Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования «Новосибирский национальный исследовательский
государственный университет» (Новосибирский государственный университет, НГУ)

Факультет информационных технологий

СОГЛАСОВАНО
Декан ФИТ НГУ
_____ М.М. Лаврентьев
«25» апреля 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ данных и машинное обучение

Направление подготовки: 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
Направленность (профиль): Искусственный интеллект и Data Science
Форма обучения: очная
Год обучения: 1, семестр: 2

№	Вид деятельности	Семестр
		2
1	Лекции, час.	32
2	Практические занятия, час.	32
3	Лабораторные занятия, час.	
4	Занятий в контактной форме без учета промежуточной аттестации, час, из них	66
5	в электронной форме, час.	
6	из них аудиторных занятий, час.	64
7	из них в активной и интерактивной форме, час.	64
8	консультаций, час.	2
9	Самостоятельная работа, час.	76
10	в том числе на выполнение письменных работ, час	42
11	Форма аттестации (экзамен, зачет, дифференцированный зачет), час	Э, 2
12	Всего зачетных единиц[1]	4

Новосибирск 2023

[1] С учетом выделенных часов на промежуточную аттестацию

Рабочая программа дисциплины составлена на основании федерального государственного образовательного стандарта (ФГОС) высшего образования - магистратура по направлению подготовки 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА.

Федеральный государственный образовательный стандарт (ФГОС) высшего образования - магистратура по направлению подготовки 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА введен в действие приказом Минобрнауки от 19.09.2017 № 918.

Место дисциплины в структуре учебного плана: Блок 1 Дисциплины (модули); обязательная часть, обязательная дисциплина.

Рабочая программа дисциплины утверждена решением Ученого совета факультета информационных технологий от 24.04.2023, протокол №91.

Программу разработал:

Доцент кафедры систем информатики ФИТ,
кандидат физико-математических наук Д.С. Мигинский

Заведующий кафедрой систем информатики ФИТ,
доктор физико-математических наук М.М. Лаврентьев

Ответственный за образовательную программу:

Заведующий кафедрой систем информатики ФИТ,
доктор физико-математических наук М.М. Лаврентьев

Аннотация к рабочей программе дисциплины
«Анализ данных и машинное обучение»

Дисциплина «**Анализ данных и машинное обучение**» реализуется в рамках образовательной программы высшего образования – программы магистратуры 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА, направленность (профиль): ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И DATA SCIENCE по очной форме обучения на русском языке.

Место в образовательной программе: Дисциплина «**Анализ данных и машинное обучение**» развивает знания, умения и навыки, сформированные у обучающихся по результатам изучения следующих дисциплин: Методологии анализа данных, является базовой для освоения дисциплины «Нейросети и машинное обучение».

Дисциплина «**Анализ данных и машинное обучение**» реализуется во 2 семестре в рамках обязательной части дисциплин (модулей) Блока 1 и является обязательной дисциплиной.

Дисциплина «**Анализ данных и машинное обучение**» направлена на формирование компетенций:

Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте (ОПК-1), в части следующих индикаторов достижения компетенции:

ОПК-1.1. Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности

ОПК-1.2. Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний

ОПК-1.3. Владеть: навыками теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте

Способен применять на практике новые научные принципы и методы исследований (ОПК-4), в части следующих индикаторов достижения компетенции:

ОПК-4.1. Знать: новые научные принципы и методы исследований

ОПК-4.2. Умеет: применять на практике новые научные принципы и методы исследований

ОПК-4.3. Владеть: навыками применения новых научных принципов и методов исследования для решения профессиональных задач

Перечень основных разделов дисциплины:

В рамках данного курса студенты освоят основы интеллектуального анализа данных, включая преобразование и очистку данных, работу с пропущенными значениями, основные способы визуализации данных (гистограммами, диаграммами плотности, диаграммами рассеяния, ящиками с усами и т.п.), корреляционный анализ. Освоят различные методы отбора признаков. Научатся решать различные задачи снижения размерности данных, кластеризации, классификации, регрессии. Студенты освоят работу со специализированными программными библиотеками для визуализации и анализа

данных и научатся применять полученные знания для решения практических задач, в том числе, загружать данные, сохраненные в разных форматах, выбирать и группировать нужные записи по заданным критериям, строить предсказательные модели и оценивать их качество.

Дисциплина «Анализ данных и машинное обучение» предусматривает проведение лекций и практических занятий (семинаров) в интерактивной форме. Студенты выполняют ряд заданий, входящих в рамки портфолио.

Самостоятельная работа включает: подготовку к практическим занятиям по разделам дисциплины, решение заданий, подготовку к экзамену.

Предусмотрено проведение занятий с использованием дистанционных образовательных технологий.

Общий объем дисциплины – 4 зачетных единиц (144 часа).

Правила аттестации по дисциплине. Текущий контроль работы в семестре осуществляется в форме портфолио (выполнение заданий). Всего предусмотрено 5 заданий. Задания выкладываются на странице курса и в группе курса.

Задания нацелены на практическое применение изученных на занятиях методов и алгоритмов. Выполненные задания сдаются в электронном виде. На решение заданий отводится не менее 2 недель. За сдачу задания после 21 дня с даты получения итоговая оценка уменьшается на 10 %. В каждом задании есть теоретическая и практическая часть.

Промежуточная аттестация по дисциплине проводится в виде экзамена. Оценка выставляется на основе суммы баллов за портфолио (выполненные задания)

Суммарное значение баллов, составляющее не менее 85 % от максимального, соответствует оценке «отлично», 70 % – «хорошо», 55 % – «удовлетворительно». Оценки «отлично», «хорошо», «удовлетворительно» соответствуют успешному прохождению промежуточной аттестации.

Учебно-методическое обеспечение дисциплины.

Воронцов К.В. Машинное обучение Школа Анализа данных Яндекс. МФТИ. национальный открытый университет Интуит, 2015 – Режим доступа: свободный – URL: https://www.intuit.ru/studies/courses/13844/1241/info

1. Внешние требования к дисциплине

Таблица 1.1

Компетенция ОПК-1. Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте, *в части следующих индикаторов достижения компетенции:*
ОПК-1.1. Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности
ОПК-1.2. Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний
ОПК-1.3. Владеть: навыками теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте
Компетенция ОПК-4. Способен применять на практике новые научные принципы и методы исследований, *в части следующих индикаторов достижения компетенции:*
ОПК-4.1. Знать: новые научные принципы и методы исследований
ОПК-4.2. Умеет: применять на практике новые научные принципы и методы исследований
ОПК-4.3. Владеть: навыками применения новых научных принципов и методов исследования для решения профессиональных задач

2. Требования к результатам освоения дисциплины

Таблица 2.1

Результаты изучения дисциплины по уровням освоения (иметь представление, знать, уметь, владеть)	**Формы организации занятий**		
	Лекции	**семинары**	**Самостоятельная работа**
ОПК-1.1.Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности			
1. Знать основы интеллектуального анализа данных	+	+	+
2. Знать основные способы визуализации данных (гистограммами, диаграммами плотности, диаграммами рассеяния, ящиками с усами и т.п.), реализованные в библиотеках matplotlib, seabron.	+	+	+
3. Знать методы понижения размерности данных, реализованные в библиотеке sklearn.	+	+	+
ОПК-1.2.Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний			
4. Уметь проводить разведочный анализ данных, проводить предобработку и очистку данных, работать с пропущенными значениями.		+	+
5. Уметь визуализировать данные, в том числе, с использованием методов снижения размерности.		+	+
6. Уметь обоснованно выбирать наиболее подходящие алгоритмы решения задач машинного обучения и оценивать качество построенных моделей		+	+

ОПК-1.3. Владеть: навыками теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте			
7. Уверенно владеть базовыми инструментами анализа данных и решения задач машинного обучения, реализованными в библиотеках pandas и sklearn.		+	+
ОПК-4.1.Знать: новые научные принципы и методы исследований			
8. Знать особенности работы со специализированными программными библиотеками языка программирования Python для анализа данных и решения задач машинного обучения.	+	+	+
ОПК-4.2.Умеет: применять на практике новые научные принципы и методы исследований			
9. Уметь составлять конвейеры для предобработки данных, построения и подбора оптимальных гиперпараметров моделей.		+	+
ОПК-4.3. Владеть: навыками применения новых научных принципов и методов исследования для решения профессиональных задач			
10. Уметь составлять композиции моделей (блендинг, стеккинг), проводить отбор признаков.		+	+

3. Содержание и структура учебной дисциплины

Таблица 3.1

Темы лекций	Активные формы, час.	Часы	Ссылки на результаты обучения
Семестр: 2			
1. Введение в предметную область. Примеры использования методов машинного обучения для решения прикладных задач. Повторение основ программирования на языке Python.	4	4	1, 2, 3, 8
2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	4	4	1, 2, 3, 8
3. Знакомство с различными методами предобработки данных, описательными статистиками и основными способами визуализации данных, методами снижения размерности. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.	4	4	1, 2, 3, 8
4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	2	2	1, 2, 3, 8
5. Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация. Метод K-средних, DBSCAN и др. Обзор методов кластеризации, реализованных в библиотеке sklearn.	2	2	1, 2, 3, 8
6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые. Нормировка данных. Определение переобученности модели. Критерии оценки качества полученных моделей.	4	4	1, 2, 3, 8
7. Постановка задачи регрессии. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные	4	4	1, 2, 3, 8

наблюдения, анализ остатков. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.			
8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация. Логистическая регрессия. Решающие деревья. Метрики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой).	4	4	1, 2, 3, 8
9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.	4	4	1, 2, 3, 8
Итого	32	32	

Таблица 3.2

Темы практических занятий	Активные формы, час.	Часы	Ссылки на результаты обучения	Учебная деятельность
Семестр: 2				
1. Введение в предметную область. Примеры использования методов машинного обучения для решения прикладных задач. Повторение основ программирования на языке Python.	4	4	1, 8	Примеры использования методов машинного обучения для практических задач. Краткий обзор синтаксиса языка Python. Встроенные операции и функции, типы и структуры данных.
2. Знакомство со специализированными библиотеками языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.	4	4	1, 8	Библиотеки NumPy и SciPy. Матрицы. Разреженные матрицы. Индексирование, срезы. Объединение массивов. Библиотека pandas. Запросы к таблицам: выборка строк/столбцов по заданным критериям. Модификация элементов таблицы. Добавление строк/столбцов. Группировка и агрегирование. Объединение таблиц (различные виды join). Многомерные данные: мультииндексы. Операции stack-unstack. Построение сводных таблиц (pivot tables).
3. Знакомство с различными методами предобработки данных, описательными статистиками и основными способами визуализации данных, методами снижения размерности.	4	4	1-5	Описательные статистики. Обзор библиотек matplotlib, seaborn, bokeh. Базовые типы визуализации данных. Знакомство с библиотекой scikit-learn (sklearn). Предобработка данных.

	Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.				Метод главных компонент. Работа с пропущенными значениями.
4.	Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.	2	2	1-8	Дальнейшее знакомство студентов с пакетом sklearn. Основные функции. Работа с данными из набора MNIST (рукописные цифры). Работа с синтетическими данными.
5.	Обучение на неразмеченных данных. Нормировка данных. Кластеризация. Иерархическая кластеризация. Метод K-средних, DBSCAN и др. Обзор методов кластеризации, реализованных в библиотеке sklearn.	2	2	1-8	Использование методов снижения размерности и методов кластеризации в задаче распознавания рукописных цифр (MNIST). Работа с синтетическими данными.
6.	Задачи обучения с учителем. Разделение данных на обучающие и тестовые. Определение переобученности модели. Критерии оценки качества полученных моделей.	4	4	1-8	Примеры задач обучения с учителем. Важность определения целевой метрики качества. Сравнение различных метрик качества моделей. Работа с несбалансированными наборами.
7.	Постановка задачи регрессии. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные наблюдения, анализ остатков. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.	4	4	1-8	Объединение алгоритмов, реализованных в sklearn, в цепочки и конвейеры с помощью класса Pipeline. Реализация регрессионных и классификационных моделей с помощью sklearn. Работа с синтетическими данными. Самостоятельная реализация метода градиентного спуска.
8.	Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация. Логистическая регрессия. Решающие деревья. Метрики качества	4	4	1-8	Реализация классификационных моделей с помощью sklearn. Реализация моделей на основе метода k-ближайших соседей. Метод логистической регрессии. Самостоятельная реализация метода градиентного спуска.

	классификации (точность/специфичность, ROC-кривая, площадь под кривой).			Реализация решающего дерева.
9.	Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.	4	4	8-10	Реализация моделей с помощью метода градиентного бустинга, метода случайного леса. Блендинг и стеккинг. Методы отбора признаков. Оптимизация гиперпараметров.
Итого		32	32		

4. Самостоятельная работа студентов

Таблица 4.1

№	Виды самостоятельной работы	Ссылки на результаты обучения	Часы на выполнение	Часы на консультации
Семестр: 3				
1	Самостоятельная работа с учебными материалами, разбор тем, изученных на лекциях и практических занятиях, разбор решенных заданий.	1-10	10	0
	Обучающиеся изучают источники из списка основной и дополнительной литературы. Предполагается самостоятельное изучение ресурсов по предметной области курса в сети интернет. Методические рекомендации по подготовке к практическим занятиям представлены в приложении к рабочей программе дисциплины.			
2	Выполнение заданий в рамках портфолио	1-10	32	0
	Обучающиеся решают практические задачи, входящие в портфолио. Методические рекомендации по выполнению домашнего задания представлены в приложении к рабочей программе дисциплины.			
3	Выполнение и защита итогового задания.	1-10	10	0
	Обучающиеся выбирают темы итогового задания – самостоятельно (обязательно согласование с преподавателем) или из списка предложенных тем. Используют полученные знания для разведочного анализа данных, проводят выбор способа предобработки данных, выбор способа решения поставленной задачи, проводят оптимизацию гиперпараметров. Методические рекомендации по самостоятельному изучению теоретического материала представлены в приложении к рабочей программе дисциплины.			
4	Подготовка к экзамену	1,2,3,4,5,7	24	2
	Подготовка к экзамену по вопросам, представленным в фонде оценочных средств.			
	Итого		76	2

5. Образовательные технологии

В ходе реализации учебного процесса по дисциплине проводятся лекционные и практические занятия. Темы, рассматриваемые на лекциях и изучаемые самостоятельно, закрепляются на практических занятиях, по вопросам, вызывающим затруднения, проводятся консультации.

Предусмотрено проведение занятий с использованием дистанционных образовательных технологий. При проведении практических занятий студенты подключаются к онлайн-сессии. На занятии разбираются теоретические темы и формулировки практических заданий. Для сдачи выполненного задания студент включает демонстрацию экрана, показывает результаты, обосновывает решение, отвечает на вопросы преподавателя

В ходе реализации учебного процесса по дисциплине также применяются следующие интерактивные формы обучения (таблица 5.1).

Таблица 5.1

1	Лекция в форме дискуссии	ОПК-1.1, 1.2, 1.3, ОПК-4.1, 4.2, 4.3
Формируемые умения: При решении практических и теоретических задач из домашних заданий студенты закрепляют знания и навыки, полученные в ходе лекций и семинарских занятий. Учатся использовать инструменты анализа данных и решения задач машинного обучения, реализованные в библиотеках pandas и sklearn. Учатся проводить разведочный анализ данных, предобработку и очистку данных. Учатся визуализировать данные, в том числе, с использованием методов снижения размерности. Учатся обоснованно выбирать наиболее подходящие алгоритмы решения задач машинного обучения и оценивать качество построенных моделей.		
Краткое описание применения: Представляется теория, проблематика вопросов, связанных с изучаемой предметной областью		
2	Портфолио	ОПК-1.1, 1.2, 1.3, ОПК-4.1, 4.2, 4.3
Формируемые умения: Студенты с помощью полученных знаний и навыков выполняют задания по анализу данных и построению классификационных или регрессионных моделей (в зависимости от поставленной задачи), начиная с разведочного анализа данных, их предобработки и очистки до построения итоговых моделей и оценки их качества		
Краткое описание применения: студенты ведут портфолио (оценки за задания), которое является основой для проведения аттестации по дисциплине		

Для организации и контроля самостоятельной работы студентов, а также проведения консультаций применяются информационно-коммуникационные технологии. (таблица 5.2).

Таблица 5.2

Информирование	antonec@yandex.ru, тема: ММО.информ, группа ВКонтакте
Консультирование	antonec@yandex.ru, тема: ММО.информ, группа ВКонтакте
Контроль	antonec@yandex.ru, тема: ММО.задачи, группа ВКонтакте
Размещение учебных материалов	группа ВКонтакте https://www.intuit.ru/studies/courses/13844/1241/info

6. Правила аттестации студентов по учебной дисциплине

По дисциплине «Анализ данных и машинное обучение» проводится текущая и промежуточная аттестация (итоговая по дисциплине).

Текущая аттестация по дисциплине осуществляется в форме портфолио (выполнение заданий). Всего предусмотрено 5 заданий, последнее из заданий - итоговое. Задания выкладываются на странице курса и в группе курса.

Задания нацелены на практическое применение изученных на занятиях методов и алгоритмов. Выполненные задания сдаются в электронном виде. На решение заданий отводится не менее 2 недель. За сдачу задания после 21 дня с даты получения итоговая оценка уменьшается на 10 %. В каждом задании есть теоретическая и практическая часть.

Итоговое задание представляет решение выбранной и согласованной с преподавателем задачи из предметной области курса (анализ данных и машинное обучение). Задание выполняется с использованием языка программирования Python и специализированных библиотек, обязательно должно присутствовать описание использованных данных, указание источника, загрузку данных, разведочный анализ, визуализацию описательных статистик, исследование с помощью методов снижения размерности, описание постановки задачи, построение классификационной или регрессионной модели (в зависимости от типа задачи), обоснование выбора метода и анализ качества полученной модели. Работа должна быть выполнена методологически корректно, без грубых ошибок. Программный код должен исполняться без фатальных ошибок. Студент должен понимать суть выполненной работы и быть готов дать пояснения по программному коду.

Итоговое задание сдается либо в виде файла *.ipynb – iPython Notebook, либо в виде презентации с приложением программного кода. Для решения задания допускается использовать язык программирования R.

Необходимым условием для прохождения промежуточной аттестации является оценка «зачтено» за портфолио. Оценка «зачтено» за портфолио выставляется при условии выполнения и защиты всех заданий, входящих в портфолио.

Промежуточная аттестация по дисциплине проводится в виде экзамена. Оценки «отлично», «хорошо», «удовлетворительно» соответствуют успешному прохождению промежуточной аттестации.

В таблице 6.1 представлено соответствие форм аттестации заявляемым требованиям к результатам освоения дисциплины.

Таблица 6.1

Коды компетенций ФГОС	**Результаты обучения**	**Формы аттестации**	
		1 этап - портфолио	**2 этап - экзамен**
ОПК-1	ОПК-1.1. Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности	+	+
ОПК-1	ОПК-1.2. Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний	+	+
ОПК-1	ОПК-1.3. Владеть: навыками теоретического и экспериментального исследования объектов	+	

	профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте		
ОПК-4	ОПК-4.1. Знать: новые научные принципы и методы исследований	+	+
ОПК-4	ОПК-4.2. Умеет: применять на практике новые научные принципы и методы исследований	+	+
ОПК-4	ОПК-4.3. Владеть: навыками применения новых научных принципов и методов исследования для решения профессиональных задач	+	

Требования к структуре и содержанию портфолио, оценочные средства, а также критерии оценки сформированности компетенций и освоения дисциплины в целом, представлены в Фонде оценочных средств, являющемся приложением 1 к настоящей рабочей программе дисциплины.

7. Литература

1. Буйначев, С.К. Основы программирования на языке Python : учебное пособие / С.К. Буйначев, Н.Ю. Боклаг ; Министерство образования и науки Российской Федерации, Уральский федеральный университет им. первого Президента России Б. Н. Ельцина. – Екатеринбург : Издательство Уральского университета, 2014. – 92 с. : табл., ил. – Режим доступа: по подписке. – URL: http://biblioclub.ru/index.php?page=book&id=275962– Библиогр. в кн. – ISBN 978-5-7996-1198-9. – Текст : электронный.

Интернет-ресурсы

Таблица 7.1

№ п/п	Наименование Интернет-ресурса	Краткое описание
1	http://www.machinelearning.ru/	Большая коллекция материалов по машинному обучению на русском языке.
2	http://anaconda.org	Дистрибутив Python с большинством необходимых библиотек.
3	http://scipy.org/	Библиотека для научных вычислений для языка программирования Python.
4	http://pandas.pydata.org/	Библиотека для анализа данных pandas.
5	http://scikit-learn.org/stable/user_guide.html	Документация библиотеки sklearn.
6	http://scikit-learn.org/stable/tutorial/index.html	Примеры решения некоторых задач.
7	http://kaggle.com	Платформа для проведения конкурсов по решению задач машинного обучения. Содержит обучающие ресурсы с примерами решений задач и их обсуждением.
8	http://archive.ics.uci.edu/ml/	Коллекция данных и задач.
9	https://stepik.org/course/67	Курс «Программирование на Python» по основам программирования на языке Python.
10	https://www.coursera.org/learn/machine-learning	Курс по основам машинного обучения от Эндрю ЫНа (Andrew Ng). Преподается на английском языке.

11	https://ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie	Курс по основам машинного обучения с использованием Python + pandas + sklearn. Преподается на русском языке. Преподаватель: Константин Вячеславович Воронцов.

8. Учебно-методическое и программное обеспечение дисциплины

8.1. Учебно-методическое обеспечение

Воронцов К.В. Машинное обучение Школа Анализа данных Яндекс. МФТИ. национальный открытый университет Интуит, 2015 – Режим доступа: свободный – URL: https://www.intuit.ru/studies/courses/13844/1241/info

8.2. Программное обеспечение

Для обеспечения реализации дисциплины используется стандартный комплект программного обеспечения (ПО), включающий регулярно обновляемое лицензионное ПО Windows и MS Office.

Перечень специализированного программного обеспечения для изучения дисциплины представлен в таблице 8.1.

Специализированное программное обеспечение

Таблица 8.1

№	Наименование ПО	Назначение
1	Python 3.7.1 (Anaconda3 2018.12 64-bit)	Среда разработки приложений
2	Notepad++	Программа для работы с текстовыми файлами

9. Профессиональные базы данных и информационные справочные системы

1. Полнотекстовые журналы Springer Journals за 1997-2015 г., электронные книги (2005-2016 гг.), реферативная БД по чистой и прикладной математике zbMATH.
2. Электронные ресурсы Web of Science Core Collection (Thomson Reuters Scientific LLC.), Journal Citation Reports + ESI
3. БД Scopus (Elsevier)

10. Материально-техническое обеспечение

Таблица 10.1

№	Наименование	Назначение
1	Презентационное оборудование (мультимедиа-проектор, экран, компьютер для управления)	Для проведения лекционных занятий
2	Компьютерный класс (с выходом в Internet)	Для организации самостоятельной работы и проведения практических занятий обучающихся

Материально-техническое обеспечение образовательного процесса по дисциплине для обучающихся из числа лиц с ограниченными возможностями здоровья осуществляется согласно «Порядку организации и осуществления образовательной деятельности по образовательным программам для инвалидов и лиц с ограниченными возможностями здоровья в Новосибирском государственном университете».

Лист актуализации рабочей программы дисциплины
«Анализ данных и машинное обучение»

№	Характеристика внесенных изменений (с указанием пунктов документа)	Дата и № протокола Ученого совета ФИТ	Подпись ответственного
1.	Внесены изменения в п.5 в части проведения занятий с использованием дистанционных образовательных технологий	05.02.2024 №94	

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования «Новосибирский национальный исследовательский
государственный университет» (Новосибирский государственный университет, НГУ)

Факультет информационных технологий

СОГЛАСОВАНО
Декан ФИТ НГУ
_____ М.М. Лаврентьев
«25» апреля 2023 г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ
по дисциплине Анализ данных и машинное обучение

Направление подготовки: 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Направленность (профиль): Искусственный интеллект и Data Science

Квалификация: магистр

Форма обучения: очная Год обучения: 1, семестр 2

Форма аттестации	Семестр
Экзамен	2

Новосибирск 2023

Фонд оценочных средств промежуточной аттестации по дисциплине является **Приложением 1** к рабочей программе дисциплины «Анализ данных и машинное обучение», реализуемой в рамках образовательной программы высшего образования – программы магистратуры 09.04.01 Информатика и вычислительная техника, направленность (профиль): Искусственный интеллект и Data Science.

Фонд оценочных средств промежуточной аттестации по дисциплине утвержден решением Ученого совета факультета информационных технологий, протокол №91 от 24.04.2023.

Разработчики:
Доцент кафедры систем информатики ФИТ,
кандидат физико-математических наук Д.С.Мигинский

Заведующий кафедрой систем информатики ФИТ,
доктор физико-математических наук М.М. Лаврентьев

Ответственный за образовательную программу:

Заведующий кафедрой систем информатики ФИТ,
доктор физико-математических наук М.М. Лаврентьев

Содержание и порядок проведения промежуточной аттестации по дисциплине

1.1. Общая характеристика содержания промежуточной аттестации

Промежуточная аттестация по дисциплине «Анализ данных и машинное обучение» проводится по завершению периода освоения образовательной программы (семестра) для оценки сформированности компетенций в части следующих индикаторов достижения компетенции (таблица П1.1).

Таблица П1.1

Код	Компетенции, формируемые в рамках дисциплины «Нейросети и машинное обучение»	2 семестр	
		1 этап - портфолио	2 этап – экзамен
ОПК-1 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте			
ОПК-1	ОПК-1.1. Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности	+	+
ОПК-1	ОПК-1.2. Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математических, естественнонаучных, социально-экономических и профессиональных знаний	+	+
ОПК-1	ОПК-1.3. Владеть: навыками теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте	+	
ОПК-4 Способен применять на практике новые научные принципы и методы исследований			
ОПК-4	ОПК-4.1. Знать: новые научные принципы и методы исследований	+	+
ОПК-4	ОПК-4.2. Умеет: применять на практике новые научные принципы и методы исследований	+	+
ОПК-4	ОПК-4.3. Владеть: навыками применения новых научных принципов и методов исследования для решения профессиональных задач	+	

Промежуточная аттестация по дисциплине включает 2 этапа: портфолио и экзамен. Тематика вопросов к экзамену включает следующие темы (разделы):

1. Специализированные библиотеки языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.
3. Методы предобработки данных, описательные статистикеи и основные способы визуализации данных, методы снижения размерности. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.
4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.

5. Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация. Метод K-средних, DBSCAN и др. Обзор методов кластеризации, реализованных в библиотеке sklearn.

6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые. Нормировка данных. Определение переобученности модели. Критерии оценки качества полученных моделей.

7. Постановка задачи регрессии. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные наблюдения, анализ остатков. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.

8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация. Логистическая регрессия. Решающие деревья. Метрики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой).

9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.

1.2. Порядок проведения промежуточной аттестации по дисциплине

Промежуточная аттестация проводится в форме экзамена и включает 2 этапа: портфолио и экзамен. Необходимым условием для прохождения промежуточной аттестации является оценка «зачтено» за портфолио. Оценка «зачтено» за портфолио выставляется при условии выполнения и защиты всех заданий, входящих в портфолио.

Экзамен проводится в устной форме, по билетам. Билет выбирается обучающимся случайным образом. При подготовке ответа на вопросы билета не разрешается использование каких-либо источников информации. В процессе ответа обучающегося на вопросы билета преподаватель может задавать дополнительные вопросы по темам дисциплины. Результаты промежуточной аттестации определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». Оценки «отлично», «хорошо», «удовлетворительно» означают успешное прохождение промежуточной аттестации.

2. Требования к структуре и содержанию фонда оценочных средств промежуточной аттестации по дисциплине

Перечень оценочных средств, применяемых на каждом этапе проведения промежуточной аттестации по дисциплине, представлен в таблице П1.2.

Таблица П1.2

№ п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в фонде
Этап 1 - портфолио			
1	Портфолио	Целевая подборка работ студента, раскрывающая его индивидуальные образовательные достижения в одной или нескольких учебных дисциплинах.	Структура портфолио
Этап 2 - экзамен			
2	Билет	Комплекс вопросов	Список теоретических вопросов

2.1 Требования к структуре и содержанию оценочных средств аттестации

2.1.1 Требования к структуре и содержанию портфолио

Портфолио по дисциплине включает 5 заданий, последнее из заданий - итоговое. Задания выкладываются на странице курса и в группе курса.

Задания нацелены на практическое применение изученных на занятиях методов и алгоритмов. Выполненные задания сдаются в электронном виде. На решение заданий отводится не менее 2 недель. За сдачу задания после 21 дня с даты получения итоговая оценка уменьшается на 10 %. В каждом задании есть теоретическая и практическая часть.

Итоговое задание представляет решение выбранной и согласованной с преподавателем задачи из предметной области курса (анализ данных и машинное обучение). Задание выполняется с использованием языка программирования Python и специализированных библиотек, обязательно должно присутствовать описание использованных данных, указание источника, загрузку данных, разведочный анализ, визуализацию описательных статистик, исследование с помощью методов снижения размерности, описание постановки задачи, построение классификационной или регрессионной модели (в зависимости от типа задачи), обоснование выбора метода и анализ качества полученной модели. Работа должна быть выполнена методологически корректно, без грубых ошибок. Программный код должен исполняться без фатальных ошибок. Студент должен понимать суть выполненной работы и быть готов дать пояснения по программному коду.

Итоговое задание сдается либо в виде файла *.ipynb – iPython Notebook, либо в виде презентации с приложением программного кода. Для решения задания допускается использовать язык программирования R.

Оценка за портфолио выставляется на основе суммы баллов за выполненные задания. Суммарное значение баллов, составляющее не менее 85 % от максимального, соответствует оценке «отлично», 70 % – «хорошо», 55 % – «удовлетворительно». Оценки «отлично», «хорошо», «удовлетворительно» соответствуют успешному прохождению аттестации.

2.1.2 Форма и перечень вопросов экзаменационного билета 2 семестра

Форма экзаменационного билета

Таблица П1.3

Новосибирский государственный университет
Экзамен

Анализ данных и машинное обучение

09.04.01 Информатика и вычислительная техника.
Искусственный интеллект и Data Science

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №

1. Вопрос из категории 1
2. Вопрос из категории 2

Составитель _____ Д.С.Мигинский

Ответственный за образовательную программу

_____М.М.Лаврентьев
(подпись)

«____»_____20 г.

Перечень вопросов для экзамена, структурированный по категориям, представлен в таблице П1.4

Таблица П1.4

Семестр 2	Формулировка вопроса
Категория 1 (ОПК-1.1, 1.2, 1.3, ОПК-4.1, 4.2, 4.3)	1. Специализированные библиотеки языка программирования Python для научных расчетов и анализа данных. NumPy, SciPy, pandas.
	2. Методы предобработки данных, описательные статистикеи и основные способы визуализации данных, методы снижения размерности.
	3. Метод главных компонент. Важность нормировки данных. Предобработка данных. Работа с пропущенными значениями.
	4. Основы машинного обучения и основные типы задач. Классификация задач машинного обучения
	5. Обучение на неразмеченных данных. Кластеризация. Иерархическая кластеризация. Метод K-средних, DBSCAN и др. Обзор методов кластеризации, реализованных в библиотеке sklearn.

Категория 2 (ОПК-1.1, 1.2, 1.3, ОПК-4.1, 4.2, 4.3)	6. Задачи обучения с учителем. Разделение данных на обучающие и тестовые. Нормировка данных. Определение переобученности модели. Критерии оценки качества полученных моделей.
	7. Постановка задачи регрессии. Линейный регрессионный анализ. Отбор признаков, коллинеарность, влиятельные наблюдения, анализ остатков. Непараметрическая регрессия (ядерное сглаживание). L1 и L2 регуляризация. Метрики качества.
	8. Постановка задачи классификации, обзор основных методов ее решения. Бинарная и многоклассовая классификация. Логистическая регрессия. Решающие деревья. Метрики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой).
	9. Ансамбли алгоритмов машинного обучения. Агрегирование моделей. Ансамбли решающих деревьев. Метод случайного леса. Градиентный бустинг.

3. Критерии оценки сформированности компетенций в рамках промежуточной аттестации по дисциплине

Таблица П1.5

Шифр компетенций	Структурные элементы оценочных средств	Показатель сформированности	Не сформирован (2 балла)	Пороговый уровень (3 балла)	Базовый уровень (4 балла)	Продвинутый уровень (5 баллов)
ОПК-1	Портфолио (этап 1), Экзамен (этап 2)	ОПК-1.1. Знать: математические, естественнонаучные и социально-экономические методы для использования в профессиональной деятельности	не знает основные способы визуализации данных (гистограммами, диаграммами плотности, диаграммами рассеяния, ящиками с усами и т.п.), реализованные в библиотеках matplotlib, seabron.	демонстрирует фрагментарные знания основных способов визуализации данных (гистограммами, диаграммами плотности, диаграммами рассеяния, ящиками с усами и т.п.), реализованные в библиотеках matplotlib, seabron.	демонстрирует базовые знания основных способов визуализации данных (гистограммами, диаграммами плотности, диаграммами рассеяния, ящиками с усами и т.п.), реализованные в библиотеках matplotlib, seabron	демонстрирует углубленные знания основных способов визуализации данных (гистограммами, диаграммами плотности, диаграммами рассеяния, ящиками с усами и т.п.), реализованные в библиотеках matplotlib, seabron
ОПК-1	Портфолио (этап 1), Экзамен (этап 2)	ОПК-1.2. Уметь: решать нестандартные профессиональные задачи, в том числе в новой или незнакомой среде и в междисциплинарном контексте, с применением математиче-	не умеет проводить разведочный анализ данных, проводить предобработку и очистку данных, работать с пропущенными значениями	демонстрирует поверхностное проводить разведочный анализ данных, проводить предобработку и очистку	демонстрирует умение проводить разведочный анализ данных, проводить предобработку и очистку данных, работать с про-	демонстрирует глубоко развитое умение проводить разведочный анализ данных, проводить предобработку и очистку данных, работать с пропущенными значениями

		ских, естественнонаучных, социально-экономических и профессиональных знаний		данных, работать с пропущенными значениями	пущенными значениями	
ОПК-1	Портфолио (этап 1)	ОПК-1.3. Владеть: навыками теоретического и экспериментального исследования объектов профессиональной деятельности, в том числе в новой или незнакомой среде и в междисциплинарном контексте	не владеет базовыми инструментами анализа данных и решения задач машинного обучения, реализованными в библиотеках pandas и sklearn.	демонстрирует неуверенное владение базовыми инструментами анализа данных и решения задач машинного обучения, реализованными в библиотеках pandas и sklearn.	демонстрирует владение базовыми инструментами анализа данных и решения задач машинного обучения, реализованными в библиотеках pandas и sklearn.	демонстрирует уверенное владение базовыми инструментами анализа данных и решения задач машинного обучения, реализованными в библиотеках pandas и sklearn.
ОПК-4	Портфолио (этап 1), Экзамен (этап 2)	ОПК-4.1. Знать: новые научные принципы и методы исследований	не знает особенности работы со специализированными программными библиотеками языка программирования Python для анализа данных и решения задач машинного обучения	демонстрирует фрагментарные знания особенностей работы со специализированными программными библиотеками языка программирования Python для анализа	демонстрирует базовые знания особенностей работы со специализированными программными библиотеками языка программирования Python для анализа данных и решения задач	демонстрирует углубленные знания особенностей работы со специализированными программными библиотеками языка программирования Python для анализа данных и решения задач машинного обучения

				данных и решения задач машинного обучения	машинного обучения	
ОПК-4	Портфолио (этап 1), Экзамен (этап 2)	ОПК-4.2. Уметь: применять на практике новые научные принципы и методы исследований	не умеет составлять конвейеры для предобработки данных, построения и подбора оптимальных гиперпараметров моделей.	демонстрирует поверхностное умение составлять конвейеры для предобработки данных, построения и подбора оптимальных гиперпараметров моделей.	демонстрирует умение составлять конвейеры для предобработки данных, построения и подбора оптимальных гиперпараметров моделей.	демонстрирует твердое умение составлять конвейеры для предобработки данных, построения и подбора оптимальных гиперпараметров моделей.
ОПК-4	Портфолио (этап 1)	ОПК-4.3. Владеть: навыками применения новых научных принципов и методов исследования для решения профессиональных задач	не умеет составлять композиции моделей (блендинг, стеккинг), проводить отбор признаков	демонстрирует неуверенное умение составлять композиции моделей (блендинг, стеккинг), проводить отбор признаков	демонстрирует умение составлять композиции моделей (блендинг, стеккинг), проводить отбор признаков	демонстрирует уверенное умение составлять композиции моделей (блендинг, стеккинг), проводить отбор признаков

4. Критерии выставления оценок по результатам промежуточной аттестации по дисциплине

Результаты промежуточной аттестации в семестре определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». Оценки «отлично», «хорошо», «удовлетворительно» означают успешное прохождение промежуточной аттестации.

Решение об окончательной оценке принимается по результатам 2-го этапа (экзамен)

Оценка «отлично» соответствует продвинутому уровню сформированности компетенции.

Оценка «хорошо» соответствует базовому уровню сформированности компетенции.

Оценка «удовлетворительно» соответствует пороговому уровню сформированности компетенции.

Оценка «неудовлетворительно» выставляется при неудовлетворительном прохождении одного или двух этапов промежуточной аттестации.