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.
29.05.2023

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.01 Прикладное машинное обучение на языке Python

1. Код и наименование направления подготовки/специальности:

02.04.02 Фундаментальная информатика и информационные технологии

2. Профиль подготовки/специализация:

Машинное обучение и интеллектуальные информационные технологии

3. Квалификация выпускника: магистр

4. Форма обучения: очная

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

6. Составители программы:
Каширина Ирина Леонидовна, доктор техн. наук, профессор

7. Рекомендована: НМС факультета прикладной математики, информатики и механики, Протокол № 7 от 26.05.2023 г.

8. Учебный год: 2023/2024 **Семестр(-ы):** 1

9.Цели и задачи учебной дисциплины

Цель изучения дисциплины:

ознакомление будущих специалистов в области Data Science с процессами, алгоритмами и инструментами, относящимися к основным принципам машинного обучения

Задачи учебной дисциплины:

−сформировать теоретические знания по основам машинного обучения для построения формальных математических моделей, анализа и обработки информации по тематике исследований;

−выработать умения по практическому применению методов машинного обучения при решении прикладных задач в различных областях, в том числе при работе с большими данными;

−выработать умения и навыки использования библиотек языка Python для разработки прикладного программного обеспечения на основе алгоритмов машинного обучения.

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к обязательным дисциплинам вариативной части базового цикла. Для изучения курса необходимы базовые знания информатики, линейной алгебры, математического анализа, 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
ПК-1	Способен проводить работы по обработке и анализу научно-технической информации результатов исследований	ПК-1.2	Анализирует и обрабатывает информацию по тематике исследований.	знать: – методы предварительной обработки данных (кодирование, стандартизация и нормализация, устранение выбросов, заполнение пропусков); – методы отбора информативных признаков; – методы классификации; – методы регрессионного анализа – методы анализа текстовых данных. уметь: – анализировать многомерные данные и преодолевать вычислительные проблемы, связанные с высокой размерностью данных; владеть (иметь навык(и)): – построения и проверки качества моделей машинного обучения; интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов;
ПК-4;	Способен разрабатывать профессионально-ориентированные программные средства и приложения на	ПК-4.2	Использует знания в области искусственного интеллекта, инженерии знаний, машинного	знать: – возможности актуальных алгоритмов машинного обучения, которые широко используются на практике, основные сферы их применения; уметь:

	основе интеллектуальных информационных технологий.		обучения для разработки прикладного программного обеспечения.	– применять методы машинного обучения при решении задач в различных прикладных областях; – использовать библиотеки языка Python для построения моделей машинного обучения; владеть (иметь навык(и)): использования библиотек языка Python для построения систем, обучающихся по прецедентам.
ПК-5	Способен совершенствовать и разрабатывать новые методы, модели, алгоритмы, технологии работы с большими данными.	ПК-5.1	Совершенствует и разрабатывает модели и алгоритмы машинного обучения для работы с большими данными.	знать: – модели и алгоритмы машинного обучения для работы с большими данными; уметь: – применять методы машинного обучения для работы с большими данными;

12. Объем дисциплины в зачетных единицах/час.(*в соответствии с учебным планом)* — 5/180.

Форма промежуточной аттестации *экзамен*

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость	
		Всего	По семестрам
			№ семестра
Контактная работа			
в том числе:	лекции	32	32
	практические		
	лабораторные	32	32
Самостоятельная работа		80	80
Промежуточная аттестация (экзамен)		36	36
Итого:		180	180

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
		1. Лекции	
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	Машинное обучение на языке Python

		данных,переобучение	
1.2	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.	Машинное обучение на языке Python
1.3	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	Машинное обучение на языке Python
1.4.	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	Машинное обучение на языке Python
1.5	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	Бэггинг, случайный лес как пример бэггинга. Бэггинг линейных классификаторов. Бустинг. Adaboost для ансамбля из простых деревьев (пней). Сравнение результатов бустинга для слабых и сильных моделей. Градиентный бустинг. Градиентный бустинг в задаче регрессии. Градиентный бустинг в задаче классификации. Градиентный бустинг над деревьями.	Машинное обучение на языке Python
1.6	Анализ текстовых данных	Представление текстовых данных в виде «мешка слов». Стоп-слова. Масштабирование данных с помощью tf-idf. Модель «мешка слов» для последовательностей из нескольких слов (n-грамм) Продвинутая токенизация, стемминг и лемматизация Моделирование тем и кластеризация документов. Латентное размещение Дирихле	Машинное обучение на языке Python
2. Лабораторные занятия			
2.1	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 помощью библиотеки Matplotlib. Линейные графики, диаграммы рассеяния, гистограммы, трехмерные графики. Знакомство с	Машинное обучение на языке Python

		библиотекой машинного обучения Scikit-Learn. Гиперпараметры и проверка качества модели	
2.2	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	Машинное обучение на языке Python
2.3.	Решение задачи регрессии	Разбор примера построения модели линейной регрессии для задачи предсказания велосипедного трафика Отбор и кодирование признаков. Визуальное сравнение общего и предсказанного моделью трафика. Проверка качества	Машинное обучение на языке Python
2.4	Решение задачи классификации.	Разбор примера построения модели логистической регрессии для задачи предсказания оттока клиентов мобильного оператора. Отбор и кодирование признаков. Проверка качества модели с помощью перекрёстной проверки.	Машинное обучение на языке Python
2.5	Древовидные модели: деревья решений, случайный лес	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.	Машинное обучение на языке Python
2.6	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	Разбор примера построения модели градиентного бустинга для задачи распознавания рукописных цифр из библиотеки MNIST.	Машинное обучение на языке Python
2.7	Анализ текстовых данных	Разбор примера построения модели анализа текстовых данных для задачи определения тональности киноотзывов.	Машинное обучение на языке Python

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Введение в машинное обучение. Основные определения и постановки задач.	4		4	10	18
2	Обзор основных необходимых библиотек языка Python	4		4	10	18
3	Построение и отбор признаков	4		4	10	18
4	Решение задачи регрессии	4		4	10	18
5	Решение задачи классификации.	4		4	10	18
6	Древовидные модели: деревья решений, случайный лес	4		4	10	18
7	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	4		4	10	18
8	Анализ текстовых данных	4		4	10	18
	Итого	32		32	80	

14. Методические указания для обучающихся по освоению дисциплины

Для лучшего усвоения материала студентам рекомендуется домашняя работа с конспектами лекций, презентациями, выполнение практических заданий для самостоятельной работы, выполнение лабораторных работ, использование рекомендованной литературы и методических материалов, в том числе размещенных на странице курса «Машинное обучение» на портале «Электронный университет ВГУ» https://edu.vsu.ru/course/view.php?id=3579, автор Каширина И.Л.
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 лекционных и практических занятия
При использовании дистанционных образовательных технологий и электронного обучения выполнять все указания преподавателей по работе на LMS-платформе, своевременно подключаться к online-занятиям, соблюдать рекомендации по организации самостоятельной работы

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины *(список литературы оформляется в соответствии с требованиями ГОСТ и используется общая сквозная нумерация для всех видов источников)*

а) основная литература:

№ п/п	Источник
1	Рашка, С. Python и машинное обучение: [Электронный ресурс] : руководство / С. Рашка ; пер. с англ. Логунова А.В.. — Электрон. дан. — Москва : ДМК Пресс, 2020. — 418 с. — Режим доступа: https://e.lanbook.com/book/100905
2	Коэльо, Л.П. Построение систем машинного обучения на языке Python [Электронный ресурс] / Л.П. Коэльо, В. Ричарт ; пер. с англ. Слинкин А. А.. — Электрон. дан. — Москва : ДМК Пресс, 2018. — 302 с. — Режим доступа: https://e.lanbook.com/book/82818
3	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс] / П. Флах. — Электрон. дан. — Москва : ДМК Пресс, 2019. — 400 с. — Режим доступа: https://e.lanbook.com/book/69955

б) дополнительная литература:

№ п/п	Источник
4	Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с. Материалы к книге: https://github.com/jakevdp/PythonDataScienceHandbook
5	Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. -СПб.: Питер, 2017. -336 с.
6	Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. -СПб.: Питер, 2017. -336 с.:
7	Шарден, Б. Крупномасштабное машинное обучение вместе с Python [Электронный ресурс] : учебное пособие / Б. Шарден, Л. Массарон, А. Боскетти ; пер. с анг. А. В. Логунова. — Электрон. дан. — Москва : ДМК Пресс, 2018. — 358 с. — Режим доступа: https://e.lanbook.com/book/10583
8	Вьюгин, В.В. Математические основы машинного обучения и прогнозирования [Электронный ресурс] : учебное пособие / В.В. Вьюгин. — Электрон. дан. — Москва : МЦНМО, 2013. — 304 с. — Режим доступа: https://e.lanbook.com/book/56397
9	Кук, Д. Машинное обучение с использованием библиотеки H2O [Электронный ресурс] / Д. Кук ; пер. с англ. Огурцова А.Б.. — Электрон. дан. — Москва : ДМК Пресс, 2018. — 250 с. — Режим доступа: https://e.lanbook.com/book/97353

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

№ п/п	Ресурс
10	А.Мюллер, С.Гвидо - Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными – 2017 электронный ресурс свободного доступа: https://owlweb.ru/wp-content/uploads/2017/06/a.myuller-s.gvido-vvedenie-v-mashinnoe-obuchenie-s-pomoshhyu-python.-rukovodstvo-dlya-specialistov-po-rabote-s-dannymi-2017.compressed-1.pdf материалы к книге: https://github.com/amueller/introduction_to_ml_with_python
11	Машинное обучение (курс лекций, К.В.Воронцов) http://www.machinelearning.ru/wiki/index.php?title=Машинное обучение (курс лекций, К.В.Воронцов)
12	https://www.kaggle.com/ - онлайн платформа для проектов в области науки о данных
13	UCI Machine Learning Repository — репозиторий наборов данных для машинного обучения

	- http://archive.ics.uci.edu/ml/
14	Ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. - http://machinelearning.ru
15	Открытый курс машинного обучения https://habr.com/company/ods/blog/322626/
16	Курс «Машинное обучение на языке Python» на портале «Электронный университет ВГУ», автор Каширина И.Л. *https://edu.vsu.ru/course/view.php?id=3579*

16. Перечень учебно-методического обеспечения для самостоятельной работы *(учебно-методические рекомендации, пособия, задачники, методические указания по выполнению практических (контрольных), курсовых работ и др.)*

№ п/п	Источник
1	Курс «Машинное обучение на языке Python» на портале «Электронный университет ВГУ», автор Каширина И.Л. https://edu.vsu.ru/course/view.php?id=3579
2	Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. -СПб.: Питер, 2017. -336 с.: Материалы к книге: *https://github.com/brinkar/real-world-machine-learning*
3	Каширина И.Л. Искусственные нейронные сети, Воронеж, Из-во ВГУ, 2014-96 с.

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

Дисциплина реализуется с применением электронного обучения и дистанционных образовательных технологий. Для организации самостоятельной работы обучающихся используется онлайн-курс электронный учебный Курс «Машинное обучение на языке Python» на портале «Электронный университет ВГУ», автор Каширина И.Л. https://edu.vsu.ru/course/view.php?id=3579 , размещенный на платформе Электронного университета ВГУ (LMS moodle), а также Интернет-ресурсы, приведенные в п.15в.

18. Материально-техническое обеспечение дисциплины:

Лекции: лекционная аудитория, учебная мебель, компьютер (ноутбук), мультимедийное оборудование (проектор, экран, средства звуковоспроизведения).

Практические и лабораторные занятия: специализированная аудитория, оснащенная учебной мебелью и персональными компьютерами для индивидуальной работы с возможностью подключения к сети «Интернет» (компьютерные классы, студии), мультимедийное оборудование (проектор, экран, средства звуковоспроизведения).

Самостоятельная работа: учебная мебель, компьютерный класс, компьютер с возможностью подключения к сети «Интернет» и платформе Электронного университета ВГУ (LMS moodle).

Программное обеспечение:
- ОС Windows 8 (10),
- интернет-браузер (Mozilla Firefox);
- ПО Adobe Reader;
- пакет стандартных офисных приложений для работы с документами, таблицами (МойОфис, LibreOffice);
- специализированное ПО, допускается демоверсия или виртуальный аналог ПО.)

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.	ПК-1	ПК-1.2	Тест
2.	Обзор основных необходимых библиотек языка Python	ПК-1	ПК-1.2	Тест
3	Построение и отбор признаков	ПК-1	ПК-1.2	Тест
4	Решение задачи регрессии	ПК-4	ПК-4.2	Задание для лабораторной работы
5	Решение задачи классификации.	ПК-4	ПК-4.2	Задание для лабораторной работы
6	Древовидные модели: деревья решений, случайный лес	ПК-4	ПК-4.2	Задание для лабораторной работы
7	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	ПК-5	ПК-5.1	Тест
8	Анализ текстовых данных	ПК-5	ПК-5.1	Задание для лабораторной работы
Промежуточная аттестация форма контроля - экзамен				*Перечень вопросов*

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Перечень заданий для лабораторных работ

Лабораторная работа № 1 (по теме линейная 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 отмасштабировав данные

- LinearRegression;
- Lasso с коэффициентом регуляризации, равным 0.01.

3) Проанализируйте качество получившихся моделей и сравните количество строго нулевых весов в них.

Лабораторная работа № 2 (по теме логистическая регрессия)

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
4) Вычислите значения метрик: recall, presision, F1-мера, AUC-ROC. Постройте ROC-кривую.

Лабораторная работа №3 (по теме Древовидные модели)

1) Разбейте предоставленный Вам преподавателем набор данных на обучающую и тестовую части в соотношении 8:2.
2) Проведите предобработку данных: заполнение пропусков, кодирование

3). Обучите, а затем провалидируйте на тестовых данных модели дерева решений, случайного леса, градиентного бустинга
4) Вычислите значения метрики точности, визуализируйте дерево решений.

Лабораторная работа № 4 (по теме Анализ текстов)

Практическое задание 2 посвящено работе с текстовыми данными и категориальными признаками и задачам бинарной классификации.
В рамках данного задания нужно решить задачу бинарной классификации для предсказания уровня заработной платы по тексту объявления о вакансии на примере набора данных с соревнования на Kaggle. Данные доступны по ссылке.
1) Разбейте получившуюся выборку на обучающую и контрольную в соотношении 70/30
2) Создайте текстовое описание объектов обучающей и контрольной выборок, объединив значения всех признаков каждого объекта выборки через символы пробела. После этого получите признаковое описание объектов, осуществив векторизацию получившихся текстов при помощи CountVectorizer, обучив его на обучающей выборке и применив на тестовой.
3) Обучите логистическую регрессию из модуля sklearn с параметрами по умолчанию на обучающей выборке:
4) Вычислите значения ROC-AUC, F-меры, а также постройте матрицу ошибок на тестовой выборке.
5) Отсортируйте веса признаков для модели. Какие слова из встречающихся в выборке имеют наибольшее/наименьшее влияние на значение целевой переменной? Проинтерпретируйте полученный результат.
6) Создайте текстовое описание объектов обучающей и контрольной выборок, объединив значения всех признаков каждого объекта выборки через символы пробела. После этого получите признаковое описание объектов, вычислив вектор tf-idf для каждого объекта помощи TfidfVectorizer, обучив его на обучающей выборке и применив на тестовой.
7) Заново обучите модель
8) Вычислите значения ROC-AUC, F-меры, а также постройте матрицу ошибок на контрольной выборке..
9). Сравните значения метрик из п. 8 со значениями, полученными в п. 4, и сравните соответствующие модели по качеству из работы.
11. Отсортируйте веса признаков для модели логистической регрессии из scikit-learn, полученной в п. 7. Какие слова из встречающихся в выборке имеют наибольшее/наименьшее влияние на значение целевой переменной? Проинтерпретируйте полученный результат.

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств:

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.

Для оценивания результатов обучения на экзамене 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.

Перечень вопросов к экзамену

1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
2. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.

3. Метрики качества алгоритмов регрессии и классификации.
4. Линейная регрессия. Простая многомерная регрессия. Регрессия с полиномиальными признаками. Методы регуляризации: Ridge, Lasso, ElasticNet.
5. Логистическая регрессия.
6. Деревья решений. Методы построения деревьев. Их регуляризация.
7. Композиции алгоритмов. Разложение ошибки на смещение и разброс.
8. Случайный лес, его особенности.
9. Градиентный бустинг, его особенности при использовании деревьев в качестве базовых алгоритмов.
10. Анализ текстов. Масштабирование данных с помощью tf-idf. Модель «мешка слов» для n-грамм.

Для оценивания результатов обучения на экзамене используется 4-балльная шала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».
Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания	Шкала оценок
Обучающийся в полной мере владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), сдал все лабораторные работы, среднее количество правильных ответов на вопросы тестов превышает 80%.	*Отлично*
Обучающийся владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), но не сдал одну лабораторную работу, среднее количество правильных ответов на вопросы тестов находится в диапазоне 60-80%.	*Хорошо*
Обучающийся демострирует неуверенное владение понятийным аппаратом данной области науки (теоретическими основами дисциплины), не сдал две практических или лабораторных работы, среднее количество правильных ответов на вопросы тестов находится в диапазоне40-60%.	*Удовлетвори-тельно*
Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не сдал более двух практических или лабораторных работы, среднее количество правильных ответов на вопросы тестов менее 40%.	*Неудовлетвори-тельно*

Студент, выполнивший в полном объеме программу курса, лабораторные работы зачтены, и имеющий посещаемость занятий 90% и более, на усмотрение преподавателя может быть освобожден от вопросов к экзамену. Итоговая оценка в этом случае, выставляется в соответствии только со средним баллом по тестам.

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

Компетенция-№ ПК-1 Способен проводить работы по обработке и анализу научно-технической информации результатов исследований

Вопросы с вариантами ответов

1.Что такое токенизация в машинном обучении?

a) разбиение анализируемого текста на отдельные слова (или устойчивые сочетания слов)
b) формирование словаря на основе текстового корпуса
c) алгоритм кодирования текстовых данных
d) выделение основы слов

2. Что такое категориальный признак?
   a) любой числовой признак
   b) нечисловой признак, значения которого можно упорядочить
   c) признак, значения которого берутся из некоторого конечного фиксированного набора
   d) текстовый признак

3. Выделяют такие этапы построения модели “мешок слов” (Выберите все подходящие ответы из списка)
   a) токенизация
   b) построение словаря
   c) моделирование тем
   d) формирование разреженной матрицы
   e) лемматизация

4. Что такое лемматизация?
   a) мера оценки важности слова в контексте документа, являющегося частью текстового корпуса
   b) процесс приведения слова к нормальной словарной форме
   c) процесс выделения основы слова
   d) числовое кодирование слов

5. К какому типу задач относится задача определения рейтинга фильма по текстам отзывов?
   a) задача регрессии
   b) задача классификации
   c) задача анализа тональности
   d) задача кластеризации

6. Из каких этапов состоит алгоритм построения дерева решений (Выберите все подходящие ответы из списка)?
   a) Выбор критерия точности прогноза
   b) Выбор критерия ветвления
   c) Выбор критерия прекращения ветвлений (останова)
   d) Определение "подходящих" размеров дерева (обрезка)
   e) Выбор структуры дерева

7. Что такое бэггинг?
   a) один из методов регуляризации в задачах машинного обучения
   b) процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов.
   c) процедура независимого построения композиции алгоритмов машинного обучения.
   d) метод получения векторного представления текстовых данных

8. С помощью метода деревьев решений возможно решение задач:
   a) только регрессии
   b) классификации и регрессии
   c) только классификации
   d) только кластеризации

9. Что такое AUC-ROC (Выберите все подходящие ответы из списка)?
   a) агрегированная характеристика качества бинарной классификации
   b) площадь под ROC-кривой

c) дисперсия ошибки регрессионной модели
d) критерий ветвления в алгоритме построения дерева решений
e) метрика точности в задачах машинного обучения

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

11. Классификация – это ..
a) отнесение объектов к одному из заранее известных классов
b) процесс формирования классов и отнесения объектов к одному из них
c) отнесение объектов к одному из заранее неизвестных классов
d) поиск кластеров в данных

Вопросы с кратким текстовым ответом

1. Дана таблица, содержащая истинное значение целевого признака Y и предсказанное значение Y'. Значение ошибки MAE равно ...

Объект	Нецелевые признаки	Y	Y'
A	,,,	1	0
B	,,,	2	2
C	,,,	3	2
D	,,,	4	5
E	,,,	5	3

Ответ: 1

2. Чему равно значение функционала ошибки MAE для для обучающей выборки из трех примеров, если реальные выходы модели Y= (0.1, 0.6, 0.8) , а целевые (требуемые) выходы D=(0, 0.5, 0.4) ? В ответе укажите 10*MAE.
Ответ: 2

3. Укажите максимально возможное значение метрики R^2
Ответ: 1

4. Чему равно значение ошибки MSE для выборки из двух обучающих векторов , если реальные выходы модели $Z =(z_1,z_2)= (0.27, 0.43)$, а целевые (правильные) выходы $Y =(y_1,y_2)= (0.29, 0.23)$? В ответе укажите 10000* MSE
Ответ: 202

Компетенция-№ ПК-4 Способен разрабатывать профессионально-ориентированные программные средства и приложения на основе интеллектуальных информационных технологий

Вопросы с вариантами ответов

1. Для чего нужна регуляризация в задачах машинного обучения ((Выберите все подходящие ответы из списка)?

a) чтобы предотвратить переобучение модели
b) чтобы обнулить коэффициенты при незначимых признаках
c) чтобы добавить в целевую функцию штраф, который бы наказывал модель за слишком большие коэффициенты
d) чтобы составить план проведения исследования
e) для формирования документации о ходе исследования

2. Что такое Эластичная сеть?

a) Метод регуляризации, сочетающий использование Гребневой регрессии и Лассо
b) метод нормализации данных
c) метод разделения выборки на обучающую и тестовую
d) метод поиска параметров регрессии путем перебора
e) метод формирования плана исследования

3. Как называется задача, направленная на предсказание значения той или иной непрерывной числовой величины для входных данных?

a) Регрессия
b) Классификация
c) Кластеризация
d) Токенизация

4. Какое из перечисленных значений НЕ может принимать метрика R^2

a) 0
b) 1
c) 1.1
d) -0.1

5. Множество примеров, используемое для конструирования модели, называется...

a) обучающим множеством
b) тестовым множеством
c) валидационным множеством
d) конструктивным множеством

6. Множество примеров, используемое для проверки работы сконструированной модели, называется...

a) обучающей выборкой
b) тестовой выборкой
c) синтетической выборкой
d) проверочной выборкой

7. Переобучение - это …

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

8. Что из перечисленного является признаком переобучения модели

a) когда ошибка на обучающей выборке низкая, а на тестовой высокая

b) когда ошибка на обучающей выборке высокая, а на тестовой низкая
c) когда ошибка на обучающей и на тестовой выборке низкая
d) когда ошибка на обучающей и на тестовой выборке высокая

9. Решение задачи регрессии ...

a) Является решением задачи "обучения без учителя"
b) Возможно без обучающей выборки данных
c) Требует некоторой обучающей выборки данных
d) Не требует предварительной подготовки обучающих данных

10. С помощью какого метода определяют коэффициенты уравнения линейной регрессии:

a) метода наименьших квадратов
b) симплекс-метода
c) метода Кронекера
d) метода включений и исключений

11. Согласно методу наименьших квадратов, в качестве оценок коэффициентов регрессии следует использовать величины, которые минимизируют сумму квадратов отклонений:

a) фактических значений зависимой переменной от ее среднего значения
b) расчетных значений зависимой переменной от ее среднего значения
c) фактических значений зависимой переменной от ее расчетных значений
d) максимального значения зависимой переменной от ее среднего значения

12. Что такое линейная регрессия?

a) это произвольная функциональная зависимость, которая позволяет прогнозировать изменения непрерывных числовых параметров;
b) модель зависимости непрерывной переменной y от объясняющих ее факторов, в которой функция зависимости является линейной
c) модель зависимости категориальной переменной y от объясняющих ее факторов, в которой функция зависимости является линейной
d) один из методов классификации

13. 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) соотношения между зависимыми и независимыми показателями и переменными в наглядной и понятной человеку форме;
d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.

Вопросы с кратким текстовым ответом

1. Дана матрица ошибок алгоритма классификации. Вычислите precision (используйте в качестве разделителя целой и дробной части точку)

		Истинный класс	
		1	-1
Предсказанный класс	1	30	10
	-1	20	20

Ответ: 0.75

2. Дана матрица ошибок алгоритма классификации. Вычислите recall (используйте в качестве разделителя целой и дробной части точку)

		Истинный класс	
		1	-1
Предсказанный класс	1	30	10
	-1	20	20

Ответ: 0.6

3. Дана матрица ошибок алгоритма классификации. Вычислите accuracy (используйте в качестве разделителя целой и дробной части точку)

		Истинный класс	
		1	-1
Предсказанный класс	1	30	10
	-1	20	20

Ответ: 0.625

4. Дана матрица ошибок, построенная по результатам работы некоторого алгоритма классификации. Общая точность (accuracy) равна...(используйте в качестве разделителя целой и дробной части точку)

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

Ответ: 0.5

5. Дана матрица ошибок, построенная по результатам работы некоторого алгоритма классификации. Общая точность (accuracy) равна... (используйте в качестве разделителя целой и дробной части точку)

		Истинный класс	
		1	-1
Предсказанный класс	1	50	20
	-1	5	50

Ответ: 0.8

6. Чему равно значение функционала ошибки MAE для для обучающей выборки из трех примеров, если реальные выходы модели Y= (0.2, 0.5, 0.8) , а целевые (требуемые) выходы D=(0, 0.5, 0.7) ? В ответе укажите 10*MAE.
Ответ: 1

Компетенция-№ ПК-5 Способен совершенствовать и разрабатывать новые методы, модели, алгоритмы, технологии работы с большими данными

Вопросы с вариантами ответов

1. Accuracy - это...
   a) доля правильных ответов алгоритма классификации
   b) величина ошибки алгоритма регрессии
   c) доля неправильных ответов алгоритма классификации
   d) метрика в задаче кластеризации

2. Выберите все подходящие ответы из списка

a) Высокая accuracy не всегда говорит о качестве модели в случае неравномерного распределения классов.
b) Precision и recall не могут быть равны
c) Precision может быть равен нулю
d) Precision, recall и accuracy могут быть равны 1 в случае безошибочной классификации
e) Precision всегда больше recall

3. В задаче медицинской диагностики класс "1" -это больные пациенты, класс "-1"- здоровые. Тогда precision - это…

a) доля больных среди пациентов, распознанных алгоритмом как больные
b) доля здоровых среди пациентов, распознанных алгоритмом как больные
c) доля пациентов, распознанных алгоритмом как больные, среди больных
d) доля пациентов, распознанных алгоритмом как здоровые, среди больных

4. В задаче медицинской диагностики класс "1" -это больные пациенты, класс "-1"- здоровые. Тогда recall - это…

a) доля больных среди пациентов, распознанных алгоритмом как больные
b) доля здоровых среди пациентов, распознанных алгоритмом как больные
c) доля пациентов, распознанных алгоритмом как больные, среди действительно больных
d) доля пациентов, распознанных алгоритмом как здоровые, среди больных

5. Задача классификации сводится к …

a) нахождению частых зависимостей между объектами или событиями;
b) определению класса объекта по его характеристикам;
c) определению по известным характеристикам объекта значения некоторого его непрерывного параметра
d) поиску независимых групп и их характеристик в всем множестве анализируемых данных

6. Допустим, установили сканер отпечатков пальцев на вход в ВГУ, теперь любой студент может приложить палец и попасть внутрь. Однако сканер иногда допускает ошибки. Служба охраны переживает и очень не хочет, чтобы посторонние люди были распознаны, как студенты. Какую метрику необходимо максимизировать в таком случае? (Будем считать, что класс 1- студенты ВГУ, класс "-1" - не студенты)

a) accuracy
b) Recall
c) Precision
d) R^2

7. Классификация относится к стратегии:

a) обучения без учителя
b) обучения с учителем
c) оба ответа неверны
d) оба ответа верны

8. Что такое логистическая регрессия?

a) метод классификации, предсказывающий целевой класс на базе количественных признаков
b) метод регрессии, предсказывающий значение вещественной целевой переменной
c) метод построения ансамбля моделей машинного обучения

d) метод кластеризации

9. Что такое ROC-кривая (выберите все верные ответы)?

a) кривая, показывающая сходимость метода в зависимости от номера итерации
b) график функции потерь (ошибок)
c) график, позволяющий оценить качество бинарной классификации
d) зависимость количества верно классифицированных примеров положительного класса от количества неверно классифицированных примеров отрицательного класса
e) график изменения параметра регуляризации

10. Какие метрики используются в задачах оценки качества классификации (выберите все верные ответы)?

a) accuracy
b) Recall
c) Precision
d) MSE
e) R^2

Вопросы с кратким текстовым ответом

1. Некоторый классификатор относит все объекты к классу 1. Допустим, что выборка состоит из 50 объектов: 20 из них принадлежат классу -1, а 30 из них действительно принадлежат классу 1. Общая точность (accuracy) равна... (используйте в качестве разделителя целой и дробной части точку)
Ответ: 0.6

2. Некоторый классификатор относит все объекты к классу 1. Допустим, что выборка состоит из 50 объектов: 20 из них принадлежат классу -1, а 30 из них действительно принадлежат классу 1. Метрика precision равна... (используйте в качестве разделителя целой и дробной части точку)
Ответ: 0.6

3. Некоторый классификатор относит все объекты к классу 1. Допустим, что выборка состоит из 50 объектов: 20 из них принадлежат классу -1, а 30 из них действительно принадлежат классу 1. Метрика recall равна... (используйте в качестве разделителя целой и дробной части точку)

Ответ: 1

4. Некоторый классификатор относит все объекты к классу -1. Допустим, что выборка состоит из 50 объектов: 20 из них действительно принадлежат классу -1, а 30 из них принадлежат классу 1. Общая точность (accuracy) равна... (используйте в качестве разделителя целой и дробной части точку)

Ответ: 0.4

Критерии и шкалы оценивания заданий ФОС:

Для оценивания выполнения заданий используется балльная шкала:

1) закрытые задания (тестовые с вариантами ответов, средний уровень сложности):

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

2) открытые задания (тестовые с кратким текстовым ответом, повышенный уровень сложности):

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

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