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

УТВЕРЖДАЮ
Заведующий кафедрой экономической теории
и мировой экономики
_____ д.э.н., проф. Т.Н.Гоголева
20.04.2021 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.17 Прикладное машинное обучение на языке Python

1. Код и наименование направления подготовки: 38.04.01 «Экономика»

2. Профиль подготовки: Экономика рынков

3. Квалификация выпускника: бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: экономической теории и мировой экономики

6. Составители программы: Костылева В.И., ст. преподаватель кафедры экономической теории и мировой экономики

7. Рекомендована: Научно-методическим советом экономического факультета ВГУ от 15.04.2021 г., протокол №4

8. Учебный год: 2024/2025 **Семестр:** 7

9. **Цели и задачи учебной дисциплины**

Целями освоения учебной дисциплины являются:

- теоретическая и практическая подготовка студентов к разработке приложений на языке Python;
- решения широкого круга задач машинного обучения.

Задачи учебной дисциплины:

-сформировать теоретические знания и практические навыки в области использования возможностей Python для анализа данных, качественной визуализации и алгоритмов машинного обучения;
- приобрести навыки работы с библиотеками Scikit-Learn и TensorFlow;
- приобрести опыт решения производственных задач.

10. Место учебной дисциплины в структуре ООП: дисциплина «Прикладное машинное обучение на языке Python» относится к вариативной части блока Б1.

Дисциплина является последующей для таких дисциплин как «Информационные технологии в экономике», «Информационные системы в экономике». Знания, полученные в ходе изучения дисциплины, используются для написания ВКР.

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	Коды	Индикаторы	Планируемые результаты обучения
ПК-4	Способен использовать для решения аналитических и прикладных задач современные технические средства и информационные технологии	ПК–4.3	Применяет современные программные средства хранения и обработки социально-экономической информации для решения аналитических и исследовательских задач	*Знать*: - методы предварительной обработки - 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*: - навыками построения и проверки качества моделей машинного обучения; - навыками интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов;
		ПК–4.4	Использует современное ПО для решения экономических задач оценки риска	*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; – использовать библиотеки языка Python для построения моделей машинного обучения; *владеть* (иметь навык(и)): - использования библиотек языка Python для построения систем, обучающихся по прецедентам.

12. Объем дисциплины в зачетных единицах/час. — 3 ЗЕТ/ 108 час.
Форма промежуточной аттестации зачет с оценкой

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость	
		Всего	По семестрам
			№ семестра 7
Аудиторные занятия		48	48
в том числе:	лекции	24	24
	практические	24	24
Самостоятельная работа		60	60
Форма промежуточной аттестации		-	-
Итого:		108	108

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. Обзор основных необходимых библиотек языка Python.	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	-
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.	
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	-
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	-
1.5	Ансамбли моделей Бэг-	Бэггинг, случайный лес как пример бэггинга. Бэггинг	-

	гинг, бустинг, градиентный бустинг	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.	
1.6	Анализ текстовых данных	Представление текстовых данных в виде «мешка слов». Стопслова. Масштабирование данных с помощью tf-idf. Модель «мешка слов» для последовательностей из нескольких слов (nграмм) Продвинутая токенизация, стемминг и лемматизация Моделирование тем и кластеризация документов. Латентное размещение Дирихле	-
	2. Практические занятия		
2.1	Введение в машинное обучение. Основные определения и постановки задач. 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 библиотекой машинного обучения Scikit-Learn. Гиперпараметры и проверка качества модели. 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.	-
2.2	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.	-
2.3	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.	-
2.4	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.	-
2.5	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	Разбор примера построения модели градиентного бустинга для задачи распознавания рукописных цифр из библиотеки MNIST.	-
2.6	Анализ текстовых данных	Разбор примера построения модели анализа текстовых данных для задачи определения тональности киноотзывов.	-

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (количество часов)			
		Лекции	Практические	Самостоятельная работа	Всего
1	Введение в машинное	4	4	10	18

	обучение. Основные определения и постановки задач. Обзор основных необходимых библиотек языка Python.				
2	Решение задачи регрессии	4	4	10	18
3	Решение задачи классификации.	4	4	10	18
4	Древовидные модели: деревья решений, случайный лес	4	4	10	18
5	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	4	4	10	18
6	Анализ текстовых данных	4	4	10	18
	Итого:	24	24	60	108

14. Методические указания для обучающихся по освоению дисциплины:

Успешное изучение курса требует от студентов посещения лекции и систематического конспектирования учебного материала. освоение и осмысление терминологии изучаемой дисциплины. Материалы лекционных занятии следует своевременно подкреплять проработкой соответствующих разделов в учебниках, учебных пособиях, научных статьях и монографиях, в соответствии со списком основной и дополнительной литературы. Дополнительная проработка изучаемого материала проводится во время практических занятий, в ходе которых анализируется и закрепляется основные знания, полученные по дисциплине.

На практических занятиях приветствуется активное участие в обсуждении конкретных ситуации, способность на основе полученных знании находить наиболее эффективные решения поставленных проблем, уметь находить полезный дополнительный материал по тематике занятии.

Самостоятельная работа предполагает изучение теоретического материала, написание программ по темам, изученным на лекционных и практических занятиях.

Изучение дисциплины предполагает наличие текущих и промежуточной аттестаций.

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	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

б) дополнительная литература:

№ п/п	Источник
1	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс] / П. Флах. — Электрон. дан. — Москва : ДМК Пресс, 2015. — 400 с. — Режим доступа: https://e.lanbook.com/book/69955
2	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
1	Электронно-библиотечная система "Лань" - https://e.lanbook.com/

2	Электронно-библиотечная система "Университетская библиотека online" - https://lib.vsu.ru/url.php?url=http://biblioclub.ru/
3	Образовательный портал "Электронный университет ВГУ". - https://edu.vsu.ru

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. -СПб.: Питер, 2017. -336 с.: Материалы к книге: https://github.com/brinkar/real-world-machine-learning
2	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
3	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-mashinnoeobuchenie-s-pomoshhyu-python.-rukovodstvo-dlya-specialistov-po-rabote-s-dannymi2017.compressed-1.pdf

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

Реализация учебной дисциплины предполагает применение дистанционных образовательных технологий (работу на образовательном портале «Электронный университет ВГУ» https://edu.vsu.ru).

18. Материально-техническое обеспечение дисциплины:

Учебные аудитории для проведения учебных занятий (лекционных, практических), оснащенных оборудованием и техническими средствами обучения: специализированная мебель, проектор, экран для проектора, компьютер с возможностью подключения к сети "Интернет", проводной микрофон, комплект активных громкоговорителей

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. Обзор основных необходимых библиотек языка Python.	ПК – 4	ПК – 4.3 ПК – 4.4	*КИМ №1*
2.	Решение задачи регрессии	ПК – 4	ПК – 4.3 ПК – 4.4	*КИМ №1*
3.	Решение задачи классификации.	ПК – 4	ПК – 4.3 ПК – 4.4	*КИМ №1*
4.	Древовидные модели: деревья решений, случайный лес	ПК – 4	ПК – 4.3 ПК – 4.4	*КИМ №1*
5.	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	ПК – 4	ПК – 4.3 ПК – 4.4	*КИМ №1*
6.	Анализ текстовых данных	ПК – 4	ПК – 4.3 ПК – 4.4	
Промежуточная аттестация форма контроля – зачёт с оценкой				*КИМ №2*

Уровень знаний студента определяется оценками: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Полный, исчерпывающий, аргументированный ответ на все вопросы и задания КИМа. Ответы должны отличаться логической последовательностью, четкостью в выражении мыслей и обоснованностью выводов, демонстрирующих знание источников основной и дополнительной литературы, понятийного аппарата и умения ими пользоваться при ответе.	*Повышенный уровень*	*Отлично*
Полный, исчерпывающий, аргументированный ответ на вопросы и задания КИМа. Ответы должны отличаться логичностью, четкостью, знанием понятийного аппарата и литературы по КИМу при незначительных упущениях при ответах.	*Базовый уровень*	*Хорошо*
Неполных и слабо аргументированный ответ, демонстрирующих общее представление и элементарное понимание существа поставленных вопросов, понятийного аппарата и обязательной литературы.	*Пороговый уровень*	*Удовлетворительно*
Обучающийся демонстрирует незнании и непонимании студентом существа задаваемых вопросов и задания КИМа. При выставлении неудовлетворительной оценки, преподаватель должен объяснить студенту недостатки его ответа.	–	*Неудовлетворительно*

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:

КИМ №1

Пример практического задания для текущей аттестации

Практическое задание 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.

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.

10) 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.

Критерии оценки:

Оценка «отлично» выставляется студенту, если студент показал знание теоретического материала, проведен анализ,данные актуальны, сделаны выводы.

Оценка «хорошо» выставляется студенту, если студент показал знание теоретического материала, проведен анализ, данные актуальны, сделаны выводы, отдельные неточности.

Оценка «удовлетворительно» выставляется если студент показал знание теоретического материала, данные не актуальны.

Оценка «неудовлетворительно» выставляется студенту, если не продемонстрировано владение теоретическим материалом, отсутствует анализ, представленны рассуждения общего характера.

20.2. Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств:

КИМ №2

Перечень вопросов к зачету

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.

Критерии оценки:

- оценка «отлично» выставляется студенту, если дан полный, исчерпывающий, аргументированный ответ на все вопросы и задания КИМа. Ответы должны отличаться логической последовательностью, четкостью в выражении мыслей и обоснованностью выводов, демонстрирующих знание источников основной и дополнительной литературы, понятийного аппарата и умения ими пользоваться при ответе;
- оценка «хорошо», если дан полный, исчерпывающий, аргументированный ответ на вопросы и задания КИМа. Ответы должны отличаться логичностью, четкостью, знанием понятийного аппарата и литературы по КИМу при незначительных упущениях при ответах;
- оценка «удовлетворительно» выставляется при неполных и слабо аргументированный ответ, демонстрирующих общее представление и элементарное понимание существа поставленных вопросов, понятийного аппарата и обязательной литературы;
- оценка «неудовлетворительно», если обучающийся демонстрирует незнании и непонимании студентом существа задаваемых вопросов и задания КИМа. При выставлении неудовлетворительной оценки, преподаватель должен объяснить студенту недостатки его ответа.