

Министерство науки и высшего образования
Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Донецкий государственный университет»

Факультет математики и информационных технологий
Кафедра теории упругости и вычислительной математики
имени академика А.С. Космодамианского



П.А. Машаров

« 29 » марта 2024 г.
МП

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ПРИКЛАДНЫЕ МОДЕЛИ АНАЛИЗА ДАННЫХ И МАШИННОГО ОБУЧЕНИЯ

Укрупненная группа направлений
подготовки
Программа высшего образования
Направление подготовки

Магистерская программа
Квалификация
Форма обучения

01.00.00 Математика и механика

Программа магистратуры
01.04.02 Прикладная математика и
информатика
Прикладная математика и информатика
Магистр
Очная

Рабочая программа адаптирована для лиц
с ограниченными возможностями здоровья и инвалидов

Донецк 2024

Рабочая программа дисциплины «**Прикладные модели анализа данных и машинного обучения**» для обучающихся по направлению подготовки 01.04.02 Прикладная математика и информатика (Магистерская программа: Прикладная математика и информатика), составлена на основании Федерального государственного образовательного стандарта высшего образования – магистратура по направлению подготовки 01.04.02 Прикладная математика и информатика, утвержденного приказом Министерства образования и науки Российской Федерации от 10 января 2018 г. № 13 (с изм. и доп.), Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры, утвержденного приказом Министерства науки и высшего образования Российской Федерации от 06 апреля 2021 г. № 245 (с изм. и доп.), в соответствии с учебным планом, утвержденным Ученым советом ФГБОУ ВО «ДонГУ» для набора 2024 года.

Разработчик:

доцент кафедры теории упругости и
вычислительной математики имени
академика А.С. Космодамианского
канд. физ.-мат. наук



Е.С. Глушанков

Рабочая программа одобрена на заседании кафедры теории упругости и вычислительной математики им. акад. А.С. Космодамианского.
Протокол от 26.03.2024 г. № 10.

Врио заведующего кафедрой



Р.Н. Нескородев

СОГЛАСОВАНО:

Декан факультета математики и
информационных технологий
28.03.2024 г.



И.А. Моисеенко

Учебно-методическая комиссия факультета математики и информационных технологий.
Протокол от 27.03.2024 г. № 3.
Председатель



Л.И. Селякова

Руководитель основной профессиональной
образовательной программы,
д-р физ.-мат. наук, доцент
26.03.2024 г.



Р.Н. Нескородев

1. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

1.1. Требования к предварительной подготовке обучающихся, предшествующие и сопутствующие дисциплины, на которых основывается изучение данной:

дисциплины программы бакалавриата: Математический анализ, Численные методы, Теория вероятностей и математическая статистика, Специализированный язык программирования Python в численных методах исследования моделей деформирования;

дисциплины программы магистратуры: Современные проблемы прикладной математики и информатики.

1.2. Дисциплины, курсовые работы и практики, для которых освоение данной дисциплины необходимо как предшествующее:

дисциплины программы магистратуры: Методы принятия решений,

2. ОПИСАНИЕ ДИСЦИПЛИНЫ

2.1. Общая характеристика

Наименование показателя	Значение показателя
Название образовательной программы	01.04.02 Прикладная математика и информатика (Магистерская программа: Прикладная математика и информатика)
Шифр и название в соответствии с учебным планом	Б1.В.ОД.4. Прикладные модели анализа данных и машинного обучения
Часть образовательной программы	Вариативная часть: выбор вуза
Количество зачетных единиц / всего часов	5 / 180

2.2. Распределение часов по формам и периодам обучения

Форма обучения	курс	семестр	Общее количество часов					Форма контроля
			лекционных	лабораторных	практических	самостоятельной работы + контроль	всего	
Очная	1	1	17	34	17	112	180	экзамен

3. ЦЕЛИ ДИСЦИПЛИНЫ

Целями дисциплины «Прикладные модели анализа данных и машинного обучения» являются освоение студентами основных теоретических сведений и практических навыков в области анализа данных и машинного обучения, ознакомление с современными методами решения задач классификации объектов, восстановления регрессии, кластеризации, ранжирования, основанными на современном математическом аппарате.

4. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ КОМПОНЕНТА ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ, ИХ ИНДИКАТОРЫ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

4.1. Компетенции

УК-1. Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

ОПК-3. Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности.

4.2. Индикаторы компетенций

УК-1.2. Разрабатывает стратегию действий для достижения поставленной цели с помощью методов анализа данных и машинного обучения.

ОПК-3.2. Разрабатывает и исследует полученные математические модели задач в области анализа данных и машинного обучения.

4.3. Результаты обучения

УК-1.2.1. Знает основные методы анализа данных и машинного обучения.

УК-1.2.2. Способен проводить анализ проблемных областей рассматриваемой задачи и выделять наиболее значимые признаки, факторы.

УК-1.2.3. Владеет методами анализа данных, позволяющими определять наиболее важные признаки, факторы в рассматриваемой задаче.

ОПК-3.2.1. Знаком с основными подходами к построению математических моделей в области анализа данных и машинного обучения.

ОПК-3.2.2. Умеет выбирать и использовать математические методы и программные средства для решения конкретных задач анализа данных и машинного обучения.

ОПК-3.3.3. Аргументированно выбирает методы решения конкретных задач анализа данных и машинного обучения, применяет их для построения математических моделей с последующим установлением свойств объектов и закономерностей связей между ними.

5. ПРОГРАММА ДИСЦИПЛИНЫ

Название темы	Краткое содержание темы (вопросы темы)
Раздел 1. Введение в машинное обучение	
История машинного обучения.	История машинного обучения.
Основные понятия в машинном обучении.	Постановка задачи машинного обучения. Типы задач машинного обучения. Признаки. Модель и алгоритм обучения. Обучение на основе минимизации эмпирического риска. Переобученность и недообученность модели.
Раздел 2. Задачи восстановления регрессии. Задачи классификации.	
Метрические алгоритмы.	Метрический алгоритм классификации. Метод ближайшего соседа. Метод k ближайших соседей. Метод k взвешенных ближайших соседей. Отступ объекта.
Разведочный анализ. Корреляция.	Корреляция. Коэффициенты корреляции Пирсона, Спирмена, Кенделла.
Методы восстановления регрессии.	Постановка задачи восстановления регрессии. Критерии оценки коэффициентов уравнения регрессии. Метод наименьших квадратов. Многомерная линейная регрессия. Коэффициент детерминации. Проблема мультиколлинеарности. Ridge regression. Lasso.

Линейная классификация.	Постановка задачи линейной классификации. Функционал ошибки для классификации. Оценка функции потерь. Обучение классификатора. Метрики качества классификации. Несбалансированные выборки.
Многоклассовые задачи.	One-vs-all. Матрица ошибок.
Решающие деревья. Случайные леса.	Логические правила. Понятие решающего дерева. Жадное построение решающего дерева. Энтропия. Алгоритм поиска разбиения. Неустойчивость решающих деревьев. Усреднение деревьев. Композиции алгоритмов. Majority vote и усреднение наблюдений. Бустинг. Бэггинг. Случайный лес. Поиска разбиения в случае бэггинга. Алгоритм построения случайного леса. Градиентный бустинг.
Метрики качества регрессии.	Среднеквадратичная ошибка. Средняя абсолютная ошибка. Математическое ожидание. Медиана. Мода. Оптимальный прогноз. Квантили. Коэффициент детерминации.
Метрики качества классификации.	Доля правильных ответов. Матрица ошибок. Точность. Полнота. Среднее арифметическое. Минимум. F-мера. Оценка принадлежности классу. PR-кривая. AUC PR. ROC-кривая. AUC ROC. Метрика Gini.
Анализ данных. Feature importance, feature extraction, feature selection	Постановка задачи понижения размерности. Важность признаков (feature importance). One hot encoding. Избыточные признаки. Шумовые признаки. Методы понижения размерности. Отбор признаков (feature selection) и извлечение признаков (feature extraction). Одномерные методы. Отбор с помощью моделей (embedded methods). Метод главных компонент.

Раздел 3. Задача классификации.	
Задача кластеризации	Постановка задачи кластеризации. Жёсткая и мягкая кластеризация. Метод К средних. Выбор количества кластеров. DBSCAN (Density-based clustering). Иерархическая кластеризация.
Раздел 4. Задача ранжирования.	
Задача ранжирования.	Постановка задачи ранжирования. Метрики качества ранжирования. Подходы в обучении модели ранжирования.

6. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Наименования разделов и тем	Количество часов				
	Лекц.	Лабор.	Практ.	СРС+К	Всего
Раздел 1. Введение в машинное обучение	4	6	2	8	20
История машинного обучения.	2	2	—	2	6
Основные понятия в машинном обучении.	2	4	2	6	14
Раздел 2. Задачи восстановления регрессии. Задачи классификации.	10	22	12	86	130
Метрические алгоритмы.	1	2	1	9	13
Разведочный анализ. Корреляция.	1	2	1	8	12
Методы восстановления регрессии.	1	3	2	10	16
Линейная классификация.	1	3	2	10	16
Многоклассовые задачи.	1	2	1	9	13
Решающие деревья. Случайные леса.	1	2	1	10	14
Метрики качества регрессии.	1	2	1	10	14
Метрики качества классификации.	1	2	1	10	14
Анализ данных. Feature importance, feature extraction, feature selection	2	4	2	10	18
Раздел 3. Задача классификации.	2	4	2	14	22
Задача кластеризации	2	4	2	14	2
Раздел 4. Задача ранжирования.	1	2	1	4	8
Задача ранжирования.	1	2	1	4	8
ИТОГО ПО КОМПОНЕНТУ ОПОП	17	34	17	112	180

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ (СРЕДСТВА) ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

7.1. Контрольные вопросы

Раздел 1

1. Постановка задачи машинного обучения.
2. Типы задач машинного обучения.
3. Признаки.
4. Модель и алгоритм обучения.
5. Обучение на основе минимизации эмпирического риска.
6. Степени обученности модели.
7. Переобученность и недообученность.

Раздел 2

8. Метрический алгоритм классификации.
9. Метод ближайшего соседа.
10. Метод k ближайших соседей.
11. Метод k взвешенных ближайших соседей.
12. Отступ объекта.
13. Корреляция.
14. Коэффициенты корреляции Пирсона, Спирмена, Кенделла.
15. Постановка задачи восстановления регрессии.
16. Критерии оценки коэффициентов уравнения регрессии.
17. Метод наименьших квадратов.
18. Многомерная линейная регрессия.
19. Коэффициент детерминации.
20. Проблема мультиколлинеарности.
21. Ridge regression. Lasso.
22. Постановка задачи линейной классификации.
23. Функционал ошибки для классификации.
24. Оценка функции потерь.
25. Обучение классификатора.
26. Метрики качества классификации.
27. Несбалансированные выборки.
28. One-vs-all.
29. Матрица ошибок.
30. Логические правила.
31. Понятие решающего дерева.
32. Жадное построение решающего дерева.
33. Энтропия.
34. Алгоритм поиска разбиения.
35. Неустойчивость решающих деревьев.
36. Усреднение деревьев.
37. Композиции алгоритмов. Majority vote и усреднение наблюдений.
38. Бустинг.
39. Бэггинг. Случайный лес.
40. Поиска разбиения в случае бэггинга.
41. Алгоритм построения случайного леса.
42. Градиентный бустинг.
43. Среднеквадратичная ошибка.
44. Средняя абсолютная ошибка.
45. Математическое ожидание.
46. Медиана.
47. Мода.
48. Оптимальный прогноз.
49. Квантили.
50. Коэффициент детерминации.
51. Доля правильных ответов.
52. Матрица ошибок.
53. Точность.
54. Полнота.
55. Среднее арифметическое.
56. Минимум.
57. F-мера.
58. Оценка принадлежности классу.

59. PR-кривая. AUC PR.
60. ROC-кривая. AUC ROC.
61. Метрика Gini.
62. Постановка задачи понижения размерности.
63. Избыточные признаки. Шумовые признаки.
64. Методы понижения размерности.
65. Важность признаков (feature importance).
66. Отбор признаков (feature selection) и извлечение признаков (feature extraction).
67. Одномерные методы.
68. Отбор с помощью моделей (embedded methods).
69. Метод главных компонент.

Раздел 3

70. Постановка задачи кластеризации.
71. Жёсткая и мягкая кластеризация.
72. Метод K средних.
73. Выбор количества кластеров.
74. DBSCAN (Density-based clustering).
75. Иерархическая кластеризация.

Раздел 4

76. Постановка задачи ранжирования.
77. Метрики качества ранжирования.
78. Подходы в обучении модели ранжирования.

7.2. Темы письменных работ (типы задач)

Контрольная работа по проверке теоретических знаний – по всем темам, с использованием указанных выше контрольных вопросов.

7.3. Темы индивидуальных заданий

- Восстановление линейной регрессии.
- One Hot Encoding.
- Сравнение методов градиентного бустинга из библиотек Scikit-learn и Catboost.
- Восстановление нелинейной регрессии с помощью решающих деревьев и случайного леса;
- Задача кластеризации.
- Итоговое индивидуальное задание (восстановление линейной регрессии, обучение моделей решающего дерева, бустинг и бэггинг, задача кластеризации).

7.4. Образец содержания экзаменационного билета

ФГБОУ ВО «ДОНЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Кафедра теории упругости и вычислительной математики
имени академика А.С. Космодамианского

Направление подготовки:	01.04.02 Прикладная математика и информатика
Магистерская программа:	Прикладная математика и информатика
Программа подготовки:	магистратура
Семестр:	1
Учебная дисциплина:	«Прикладные модели анализа данных и машинного обучения»

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Метод k ближайших соседей.
2. Метрики качества классификации.
3. Определить коэффициенты уравнения линейной регрессии зависимости показателя Y от признака X .

X	1	2	3	4	5
Y	2	5	9	14	19

Утверждено на заседании кафедры теории упругости и вычислительной математики имени академика А.С. Космодамианского.

Протокол № __ от «__» _____ 20__ года.

Заведующий кафедрой _____

Экзаменатор _____

В случае ведения учебного процесса с использованием электронного обучения и дистанционных образовательных технологий, содержание билета может отличаться от приведенного.

8. РАСПРЕДЕЛЕНИЕ БАЛЛОВ, КОТОРЫЕ ПОЛУЧАЮТ ОБУЧАЮЩИЕСЯ

Общая оценка знаний обучающихся по дисциплине проводится по 100-балльной шкале исходя из максимума, приведенного в таблице ниже.

По результатам работы в семестре обучающийся, набравший не менее 60 баллов, имеет право получить оценку. Те, кого набранные баллы не устраивают, сдают индивидуальные задания. Максимальное количество баллов за экзамен – 75. Оценка за семестр вычисляется как максимальная из полученных за семестр и на экзамене и выставляется согласно принятому порядку.

Номера разделов	Виды работ	Максимальное количество баллов
2	Индивидуальные задания	43
3	Индивидуальные задания	47
4	Контрольная работа по проверке теоретических знаний	10
ИТОГО		100
Экзамен		75
Общий итог за семестр		100

Соответствие баллов оценке

Количество баллов из 100	ECTS	Оценка по пятибалльной шкале
		Экзамен
90-100	A	отлично
80-89	B	хорошо
75-79	C	
70-74	D	удовлетворительно
60-69	E	
35-59	FX	неудовлетворительно
0-34	F	

9. ОБЕСПЕЧЕНИЕ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ДЛЯ ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ И ИНВАЛИДОВ

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- 1) для слепых и слабовидящих:
 - лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;
 - для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств;
 - письменные задания оформляются увеличенным шрифтом.
- 2) для глухих и слабослышащих:
 - лекции оформляются в виде электронного документа;
 - письменные задания выполняются на компьютере в письменной форме;
 - экзамен проводится в письменной форме на компьютере; возможно проведение в форме тестирования.
- 3) для лиц с нарушениями опорно-двигательного аппарата:
 - лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;
 - письменные задания выполняются на компьютере;
 - экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- 1) для слепых и слабовидящих:
 - в печатной форме увеличенным шрифтом;
 - в форме электронного документа;
- 2) для глухих и слабослышащих:
 - в печатной форме;
 - в форме электронного документа.
- 3) для обучающихся с нарушениями опорно-двигательного аппарата:
 - в печатной форме;
 - в форме электронного документа.

10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОГО ПРОЦЕССА

Учебные занятия проводятся в Главном корпусе ДонГУ (г. Донецк, пр. Гурова, 6), в Учебно-практическом вычислительном центре ДонГУ (г. Донецк, пр. Гурова, 6, корпус 12).

Для проведения лекций требуется аудитория, оборудованная меловой или маркерной доской / сенсорным экраном / мультимедийный проектор с экраном и ноутбуком, комплект учебной мебели для студентов, рабочее место преподавателя.

Для проведения практических занятий требуется аудитория, оборудованная меловой или маркерной доской / сенсорным экраном / мультимедийный проектор с экраном и ноутбук, комплект учебной мебели для студентов, рабочее место преподавателя.

Для проведения лабораторных занятий требуется аудитория, оборудованная маркерной доской или сенсорным экраном / мультимедийный проектор с экраном и ноутбук, персональные компьютеры, комплект учебной мебели для студентов, рабочее место преподавателя, выход в Интернет – проводной или с использованием Wi-Fi.

Для самостоятельной работы используются текстовые и электронные ресурсы Научной библиотеки университета и других электронных библиотечных баз данных, учебно-методическое обеспечение, представленное в аудиториях Главного корпуса (ауд. 511, 605, 610).

Обучающиеся имеют возможность использовать учебные материалы по дисциплине, размещенные на платформе Moodle Центра дистанционного образования ФГБОУ ВО «ДонГУ». При изучении дисциплины применяются электронное обучение и дистанционные образовательные технологии.

С использованием ресурсов платформы дистанционного образования осуществляется текущий контроль знаний обучающихся на основе тестирования и проверки результатов самостоятельной работы.

11. РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

11.1. Основная литература

1. Брандт З. Анализ данных : стат. и вычисл. методы для науч. работников и инженеров / З. Брандт. – М.: Мир : АСТ, 2003. – 686 с.
2. Шарден Б. Крупномасштабное машинное обучение вместе с Python / Б.Шарден, Л. Массарон, А. Боскетти. – М.: ДМК Пресс, 2017. – 358 с.

11.2. Дополнительная литература

3. Бондаренко Н.С. Анализ данных в научных исследованиях : учеб.-метод. пособие / Н.С. Бондаренко. – Донецк: ДонНУ, 2018. – 173 с.
4. Крутиков В.Н. Анализ данных : учеб. пособие / В.Н. Крутиков, В.В. Мешечкин. – Кемерово: КГУ, 2014. – 137 с.

12. ИНФОРМАЦИОННЫЕ РЕСУРСЫ

1. Национальная электронная библиотека (НЭБ): федеральная государственная информационная система / Министерство Культуры РФ; Российская государственная библиотека. – Москва, 2019- . – URL: <https://rusneb.ru/> (дата обращения: 01.09.2023). – Режим доступа: свободный, подписка. Необходима установка программного обеспечения. – Текст: электронный.
2. eLIBRARY.RU: научная электронная библиотека: сайт. – Москва, 2000- . – URL: <https://elibrary.ru> (дата обращения: 01.09.2023). – Режим доступа: для авторизов. пользователей. – Текст: электронный.
3. Научная электронная библиотека «КиберЛенинка»: сайт / Ассоциация «Открытая наука». – Москва, 2014- . – URL: <https://cyberleninka.ru/>. – Режим доступа: свободный. – Текст: электронный.
4. Электронно-библиотечная система «Лань»: [сайт]. – URL: <https://e.lanbook.com> (дата обращения: 01.09.2023). – Режим доступа: для авторизов. пользователей. – Текст: электронный.
5. ЭБС Юрайт: электронная библиотечная система: сайт. – Москва, 2013. – URL: <https://biblio-online.ru> (дата обращения: 01.09.2023). – Режим доступа: для авторизов. пользователей. – Текст: электронный.
6. Электронно-библиотечная система ДонГУ: сайт / ФГБОУ ВО «ДонГУ». – Донецк, 2016- . – URL: <http://library.donnu.ru/> (дата обращения: 01.09.2023). – Режим доступа: свободный. – Текст: электронный.

7. Электронный каталог Научной библиотеки ДонГУ: раздел сайта / НБ ДонГУ. – Текст: электронный // ЭБС ДонГУ: сайт. – URL: <http://library.donnu.ru/catalog/> (дата обращения: 01.09.2023). – Режим доступа: поиск свободный, электронные документы – для пользователей ДонГУ.

8. Электронный архив ДонГУ: раздел сайта / НБ ДонГУ. – Текст: электронный // ЭБС ДонГУ: сайт. – URL: <http://hero.donnu.ru/> (дата обращения: 01.09.2023). – Режим доступа: свободный.

13. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

1. Windows 7 PRO (корпоративная лицензия ДонГУ № 46484614)
2. Microsoft Office (корпоративная лицензия ДонГУ № 46472919)
3. Microsoft Visual Studio (лицензия программы Dream Spark для высших учебных заведений)
4. Python 3.12.2 (лицензия PSF для свободного программного обеспечения)
5. Антивирус Касперского, Adobe Acrobat Reader, xPDF (лицензии GPL, Apache, BSD для свободного программного обеспечения).