгоритмы рескалирования. Полный алгоритм анализа данных на реальном примере. Анализ влияния на моделирование. Алгоритмический отбор признаков, отбор по итогам оценки важности признаков по итогам построения простых моделей.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Задачи обучения без учителя	Решение задачи кластеризации на реальных данных. Интерпретация результата. Обнаружение аномалий в реальных данных. Интерпретация результатов. Понижение размерности для задачи снижения объема данных. Понижение размерности как инжиниринг признаков.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Практическое использование моделей машинного обучения	Сложные ансамбли моделей - XGBoost, CatBoost. Оптимизация гиперпараметров, выбор модели. Использование различных типов графики, инфографики для оптимального представления результатов моделирования. Знакомство с разными задачами обработки текста - рубрикация, перевод, анализ тональности, генерация. Знакомство с разными задачами анализа изображений - идентификация объекта, обнаружение объекта, распознавание объекта.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.

6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю

Примерные задания для контрольной работы

1. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Описать основные количественные характеристики датасета. Визуализировать распределение каждого признака, входящего в датасет.

Визуализировать совместное распределение каждого признака и целевой переменной. Сделать вывод о линейной разделимости классов.

2. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Построить модель парной классификации каждого признака и целевой переменной. Оценить метрики эффективности четырех моделей классификации. Сделать вывод о значимости каждого признака для классификации.
3. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод деревьев решений. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
4. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод логистической регрессии. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
5. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод опорных векторов. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.

6. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод многослойного перцептрона. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
7. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Построить модель классификации объектов по всем четырем признакам, используя любой метод классификации. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать вывод об улучшении эффективности классификации.
8. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Построить модель классификации объектов по всем четырем признакам, используя любой метод. Оценить эффективность модели при помощи перекрестной проверки (кросс-валидации). Сделать вывод об эффективности классификации.
9. Загрузить встроенный в библиотеку `sklearn` датасет “Ирисы”. Понизить размерность датасета до двух измерений. Визуализировать получившийся набор данных учитывая значение целевой переменной. Сделать вывод о линейной разделимости классов.
10. Загрузить встроенный в библиотеку `sklearn` датасет “Бостон”. Описать основные количественные характеристики датасета. Охарактеризовать шкалу измерения каждого признака (вид шкалы, минимальное, максимальное значение, количество значений). Визуализировать совместное распределение каждого признака и целевой переменной. Сделать вывод о значимости каждого признака для регрессии.

11. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Построить модель множественной регрессии. Оценить метрики эффективности регрессии. Сделать вывод об эффективности получившейся модели.
12. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Построить модель регрессии по всем признакам, используя любой метод регрессии. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать вывод об улучшении эффективности регрессии.
13. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Построить модель регрессии по всем признакам, используя любой метод. Оценить эффективность модели при помощи перекрестной проверки (кросс-валидации). Сделать вывод об эффективности регрессии.
14. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Понизить размерность датасета до двух измерений. Визуализировать получившийся набор данных учитывая значение целевой переменной. Сделать вывод.
15. Загрузить встроенный в библиотеку sklearn датасет “Цифры”. Построить модель классификации объектов, используя метод многослойного перцептрона. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
16. Загрузить встроенный в библиотеку sklearn датасет “Цифры”. Построить модель классификации объектов, используя любой метод классификации. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать вывод об улучшении эффективности классификации.

17. Загрузить встроенный в библиотеку sklearn датасет “Цифры”.

Построить модель классификации объектов, используя любой метод.

Оценить эффективность модели при помощи перекрестной проверки (кросс-валидации). Сделать вывод об эффективности классификации.

18. Загрузить встроенный в библиотеку sklearn датасет “Цифры”.

Понизить размерность датасета до двух измерений. Визуализировать получившийся набор данных учитывая значение целевой переменной.

Сделать вывод о линейной разделимости классов.

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях Департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных.

7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе 2. **«Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине».**

Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний

Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания

<p>ПКН-4 Способность проектировать и создавать интеллектуальные информационные системы, выбирать метод обучения в соответствии с анализом задачи</p>	<p>Демонстрирует знание основных понятий машинного обучения и интеллектуального анализа данных, понимание области и границ применимости, основные виды задач.</p>	<p>Знать: основные понятия машинного обучения и интеллектуальной обработки данных, их области применения и сравнительные характеристики</p> <p>Уметь: применять основные методы обучения с учителем и без учителя для решения типовых и практико-ориентированных задач машинного обучения и анализа данных</p>	<p>Метод градиентного спуска для парной линейной регрессии.</p>
	<p>Демонстрирует знание популярных инструментальных средств машинного обучения, собирает датасет, строит модели, проводит их анализ и диагностику, делает содержательные выводы.</p>	<p>Знать: основные инструментальные средства (языки программирования, библиотеки, фреймворки) для обработки данных и машинного обучения</p> <p>Уметь: применять современные и популярные инструментальные средства для решения практических задач машинного обучения</p>	<p>Загрузить встроенный в библиотеку sklearn датасет "Цифры". Понизить размерность датасета до двух измерений. Визуализировать полученный набор данных учитывая значение целевой переменной. Сделать вывод о линейной разделимости классов</p>
	<p>Презентабельно демонстрирует результаты анализа данных и машинного обучения в форме, доступной непрофессионалу, структурирует отчет по проведенному анализу.</p>	<p>Знать: основные средства и приемы визуализации данных и представления результатов машинного обучения, включая основные метрики и нефункциональные требования к интеллектуальным системам</p> <p>Уметь: представлять результаты проведенного анализа в форме, понятной неспециалисту; визуализировать данные и обнаруженные в них зависимости,</p>	<p>Загрузить встроенный в библиотеку sklearn датасет "Цифры". Понизить размерность датасета до двух измерений. Визуализировать полученный набор данных учитывая значение целевой переменной. Сделать вывод о линейной разделимости классов.</p>

		проводить презентацию проведенного анализа	
--	--	--	--

Примеры практико-ориентированных (ситуационных) заданий

Практическая часть представляет собой задачу по анализу конкретного датасета. Для каждого задания будет необходимо:

1. Загрузить датасет в Python.
2. Описать набор данных и решаемую задачу.
3. Выделить целевую переменную и факторные переменные.
4. Удалить ненужные данные, проанализировать отсутствующие значения.
5. Прокомментировать количественные параметры датасета.
6. Разбить выборку на обучающую и тестовую.
7. Работа по вариантам.

Вариант 1. Очистка данных и обучение моделей.

Данный вариант предполагает фокусировку на обучении нескольких видов моделей обучения с учителем. В зависимости от набора данных, может предполагаться задача классификации и регрессии. Необходимо после минимальной подготовки датасета к обучению обучить несколько моделей и сравнить их эффективность.

Вариант 2. Описательный анализ и визуализация данных.

Данный вариант предполагает фокусировку на исследовании данных и визуализации. При решении этого варианта следует провести как можно более подробный описательный анализ данных с использованием

максимального спектра средств визуализации. При этом следует делать значимые выводы об обнаруженных в данных закономерностях.

Вариант 3. Построение модели и оптимизация гиперпараметров.

Данный вариант предполагает фокусировку на процессе улучшения эффективности модели обучения с учителем. Студенту следует подготовить датасет к обучению, обучить одну из моделей с учителем со значениями гиперпараметров по умолчанию, получить значение эффективности. После этого вручную или автоматически подобрать значения гиперпараметров таким образом, чтобы получить максимальный прирост эффективности.

Вариант 4. Выбор признаков.

Данный вариант предполагает фокусировку на улучшении модели путем ввода новых признаков в модель. Следует подготовить модель к обучению, обучить модель и зафиксировать начальный уровень эффективности. Затем следует исследовать влияние исключения существующих и введения новых признаков в модель на эффективность. Как вариант можно рассматривать введение полиномиальных признаков. Следует стремиться к максимальному увеличению эффективности модели.

Вариант 5. Исследование влияния обучения без учителя на эффективность обучения.

Данный вариант предполагает фокусировку на использовании методов обучения без учителя для ускорения или повышения эффективности обучения с учителем. Следует подготовить модель к обучению, обучить модель и зафиксировать начальный уровень эффективности. Затем следует попробовать применить понижение размерности, обнаружение аномалий или кластеризацию (в любой комбинации) для трансформации исходного

датасета. В конце работы следует сделать значимый вывод об изменении скорости и эффективности обучения с учителем.

Примерные вопросы для подготовки к зачету

1. Понятие машинного обучения. Отличие машинного обучения от других областей программирования.
2. Классификация задач машинного обучения. Примеры задач из различных классов.
3. Основные понятия машинного обучения: набора данных, объекты, признаки, атрибуты, модели, параметры.
4. Структура и представление данных для машинного обучения.
5. Инструментальные средства машинного обучения.
6. Задача регрессии: постановка, математическая формализация.
7. Метод градиентного спуска для парной линейной регрессии.
8. Понятие функции ошибки: требования, использование, примеры.
9. Множественная и нелинейная регрессии.
10. Нормализация признаков в задачах регрессии.
11. Задача классификации: постановка, математическая формализация.
12. Метод градиентного спуска для задач классификации.
13. Логистическая регрессия в задачах классификации.
14. Множественная и многоклассовая классификация. Алгоритм “один против всех”.
15. Метод опорных векторов в задачах классификации.
16. Понятие ядра и виды ядер в методе опорных векторов.
17. Метод решающих деревьев в задачах классификации.
18. Метод k ближайших соседей в задачах классификации.
19. Однослойный перцептрон в задачах классификации.

Примерные вопросы для подготовки к экзамену

1. Метрики эффективности и функции ошибки: назначение, примеры, различия.
2. Понятие набора данных (датасета) в машинном обучении. Требования, представление. Признаки и объекты.
3. Шкалы измерения признаков. Виды шкал, их характеристика.
4. Понятие чистых данных. Определение, очистка данных.
5. Основные этапы проекта по машинному обучению.
6. Предварительный анализ данных: задачи, методы, цели.
7. Проблема отсутствующих данных: причины, исследование, пути решения.
8. Проблема несбалансированных классов: исследование, пути решения.
9. Понятие параметров и гиперпараметров модели. Обучение параметров и гиперпараметров. Поиск по сетке.
10. Понятие недо- и переобучения. Определение, пути решения.
11. Диагностика модели машинного обучения. Методы, цели.
12. Проблема выбора модели машинного обучения. Сравнение моделей.
13. Измерение эффективности работы моделей машинного обучения. Метрики эффективности.
14. Метрики эффективности моделей классификации. Виды, характеристика, выбор.
15. Метрики эффективности моделей регрессии. Виды, характеристика, выбор.
16. Перекрестная проверка (кросс-валидация). Назначение, схема работы.
17. Методы векторизации текстов для задач машинного обучения.
18. Представление графической информации в моделях машинного обучения.
19. Задачи без учителя. Кластеризация. Метод k средних.

20. Задачи без учителя. Обнаружение аномалий.
21. Задачи без учителя. Понижение размерности. Метод PCA.
22. Конвейеры в библиотеке sklearn. Назначение, использование.
23. Использование методов визуализации данных для предварительного анализа.
24. Исследование коррелированности признаков: методы, цели, выводы.
25. Решкалирование данных. Виды, назначение, применение. Нормализация и стандартизация данных.
26. Преобразование категориальных признаков в числовые.
27. Ансамблевые модели машинного обучения. Виды ансамблирования.
28. Конвейеризация моделей машинного обучения.
29. Методы визуализации данных для машинного обучения.
30. Задача выбора модели. Оценка эффективности, валидационный набор.

Пример экзаменационного билета

Экзаменационный билет №

1. Представление графической информации в моделях машинного обучения. (20 баллов)
2. Метрики эффективности и функции ошибки: назначение, примеры, различия. (20 баллов)
3. Вариант 3. Построение модели и оптимизация гиперпараметров.
Датасет: <https://www.kaggle.com/nandvard/microsoft-data-science-capstone>
(20 баллов)

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

Основная литература:

1. Колдаев, В. Д. Структуры и алгоритмы обработки данных : учебное пособие / В. Д. Колдаев. - Москва : РИОР : ИНФРА-М, 2021. - 296 с. - ЭБС ZNANIUM.com. - URL: <https://znanium.com/catalog/product/1230215> (дата обращения: 06.02.2023). – Текст: электронный.

Дополнительная литература:

2. Нагаева, И. А. Основы алгоритмизации и программирования: практикум : учебное пособие / И. А. Нагаева, И. А. Кузнецов. – Москва : Берлин : Директ-Медиа, 2021. – 169 с. – ЭБС Университетская библиотека ONLINE. – URL: <https://biblioclub.ru/index.php?page=book&id=598404> (дата обращения: 06.02.2023). – Текст : электронный.

3. Златопольский, Д. М. Программирование: типовые задачи, алгоритмы, методы: учебное пособие / Д. М. Златопольский. — 4-е изд. (эл.). — Москва : Лаборатория знаний, 2020. — 226 с.: ил. — ЭБС Университетская библиотека ONLINE. — URL: <https://biblioclub.ru/index.php?page=book&id=222873> (дата обращения 06.02.2023). – Текст: электронный.

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/>

2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>

3. Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>

4. Электронно-библиотечная система Znanium <http://www.znanium.com>

5. Электронно-библиотечная система издательства «ЮРАЙТ»
<https://urait.ru/>
6. Электронно-библиотечная система издательства Проспект
<http://ebs.prospekt.org/books>
7. Электронно-библиотечная система издательства Лань
<https://e.lanbook.com/>
8. Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>
9. Электронная библиотека Издательского дома «Гребенников»
<https://grebennikon.ru/>
10. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>
11. Национальная электронная библиотека <http://нэб.рф/>
12. Финансовая справочная система «Финансовый директор»
<http://www.1fd.ru/>
13. Pyru 1.0.9 [Электронный ресурс]: сайт. – Режим доступа:
<https://pypi.python.org/pypi/pyru>
14. Python Data Analysis Library [Электронный ресурс]: сайт. – Режим
доступа: <http://pandas.pydata.org/>
15. Python Documentation [Электронный ресурс]: сайт. – Режим
доступа: <http://python.org/doc/>
16. Python Standard Library [Электронный ресурс]: сайт. – Режим
доступа: <https://docs.python.org/2/library/>
17. Scikit-learn Machine Learning in Python [Электронный ресурс]:
сайт. – Режим доступа: <http://scikit-learn.org>
18. The Python Tutorial // <https://docs.python.org/3/tutorial/index.html>
19. NumPy User Guide // <http://docs.scipy.org/doc/numpy/user/index.html>
20. Pandas User Guide <http://pandas.pydata.org/pandas-docs/stable/>

10. Методические указания для обучающихся по освоению дисциплины

При изучении теоретического материала необходимо опираться на рабочую программу дисциплины, материалы лекций и литературу из основного списка. Кроме этого, необходимо активно работать с Интернет-источниками и пособиями других авторов, помогающими усвоить материал отдельных разделов программы.

Необходимо конспектировать лекции, пометая сложные и непонятные моменты с тем, чтобы задать вопросы лектору в конце лекции или же на консультации.

При подготовке к семинарским занятиям необходимо изучить вопросы, вынесенные на самостоятельное изучение, так как семинарские занятия предполагают их обсуждение и дискуссию по теме; кроме того, задания для самостоятельной работы необходимы для того, чтобы успешно выполнить самостоятельные задания на семинарах.

Индивидуальные задания для работы на компьютере, файлы с выполненными заданиями необходимо хранить в личной сетевой папке в компьютерной сети вуза.

Для выполнения кейса задачи, рекомендуется использовать язык программирования Python и специализированные библиотеки для анализа данных и машинного обучения.

Студент должен прислать рабочий notebook с решением кейса задачи, комментариями кода и аналитическими выводами до конца экзамена на электронную почту преподавателя.

Критерии оценки практической экзаменационной работы:

1. Структурированность отчета. В работе должна прослеживаться четкая структура - подготовительный этап, анализ данных,

построение простых моделей, сравнение и анализ моделей, выводы, построение моделей с учетом выводов, итоговый результат.

2. Наличие выводов. Работа должна содержать текстовые замечания, поясняющие каждый шаг работы студента: что делается, зачем и какую информацию это нам дает. Оценивается полнота и адекватность выводов.
3. Визуализация. Работа должна демонстрировать навыки студента визуализировать информацию. Особенно на этапах описательного анализа и анализа обучаемости модели. Оценивается разнообразие, наглядность и информативность визуализации.
4. Использование метрик эффективности. Оценивается разнообразие и адекватность задаче примененных метрик эффективности (включая время обучения) а также полнота сравнения и правильность выводов из сравнения моделей по разным метрикам.
5. Валидность результатов. Студент должен продемонстрировать умение оценивать достоверность измерения метрик моделей и повышать ее с использованием перекрестной проверки (кросс-валидации). Использование k-fold cross validation является предпочтительным методом измерения эффективности модели. Если происходит выбор модели, то ее итоговая эффективность должна измеряться на чистом наборе данных.

Список датасетов, использующихся на экзамене:

1. <https://www.kaggle.com/uciml/mushroom-classification>
2. <https://www.kaggle.com/lodetomasi1995/income-classification>
3. <https://www.kaggle.com/uciml/glass>
4. <https://www.kaggle.com/uciml/german-credit>
5. <https://www.kaggle.com/zaurbegiev/my-dataset>
6. <https://www.kaggle.com/kaushiksuresh147/customer-segmentation>

7. <https://www.kaggle.com/deepu1109/star-dataset>
8. <https://www.kaggle.com/vinesmsuic/star-categorization-giants-and-dwarfs>
9. <https://www.kaggle.com/shebrahimi/financial-distress>
10. <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>
11. <https://www.kaggle.com/amir75/caesarean-section-classification>
12. <https://www.kaggle.com/sachinsharma1123/performance-prediction>
13. <https://www.kaggle.com/ninzaami/loan-predication>
14. <https://www.kaggle.com/caparrini/beatsdataset>
15. <https://www.kaggle.com/zhiruo19/covid19-symptoms-classification>
16. <https://www.kaggle.com/kunalvsingh93/banking-modelclassification>
17. <https://www.kaggle.com/mansoordaku/ckdisease>
18. <https://www.kaggle.com/mnassrib/telecom-churn-datasets>
19. <https://www.kaggle.com/akshayksingh/kidney-disease-dataset>
20. <https://www.kaggle.com/henriqueyamahata/bank-marketing>
21. <https://www.kaggle.com/maajdl/yeh-concret-data>
22. <https://www.kaggle.com/hellbuoy/car-price-prediction>
23. <https://www.kaggle.com/rhuebner/human-resources-data-set>
24. <https://www.kaggle.com/loveall/appliances-energy-prediction>
25. <https://www.kaggle.com/elikplim/forest-fires-data-set>
26. <https://www.kaggle.com/nandvard/microsoft-data-science-capstone>
27. <https://www.kaggle.com/shebrahimi/financial-distress>
28. <https://www.kaggle.com/aungpyaeap/beauty>
29. <https://www.kaggle.com/vbmokin/ammonium-prediction-in-river-water>
30. <https://www.kaggle.com/veer06b/marrket-mix-dataset>

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

11.1. Комплект лицензионного программного обеспечения:

1. Антивирус;
2. Браузер.
3. Дистрибутив языка Python 3.6 (или более поздней версии)
Anaconda
4. Облачная среда разработки Google Colab или аналогичная
5. Среда разработки Jupyter Notebook

11.2. Современные профессиональные базы данных и информационные справочные системы

- не используются

11.3. Сертифицированные программные и аппаратные средства защиты информации

- не используются

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

Для проведения лекций и практических занятий необходима аудитория, оснащенная проектором и компьютерами с постоянным подключением к сети Интернет.