

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ
Заведующий кафедрой

Математических методов исследования операций



Азарнова Т.В.

29 мая 2023 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.О.31 Практикум по машинному обучению

1. Код и наименование направления подготовки/специальности:

38.03.05 Бизнес-информатика

2. Профиль подготовки/специализация:

Бизнес-аналитика и системы автоматизации предприятий

3. Квалификация (степень) выпускника: бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: *Математических методов исследования операций*

6. Составители программы: *Замятин Игорь Викторович, к. ф.-м. наук*

7. Рекомендована: Научно-методическим советом факультета прикладной математики, информатики и механики Протокол №7 от 26.05.2023

8. Учебный год: 2025/2026

Семестр(ы): 5

9. Цели и задачи учебной дисциплины:

Целью курса является ознакомление будущих специалистов в области бизнес-информатики с процессами, алгоритмами и инструментами, относящимися к основным принципам машинного обучения.

Задачи курса: сформировать теоретические знания по основам машинного обучения для построения формальных математических моделей и интерпретации результатов моделирования; выработать умения по практическому применению методов машинного обучения при решении прикладных задач в различных областях; выработать умения и навыки использования библиотек языка Python для разработки систем машинного обучения.

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к обязательным дисциплинам базового цикла (блок Б1). Для изучения курса необходимы базовые знания информатики, линейной алгебры, теории вероятностей, математической статистики, методов оптимизации.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения:

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
ОПК-4	Способен использовать информацию, методы и программные средства ее сбора, обработки и анализа для информационно-аналитической поддержки принятия управленческих решений	ОПК-4.1	Собирает и анализирует информацию для поддержки принятия решений	Знать: - возможности актуальных алгоритмов машинного обучения, которые широко используются на практике, основные сферы их применения. Уметь: - применять методы машинного обучения при решении задач в различных прикладных областях. Владеть: - навыками интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов.
		ОПК-4.2	Использует методы и программные средства обработки информации	Знать: - методы предварительной обработки данных (кодирование, стандартизация и нормализация, устранение выбросов, заполнение пропусков); - методы отбора информативных признаков. Уметь: - анализировать многомерные данные и преодолевать вычислительные проблемы, связанные с высокой размерностью данных.

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
		ОПК-4.3	Использует методы и программные средства анализа информации	<p>Знать:</p> <ul style="list-style-type: none"> - методы классификации; - методы регрессионного анализа <p>методы анализа текстовых данных.</p> <p>Уметь:</p> <ul style="list-style-type: none"> - использовать библиотеки языка Python для построения моделей машинного обучения. <p>Владеть:</p> <ul style="list-style-type: none"> - навыками построения и проверки качества моделей машинного обучения; - навыками использования библиотек языка Python для построения систем, обучающихся по прецедентам.

12. Объем дисциплины в зачетных единицах/час — 2/72.

Форма промежуточной аттестации *зачет*

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость (часы)		
		Всего	В том числе в интерактивной форме	По семестрам
Аудиторные занятия		34	34	34
в том числе:	лекции			
	практические			
	лабораторные	34	34	34
Самостоятельная работа		38		38
Итого:		72	34	72
Форма промежуточной аттестации		Зачет		Зачет

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
1. Лабораторные занятия			
1.1	Введение в машинное обучение. Основные определения и постановки задач.	Основные этапы решения задачи анализа данных. Примеры прикладных задач. Виды обучения: с учителем, без учителя, с подкреплением. Основные типы задач: задача классификации, задача регрессии, задача кластеризации, задача прогнозирования, задача ранжирования. Основные проблемы машинного обучения: недостаточный объем обучающей выборки, пропуски в данных, переобучение	Б1.О.31 Практикум по машинному обучению
1.2	Обзор основных необходимых библиотек языка Python	Библиотека NumPy для оптимизированных вычислений над массивами данных. Введение в массивы библиотеки NumPy. Выполнение вычислений над массивами библиотеки NumPy, универсальные функции. Операции над данными в библиотеке Pandas. Обработка отсутствующих данных. Агрегирование и группировка. Визуализация с	Б1.О.31 Практикум по машинному обучению

		помощью библиотеки Matplotlib. Линейные графики, диаграммы рассеяния, гистограммы, трехмерные графики. Знакомство с библиотекой машинного обучения Scikit-Learn. Гиперпараметры и проверка качества модели	
1.3	Построение и отбор признаков	Извлечение признаков (Feature Extraction). Преобразования признаков (Feature transformations): кодирование нечисловых данных, нормировка и калибровка, заполнение пропусков Выбор признаков (Feature selection): статистические подходы, визуализация, отбор с использованием моделей	Б1.О.31 Практикум по машинному обучению
1.4	Решение задачи регрессии	Метод наименьших квадратов. Измерение ошибки в задачах регрессии (MSE , $RMSE$, MAE , R^2). Многомерная регрессия, проблема мультиколлинеарности. Регрессия, линейная по параметрам, полиномиальная регрессия. Решение проблемы переобучения: L1- регуляризация (Lasso), L2- Регуляризация (гребневая регрессия), эластичная сеть. Настройка гиперпараметров алгоритма с помощью n-кратной перекрестной проверки. Разбор примера построения модели линейной регрессии для задачи предсказания велосипедного трафика Отбор и кодирование признаков. Визуальное сравнение общего и предсказанного моделью трафика. Проверка качества. Построение модели линейной регрессии с помощью библиотеки Scikit-Learn для заданного набора данных. Анализ качества построенной модели.	Б1.О.31 Практикум по машинному обучению
1.5	Решение задачи классификации	Линейная модель классификации. Логистическая регрессия как линейный классификатор. Функция потерь (ошибок классификации). Логистическая функция потерь с учетом L2-регуляризации. Использование полиномиальных признаков для нелинейного разделения. Confusion matrix (матрица ошибок классификации). Метрики качества классификации: accuracy (доля правильных ответов), precision (точность), recall (полнота), F1- мера. AUC-ROC – площадь под кривой ошибок. Метрическая классификация - метод ближайших соседей (kNN). Использование наивной байесовской модели для классификации Разбор примера построения модели логистической регрессии для задачи предсказания оттока клиентов мобильного оператора. Отбор и кодирование признаков. Проверка качества модели с помощью перекрестной проверки. Построение модели логистической регрессии с помощью библиотеки Scikit-Learn. Анализ качества построенной модели	Б1.О.31 Практикум по машинному обучению
1.6.	Древовидные модели: деревья решений, случайный лес	Этапы построения дерева решений, выбор критерия точности прогноза. типа ветвления. Метрики ветвления на основе прироста информации (алгоритм ID3), нормализованного прироста информации (алгоритм C4.5), индекса Джини (алгоритм CART). Правила разбиения. Механизм отсечения дерева. Критерии останова алгоритма (минимальное число объектов, при котором выполняется расщепление, минимальное число объектов в листьях, максимальная глубина деревьев. Переобучение решающих деревьев. Случайный лес. Обучение случайного леса. Достоинства и недостатки случайного леса. Разбор примера построения модели дерева решений для задачи предсказания исхода футбольного матча. Анализ деревьев, полученных при использовании различных метрик. Построение модели случайного леса на примере задачи кредитного скоринга. Кодирование признаков и заполнение пропущенных данных. Построение моделей деревьев решений и случайного леса с помощью библиотеки Scikit-Learn для заданного набора	Б1.О.31 Практикум по машинному обучению

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				Всего
		Лекции	Практические	Лабораторные	Самостоятельная работа	
1	Введение в машинное обучение. Основные определения и постановки задач.			6	6	12
2	Обзор основных необходимых библиотек языка Python			8	10	18
3	Построение и отбор признаков			4	6	10
4	Решение задачи регрессии			4	4	8
5	Решение задачи классификации.			6	6	12
6	Древовидные модели: деревья решений, случайный лес			6	6	12
	Итого			34	38	72

14. Методические указания для обучающихся по освоению дисциплины

Работа с конспектами занятий, презентациями, выполнение практических заданий для самостоятельной работы, выполнение лабораторных работ, использование рекомендованной литературы и методических материалов, в том числе размещенных на странице курса «Б1.О.31 Практикум по машинному обучению» на портале «Электронный университет ВГУ» <https://edu.vsu.ru/course/view.php?id=10157>, автор Замятин И.В.

В рамках общего объема часов, отведенных для изучения дисциплины, предусматривается выполнение следующих видов самостоятельных работ студентов (СРС): изучение теоретического материала, написание программ по темам, изученным на лабораторных занятиях.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

а) основная литература:

№ п/п	Источник
1	Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения [Электронный ресурс] : руководство / С. Рашка ; пер. с англ. Логунова А.В.. — Электрон. дан. — Москва : ДМК Пресс, 2017. — 418 с. — Режим доступа: https://e.lanbook.com/book/100905
2	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. — Москва: ДМК Пресс, 2015. — 400 с. — ISBN 978-5-97060-273-7. — Текст: электронный // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/69955
3	Замятин И.В. Программирование на языке Python [Электронный ресурс] : учебно-методическое пособие : [для студ. 3-го курса, обучающихся по направлению 38.03.05 - Бизнес-информатика] / И.В. Замятин ; Воронеж. гос. ун-т. — Воронеж: Издательский дом ВГУ, 2019.— Свободный доступ из интрасети ВГУ. — URL: http://www.lib.vsu.ru/elib/texts/method/vsu/m19-160.pdf

б) дополнительная литература:

№ п/п	Источник
4	Козьмо, Л. П. Построение систем машинного обучения на языке Python / Л. П. Козьмо, В.

	Ричарт; перевод с английского А. А. Слинкин. — 2-е изд. — Москва: ДМК Пресс, 2016. — 302 с. — ISBN 978-5-97060-330-7. — Текст: электронный // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/82818
5	Каширина И.Л. Нейросетевые технологии : учебно-методическое пособие для вузов / И.Л. Каширина ; Воронеж. гос. ун-т. — Воронеж : ИПЦ ВГУ, 2008. — 70 с. : ил. — <URL: http://www.lib.vsu.ru/elib/texts/method/vsu/m08-110.pdf >.
6	Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.
7	Прикладные методы анализа статистических данных [Электронный ресурс]: учебное пособие / Е.Р. Горяинова, А.Р. Панков, Е.Н. Платонов. — Электрон. дан. — М.: Издательский дом Высшей школы экономики, 2012. — 312 с. — Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=65997
8	Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы: Пер.с польск.И.Д.Рудинского. [Электронный ресурс] : / Рутковская Д., Пилиньский М., Рутковский Л. — Электрон. дан. — М.: Горячая линия-Телеком, 2013. — 384 с. — Режим доступа: http://e.lanbook.com/books/element.php?pl1_id=11843

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет):

№ п/п	Источник
9	Электронная библиотечная система «Издательства «Лань». Режим доступа: http://e.lanbook.com/
10	Б1.О.12 Интеллектуальный анализ данных / И.В. Замятин — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=6188
11	Б1.О.31 Практикум по машинному обучению / И.В. Замятин — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=10157
12	Электронная библиотечная система ВГУ. Режим доступа: http://www.lib.vsu.ru

16. Перечень учебно-методического обеспечения для самостоятельной работы

Самостоятельная работа обучающегося должна включать подготовку к лабораторным занятиям, выполнение текущих заданий по освоению соответствующих тем курса, выполнение курсовой работы и подготовку к промежуточной аттестации. Для этого рекомендуется освоить теоретический материал соответствующих тем по конспектам лекций, презентационному материалу, размещенному на ЭО ресурсах, литературу из представленного перечня, материалы с тематических ресурсов сети Интернет.

№ п/п	Источник
1	Б1.О.12 Интеллектуальный анализ данных / И.В. Замятин — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=6188
2	Б1.О.31 Практикум по машинному обучению / И.В. Замятин — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=10157
3	Электронная библиотечная система ВГУ. Режим доступа: http://www.lib.vsu.ru

17. Информационные технологии, используемые для реализации учебной дисциплины, включая программное обеспечение и информационно-справочные системы (при необходимости)

Python 3 с подключенными библиотеками (дистрибутив Anaconda).

Дисциплина реализуется с применением электронного обучения и дистанционных образовательных технологий. Для организации самостоятельной работы обучающихся используется онлайн-курс «Интеллектуальный анализ данных», размещенный на платформе Электронного университета ВГУ (LMS moodle), а также Интернет-ресурсы, приведенные в п.15в.

18. Материально-техническое обеспечение дисциплины:

Практические и лабораторные занятия должны проводиться в специализированной аудитории, оснащенной современными персональными компьютерами и программным обеспечением в соответствии с тематикой изучаемого материала, а также компьютером с подключенным к нему проектором с видеотерминала на настенный экран.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция (и)	Индикатор(ы) достижения компетенции	Оценочные средства
1.	Введение в машинное обучение. Основные определения и постановки задач.	ОПК-4	ОПК-4.1	Задание для Лабораторной работы №1
2.	Обзор основных необходимых библиотек языка Python		ОПК-4.2	
3.	Построение и отбор признаков		ОПК-4.3	Задание для Лабораторной работы №2
4.	Решение задачи регрессии		ОПК-4.3	Задание для Лабораторной работы №3
5.	Решение задачи классификации.		ОПК-4.3	Задание для Лабораторной работы №4
6.	Древовидные модели: деревья решений, случайный лес.		ОПК-4.3	Задание для Лабораторной работы №4
Промежуточная аттестация форма контроля - зачет				Задания для лабораторных работ

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

Лабораторная работа №1

1. На сайтах <https://www.kaggle.com/> , <https://archive.ics.uci.edu/ml/index.php> , или любом другом найти и загрузить произвольный датасет. Требования к датасету:

- не менее 7 столбцов
- не менее 3 столбцов, содержащих НЕчисловые данные

2. Создать проект (блокнот) Python и загрузить в него выбранный датасет.

3. Определить тип задачи машинного обучения и указать искомую (зависимую) переменную.

4. Выполнить подготовку данных датасета для дальнейшего использования в модели машинного обучения:

- Выполнить визуализацию данных датасета.
- Выполнить базовый статистический анализ.
- Выполнить необходимые преобразования типов данных.
- Выполнить очистку данных (от пропущенных значений).
- (факультативно) Выполнить масштабирование данных.

Лабораторная работа № 2 (регрессия)

1. Разбейте предоставленный Вам преподавателем набор данных на обучающую и тестовую части в соотношении 8:2.
2. Обучите, а затем провалидируйте на тестовых данных следующие модели, используя в качестве метрики качества R^2 , предварительно отмасштабируя данные
 - LinearRegression;
 - Lasso с коэффициентом регуляризации, равным 0.01.
3. Проанализируйте качество получившихся моделей и сравните количество строго нулевых весов в них.

Лабораторная работа № 3 (классификация)

1. Разбейте предоставленный Вам преподавателем набор данных на обучающую и тестовую части в соотношении 8:2.
2. Проведите предобработку данных: заполнение пропусков, кодирование, масштабирование
3. Обучите, а затем провалидируйте на тестовых данных модель логистической регрессии, наивный байесовский классификатор.
4. Вычислите значения метрик: recall, precision, F1-мера, AUC-ROC. Постройте ROC-кривую.

Лабораторная работа № 4 (деревья решений)

1. Разбейте предоставленный Вам преподавателем набор данных на обучающую и тестовую части в соотношении 8:2.
2. Проведите предобработку данных: заполнение пропусков, кодирование, масштабирование
3. Обучите, а затем провалидируйте на тестовых данных модели дерево решений и случайный лес.
4. Вычислите значения метрик: recall, precision, F1-мера, AUC-ROC. Постройте ROC-кривую.

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: Собеседование по вопросам к зачету. Контрольно-измерительные материалы промежуточной аттестации включают в себя теоретические вопросы, позволяющие оценить уровень полученных знаний и практические задания, позволяющие оценить степень сформированности умений и навыков.

Промежуточная аттестация проводится в соответствии с Положением о промежуточной аттестации обучающихся по программам высшего образования.

Для оценивания результатов обучения на зачете используются следующие показатели:

- 1) знание учебного материала и владение понятийным аппаратом теории машинного обучения;
- 2) умение анализировать многомерные данные и преодолевать вычислительные проблемы, связанные с высокой размерностью данных;
- 3) умение применять методы машинного обучения при решении задач в различных прикладных областях;
- 5) владение навыками использования библиотек языка Python для построения систем, обучающихся по прецедентам
- 6) владение навыками построения и проверки качества моделей машинного обучения;
- 7) владение навыками интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов.

По учебному плану предусмотрен зачет.

Критерии оценки «зачтено» — продемонстрировано знание теоретического материала, положительные результаты решения тестовых лабораторных работ.

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), сдал все лабораторные работы.	Базовый уровень	Зачтено
Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не сдал хотя бы одну лабораторную работу.	–	Не зачтено

20.3 Фонд оценочных средств сформированности компетенций студентов, рекомендуемый для проведения диагностических работ

ОПК-4

Вопросы с выбором ответа

1. С помощью метода деревьев решений возможно решение задач:
 - a. Регрессии.
 - b. Классификации.
 - c. Классификации и регрессии.**
 - d. Кластеризации.
2. Какой из видов задач машинного обучения направлен на предсказание значения той или иной непрерывной числовой величины на основе входных данных?
 - a. Регрессия.**
 - b. Классификация.
 - c. Кластеризация.
 - d. Переобучение.
3. Множество примеров, используемое для обучения модели, называется...
 - a. обучающим множеством**
 - b. валидационным множеством
 - c. тестовым множеством
4. Обучающая выборка в задаче обучения с учителем — это...

- a. **набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат**
- b. выявление в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности
- c. группировка объектов на основе данных, описывающих свойства объектов

5. Что такое стандартизация данных?

- a. Перевод значений всех признаков в числовую шкалу измерения
- b. Выделение значимых признаков
- c. Преобразование входных признаков, при котором все входные значения масштабируются в отрезок $[0,1]$
- d. **Преобразование входных признаков так, чтобы среднее значение каждого признака было 0, а стандартное отклонение 1**

6. Задача классификации сводится к ...

- a. поиску независимых групп и их характеристик в всем множестве анализируемых данных
- b. определению по известным характеристикам объекта значение некоторого его параметра
- c. нахождению частых зависимостей между объектами или событиями
- d. **определению класса объекта по его характеристикам**

7. Переобучение - это ...

- a. излишнее обучение модели, не дающее прироста точности
- b. повторное обучение модели для проверки ее корректности
- c. дообучение модели на новых данных
- d. **когда построенная модель хорошо объясняет примеры из обучающей выборки, но плохо работает на новых данных**

8. При решении задачи регрессии ищут...

- a. правила или набор правил в соответствии с которыми любой новый объект можно отнести к одному из классов
- b. **функциональные зависимости, которые позволяют прогнозировать изменения непрерывных числовых параметров**
- c. группы, на которые можно разделить объекты, данные о которых подвергаются анализу
- d. соотношения между зависимыми и независимыми показателями и переменными в наглядной и понятной человеку форме

9. Что такое перекрестная проверка (Cross-validation)?
(укажите все правильные варианты):
- a. Метод формирования обучающего и тестового множеств для обучения модели в условиях недостаточности исходных данных или неравномерного представления классов.
 - b. Построение нескольких моделей для одного исходного набора данных.
 - c. Метод оценки эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Открытые вопросы

10. Дана матрица ошибок, построенная по результатам работы некоторого алгоритма классификации. Общая точность (accuracy) равна...

		Истинный класс	
		1	-1
Предсказанный класс	1	25	20
	-1	20	15

a. 0,5

- b. Критерий оценивания: точное соответствие

Критерии оценивания для вопросов №1-14: указан правильный вариант – 1, в противном случае – 0.

Критерии оценивания для вопросов №15, 16: указаны оба правильных варианта – 1, указан один из правильных вариантов – 0,5, не указано ни одного правильного варианта – 0.

Задания раздела 20.3 рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных результатов освоения данной дисциплины (знаний, умений, навыков).