

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
**«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ им. В.Г.ШУХОВА»**
(БГТУ им. В.Г. Шухова)

УТВЕРЖДАЮ
Директор института энергетики,
информационных технологий и
управляющих систем

Белоусов А.В.

2021 г.

РАБОЧАЯ ПРОГРАММА
дисциплины

Методы анализа данных

направление подготовки:

10.05.03 Информационная безопасность автоматизированных систем

Специализация программы:

Безопасность открытых информационных систем

Квалификация

Специалист по защите информации

Форма обучения

очная


Институт энергетики, информационных технологий и управляющих систем

Кафедра Программного обеспечения вычислительной техники и
автоматизированных систем

Белгород 2021

Рабочая программа составлена на основании требований:

- Федерального государственного образовательного стандарта высшего образования – специалитет по специальности 10.05.03 Информационная безопасность автоматизированных систем, утвержденного приказом Минобрнауки России от 26.11.2020 №1457
- учебного плана, утвержденного ученым советом БГТУ им. В.Г. Шухова в 2021 году.

Составитель: к.ф.-м.н.  (Зуев С.В.)
(ученая степень и звание, подпись) (инициалы, фамилия)

Рабочая программа обсуждена на заседании кафедры

« 14 » 05 2021 г., протокол № 8

Заведующий кафедрой: к.т.н., доцент  (Поляков В.М.)
(ученая степень и звание, подпись) (инициалы, фамилия)


Рабочая программа согласована с выпускающей кафедрой программного обеспечения вычислительной техники и автоматизированных систем
(наименование кафедры/кафедр)

Заведующий кафедрой: к.т.н., доцент  (Поляков В.М.)
(ученая степень и звание, подпись) (инициалы, фамилия)

« 14 » 05 2021 г.

Рабочая программа одобрена методической комиссией института

« 20 » 05 2021 г., протокол № 9

Председатель к.т.н., доцент  (Семернин А.Н.)
(ученая степень и звание, подпись) (инициалы, фамилия)

1. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Категория (группа) компетенций	Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Наименование показателя оценивания результата обучения по дисциплине
Общепрофессиональные компетенции	ОПК-5.3 Способен осуществлять контроль обеспечения информационной безопасности и проводить верификацию данных в открытых информационных системах	ОПК-5.3.1 Осуществляет контроль обеспечения информационной безопасности в открытых информационных системах	<p>Знания</p> <ul style="list-style-type: none"> - критериев обеспечения безопасности в открытых ИС; - методов автоматизированного и автоматического анализа данных в открытых ИС; - методов интеллектуального анализа данных. <p>Умения</p> <ul style="list-style-type: none"> - применять методы анализа данных для решения задач практической деятельности. <p>Навыки</p> <ul style="list-style-type: none"> - использования системных журналов в качестве источников данных, - представления результатов контроля
		ОПК-5.3.2 Проводит верификацию данных в открытых информационных системах	<p>Знания</p> <ul style="list-style-type: none"> - процедур верификации данных; - методов интеллектуального анализа целостности данных. <p>Умения</p> <ul style="list-style-type: none"> - разрабатывать системные программные средства верификации данных; - разрабатывать программные анализаторы больших данных открытых ИС. <p>Навыки</p> <ul style="list-style-type: none"> - выбора методов верификации данных.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

1. Компетенция ОПК-5.3 Способен осуществлять контроль обеспечения информационной безопасности и проводить верификацию данных в открытых информационных системах

Данная компетенция формируется следующими дисциплинами.

Стадия	Наименования дисциплины
1	Методы анализа данных
2	Информационная безопасность открытых информационных систем
3	Администрирование информационных систем и служб

3. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины составляет 4 зач. единицы, 144 часа.

Форма промежуточной аттестации _____ зачет _____
(экзамен, дифференцированный зачет, зачет)

Вид учебной работы	Всего часов	Семестр № 4
Общая трудоемкость дисциплины, час	144	144
Контактная работа (аудиторные занятия), в т.ч.:	71	71
лекции	34	34
лабораторные	34	34
практические		
групповые консультации в период теоретического обучения и промежуточной аттестации	3	3
Самостоятельная работа студентов, включая индивидуальные и групповые консультации, в том числе:	73	73
Курсовой проект		
Курсовая работа		
Расчетно-графическое задание		
Индивидуальное домашнее задание		
Самостоятельная работа на подготовку к аудиторным занятиям (лекции, лабораторные занятия)	73	73
Зачет (4 семестр)	-	-

4. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Наименование тем, их содержание и объем Курс 2 Семестр 2

№ п/п	Наименование раздела (краткое содержание)	Объем на тематический раздел по видам учебной нагрузки, час			
		Лекции	Практические занятия	Лабораторные занятия	Самостоятельная работа на подготовку к аудиторным
1. Программные и математические средства анализа данных					
	Язык программирования Python. Типизация данных. Базовые конструкции. Элементы функционального и объектно-ориентированного программирования. Модули, их создание и использование. Средства визуализации библиотеки matplotlib. Элементы линейной алгебры, теории вероятностей и математической статистики, их реализация в функциях Python.	8	-	6	12
2. Сбор и обработка данных					
	Способы сбора данных в открытых системах: парсинг серверных логов, данных с веб-страниц. Обработка данных средствами pandas. Общее понятие о препроцессинге данных. Исследование и управление данными, фильтры, изменение размерностей. Понятие машинного обучения. Метрические и байесовские методы.	14	-	16	31
3. Модели и алгоритмы интеллектуального анализа данных					
	Регрессионный анализ. Деревья принятия решений. Нейронные сети. Кластеризация. Алгоритмы обработки естественного языка.	12	-	12	30
	ВСЕГО	34	-	34	73

4.2. Содержание практических (семинарских) занятий

Учебным планом не предусмотрены.

4.3. Содержание лабораторных занятий

№ п/п	Наименование раздела дисциплины	Тема лабораторного занятия	К-во часов	Самостоятельная работа на подготовку к лабораторным занятиям
семестр № 4				
1	Программные и математические средства анализа	Простые задачи на программирование на Python.	2	3
2		Функциональное и объектно-	4	6

	данных	ориентированное программирование на Python		
5	Сбор и обработка данных	Парсеры данных	4	6
6		Преоброессинг данных и трансформация размерности	4	6
7		Метод к ближайших соседей	2	3
8		Байесовский классификатор	2	3
9	Модели и алгоритмы интеллектуального анализа данных	Множественная и логистическая регрессии	4	6
10		Алгоритм ID3 построения дерева принятия решений	2	3
11		Простая распознающая нейронная сеть	4	6
12		Алгоритмы кластеризации	4	6
13		Анализатор текста на основе би- и триграмм	2	3
ИТОГО:			34	51
ВСЕГО:				85

4.4. Содержание курсового проекта/работы

Учебным планом не предусмотрено

4.5. Содержание расчетно-графического задания, индивидуальных домашних заданий

Учебным планом не предусмотрены

5. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

5.1. Реализация компетенций

1 Компетенция ОПК-5.3 Способен осуществлять контроль обеспечения информационной безопасности и проводить верификацию данных в открытых информационных системах

Наименование индикатора достижения компетенции	Используемые средства оценивания
ОПК-5.3.1 Осуществляет контроль обеспечения информационной безопасности в открытых информационных системах	Собеседование, устный опрос
ОПК-5.3.2 Проводит верификацию данных в открытых информационных системах	Прием лабораторных работ, устный опрос, дифференцированный зачет

5.2. Типовые контрольные задания для промежуточной аттестации

5.2.1. Перечень контрольных вопросов (типовых заданий) для дифференцированного зачета / зачета

№ п/п	Наименование раздела дисциплины	Содержание вопросов (типовых заданий)
1	Программные и математические средства анализа данных	<ol style="list-style-type: none"> 1. Как работает ветвление в Python? 2. Опишите синтаксис и назначение функций map, reduce. 3. Опишите способ построения функций с открытыми аргументами и аргументами по ключам. 4. Как построить цикл while? 5. Какие вы знаете типы данных в Python? 6. Какие функции для построения графических объектов из библиотеки matplotlib вы изучали? 7. Опишите создание и использование класса в Python. 8. Что такое вектор? Как представляются данные в подготовленном для анализа виде? 9. Статистическое распределение. Плотность вероятности. Функция распределения. Обратная функция распределения. 10. Уровень значимости и мощность проверки. Как связаны эти термины с ошибками 1-го и 2-го рода?
2	Сбор и обработка данных	<ol style="list-style-type: none"> 1. Какие вам известны источники данных? Что за данные в них можно получить? 2. Что такое датасет? Объясните понятие метки в датасете. 3. Какие вы знаете средства для извлечения информации из веб-страниц? 4. Как можно производить поиск по веб-странице, структура которой неизвестна? 5. Как производить поиск и фильтрацию в .xml файлах? 6. Очистка данных: методы, приемы, средства. 7. Препроцессинг: параметризация и изменение размерности с уменьшением и без уменьшения информационного содержания. 8. Бутстрапирование – назначение, способ реализации. 9. Метод k ближайших соседей: назначение, способ реализации, достоинства и недостатки. 10. Идея и назначение наивного байесовского классификатора. 11. Смысл и способ применения сглаживания в классификаторе по методу наивного Байеса. 12. Априорное и апостериорное распределения.
3	Модели и алгоритмы интеллектуального анализа данных	<ol style="list-style-type: none"> 1. Как связано понятие регрессии с решением систем линейных алгебраических уравнений? 2. Что является метрикой качества множественной регрессии? 3. В чем достоинства и недостатки классификатора на основе логистической регрессии? 4. На чем основан алгоритм ID3 построения решающего дерева?

		<p>5. Что такое функция активации? Какие требования к ней предъявляются?</p> <p>6. Какие вы знаете алгоритмы обучения нейронной сети?</p> <p>7. Что такое переобучение и недообучение?</p> <p>8. Опишите известные вам алгоритмы кластеризации.</p> <p>9. На основе какого предположения осуществляется обратное распространение ошибки при обучении?</p> <p>10. В чем основная задача тематического моделирования?</p> <p>11. Какие вы знаете алгоритмы автоматической генерации текста?</p>
--	--	---

5.2.2. Перечень контрольных материалов для защиты курсового проекта/ курсовой работы

Учебным планом не предусмотрены

5.3. Типовые контрольные задания (материалы) для текущего контроля в семестре

Наименование индикатора достижения компетенции	Типовые контрольные задания
ОПК-5.3.1 Осуществляет контроль обеспечения информационной безопасности в открытых информационных системах	<p>Написать анализатор логов для операционной системы семейства Linux.</p> <p>Выявить аномалию в данных, заданных в кейсе.</p> <p>Осуществить кластеризацию в наборе векторов состояния компьютерной системы.</p> <p>Напишите системного демона Linux, осуществляющего анализ поведения пользователя с помощью решающего дерева.</p> <p>Разработайте анализатор текстов коротких сообщений на предмет наличия в них заданных сообщений.</p> <p>Напишите ИНС, распознающую активность бота на узле.</p> <p>Предложите способ быстрого распознавания DDoS атаки на сервер.</p>
ОПК-5.3.2 Проводит верификацию данных в открытых информационных системах	<p>Построить компаратор данных на основе трансформации размерностей.</p> <p>Создать программного агента, выявляющего аномальные данные методом k ближайших соседей.</p> <p>Написать фильтр фрода для корпоративного чата.</p> <p>Разработать аутентификатор с функцией интеллектуальной проверки подлинности.</p> <p>Написать поточный анализатор данных на основе авторегрессии.</p>

Тестовые задания

№ п/п	Наименование раздела дисциплины	Содержание вопросов (типовых заданий)
1	Программные и математические средства анализа	<p>Задание 1</p> <p>Пустой список в качестве булева выражения интерпретируется в Python как ____</p>

№ п/п	Наименование раздела дисциплины	Содержание вопросов (типовых заданий)
	данных	<p><i>Выберите один из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) True 2) None 3) False 4) Не интерпретируется (выдается ошибка). <p><u>Задание 2</u> Конструкция else вместе с for _____</p> <p><i>Выберите один из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) не используется 2) может использоваться всегда для проверки возможности итерации 3) может использоваться только на одном уровне с if 4) используется только вместе с brake <p><u>Задание 3</u> Строка T = list(map(lambda x: x[::-1], s))</p> <p><i>Выберите один из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) Работать не будет 2) Будет работать со строкой s 3) Будет работать со списком s 4) Будет работать со списком или кортежем s <p><u>Задание 4</u> Выберите векторы из этих объектов</p> <p><i>Выберите несколько из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) [-1,2.2,0,-3] 2) [-3,-1,2,4] 3) ['a','b','c','d'] 4) [100,200,300,'d',400] <p><u>Задание 5</u> p-значение _____</p> <p><i>Выберите несколько из 5 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) помогает оценить вероятность ошибки 1-го рода 2) помогает оценить вероятность ошибки 2-го рода 3) есть вероятность ошибки 1-го рода 4) есть вероятность не совершить ошибку 1-го рода 5) есть вероятность не совершить ошибку 2-го рода
2	Сбор и обработка данных	<p><u>Задание 1</u> Данными является _____</p> <p><i>Выберите несколько из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) текст на естественном языке 2) изображение 3) таблица, форма значений в столбце единая 4) таблица, форма значений в столбце произвольная <p><u>Задание 2</u> Выбирать все теги, указанный атрибут которых имеет определенное значение, позволяет _____</p> <p><i>Выберите один из 3 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) BeautifulSoup 2) requests 3) re <p><u>Задание 3</u> <i>Укажите соответствие для всех вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) размерность уменьшается вместе с информацией 2) размерность изменяется, информация не меняется <p>___ Трансформация размерности ___ Метод главных компонент</p> <p><u>Задание 4</u> Как выбирается количество соседей в методе k ближайших соседей?</p>

№ п/п	Наименование раздела дисциплины	Содержание вопросов (типовых заданий)
		<p><i>Выберите один из 5 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) случайно 2) оптимально 3) перебором по возрастанию 4) перебором по убыванию 5) по количеству меток <p><u>Задание 5</u> Почему байесовский классификатор называют наивным? <i>Выберите один из 3 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) из-за простоты 2) потому что предполагается независимость вхождений слов в сообщение 3) потому что у него низкое качество классификации
3	Модели и алгоритмы интеллектуального анализа данных	<p><u>Задание 1</u> Основным отличием линейной регрессии от других методов является _____</p> <p><i>Выберите один из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) сложность реализации 2) отсутствие машинного обучения 3) предположение о линейности закономерностей 4) использование алгоритма решения СЛАУ <p><u>Задание 2</u> Логистическая регрессия _____</p> <p><i>Выберите несколько из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) делит данные на два класса 2) осуществляет линейное разделение 3) делит данные на произвольное число классов 4) осуществляет кластеризацию <p><u>Задание 3</u> Главным ориентиром в построении дерева принятия решений в алгоритме ID3 является _____</p> <p><i>Выберите один из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) расстояние 2) энтропия 3) точное соответствие 4) нет ориентира – случайный выбор <p><u>Задание 4</u> Восходящая кластеризация _____</p> <p><i>Выберите один из 3 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) не может выдать пустой кластер 2) не может выдать один кластер 3) не может выдать число кластеров, равное числу экземпляров <p><u>Задание 5</u> Тематическое моделирование _____</p> <p><i>Выберите несколько из 4 вариантов ответа:</i></p> <ol style="list-style-type: none"> 1) позволяет отнести текст к одной из тем 2) позволяет отнести текст к нескольким темам 3) позволяет отнести текст к нескольким темам и определить вероятности соответствия 4) позволяет автоматически продолжить текст

5.4. Описание критериев оценивания компетенций и шкалы оценивания

Промежуточная аттестация предусмотрена в форме зачета, используется следующая шкала оценивания: не зачтено, зачтено.

Критериями оценивания достижений показателей являются:

Наименование показателя оценивания результата обучения по дисциплине	Критерий оценивания
Знания	Знание терминологии
	Знание основных принципов и методов анализа данных
	Знания о применимости методов анализа данных в сфере профессиональной деятельности
	Объем и полнота освоенного материала
Умения	Способность уместно применять терминологию
	Способность сопоставить метод решения типовой задаче анализа данных
	Способность аргументировать предложение методов анализа данных
	Способность уверенно ориентироваться в программных кодах на языке Python
	Способность провести анализ данных и представить результат
Навыки	Программирования на Python
	Подготовки данных и проведения препроцессинга на Python
	Владения средствами визуализации
	Владения основными модулями Python, широко используемыми для решения задач анализа данных и машинного обучения: numpy, pandas

Оценка преподавателем выставляется интегрально с учётом всех показателей и критериев оценивания.

Оценка сформированности компетенций по показателю **знания**.

Критерий	Уровень освоения и оценка	
	Не зачтено	Зачтено
Знание терминологии	Не знает терминов и определений	Знает термины и определения
Знание основных принципов и методов анализа данных	Не знает принципов и методов научных исследований	Проявляет достаточно полные знания о принципах и методах анализа данных
Знания о применимости методов анализа данных в сфере профессиональной деятельности	Не имеет представлений о применении методов анализа данных в области информационной безопасности	Демонстрирует понимание применения методов анализа данных в области информационной безопасности
Объем и полнота освоенного материала	Не знает значительной части материала дисциплины	Знает материал дисциплины в достаточном объеме

Оценка сформированности компетенций по показателю **умения**.

Критерий	Уровень освоения и оценка	
	Не зачтено	Зачтено

Способность уместно применять терминологию	Не применяет терминологию или применяет ее неуместно	Применяет терминологию везде, где это необходимо
Способность сопоставить метод решения типовой задаче анализа данных	Не может корректно подобрать метод для решения проблемы	Находит нужный метод для любой предложенной проблемы
Способность аргументировать предложение методов анализа данных	Не имеет своей точки зрения или не способен ее защитить	Показывает способность выработать и обосновать применение метода анализа данных по проблеме, конструктивно обсуждать ее
Способность уверенно ориентироваться в программных кодах на языке Python	Не знает языка Python	Применяет все, упомянутые при изучении дисциплины конструкции и типы данных, уверенно владеет синтаксисом
Способность провести анализ данных и представить результат	Не способен проводить анализ данных	Умеет подготавливать данные, подбирает средства анализа, визуализирует результат

Оценка сформированности компетенций по показателю *навыки*.

Критерий	Уровень освоения и оценка	
	Не зачтено	Зачтено
Программирования на Python	Не программирует на Python	Уверенно программирует с использованием современных парадигм
Подготовки данных и проведения препроцессинга на Python	Не может корректно подобрать метод подготовки данных	Находит нужный метод для любого предложенного датасета
Владения средствами визуализации	Не умеет визуализировать данные	Строит все виды рассмотренных в курсе графических объектов
Способность уверенно ориентироваться в программных кодах на языке Python	Не знает языка Python	Применяет все, упомянутые при изучении дисциплины конструкции и типы данных, уверенно владеет синтаксисом
Владения основными модулями Python, широко используемыми для решения задач анализа данных и машинного обучения: numpy, pandas	Не владеет numpy, pandas	Использует массивы numpy любой сложности, владеет методами pandas и их комбинациями

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

6.1. Материально-техническое обеспечение

№	Наименование специальных помещений и помещений для самостоятельной работы	Оснащенность специальных помещений и помещений для самостоятельной работы
1	Аудитория 425 ГУК	Компьютерный класс с выходом в интернет
2	Аудитория 426 ГУК	Компьютерный класс с моделью автоматизированной системы

6.2. Лицензионное и свободно распространяемое программное обеспечение

№	Перечень лицензионного программного обеспечения.	Реквизиты подтверждающего документа
1	Операционная система Ubuntu 18	Свободно распространяемое ПО
2	Операционная система Linux Mint 20	Свободно распространяемое ПО

6.3. Перечень учебных изданий и учебно-методических материалов

1. Грас Дж. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил.
2. Маккинни Уэс. Python и анализ данных: Пер. с англ. - М.: ДМК-Пресс, 2015. - 484 с.: ил.
3. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман. — М.: ДМК, 2016. — 498 с.
4. Шашков В.Б. Прикладной регрессионный анализ. Многофакторная регрессия: Учебное пособие.- Оренбург: ГОУ ВПО ОГУ, 2003. – 363 с.
5. Ростовцев В.С. Искусственные нейронные сети: учебник / В.С. Ростовцев. – Киров: Изд-во ВятГУ, 2014. – 208 с.
6. Гитис Л.Х. Кластерный анализ в задачах классификации, оптимизации и прогнозирования. / Л.Х. Гитис. – М: Изд-во МГГУ, 2001. – 104 с.
7. Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. / К.В. Воронцов. – М. Изд-во МФТИ, 2021. – 107 с.

6.4. Перечень интернет ресурсов, профессиональных баз данных, информационно-справочных систем

1. Ресурс для аналитиков данных <https://kaggle.com/>
2. Система elibrary: <https://elibrary.ru/>
3. Система Google Scholar: <https://scholar.google.com/>
4. Научная библиотека БГТУ им. В.Г. Шухова: <http://ntb.bstu.ru/jirbis2/>

7. УТВЕРЖДЕНИЕ РАБОЧЕЙ ПРОГРАММЫ

Рабочая программа утверждена на 20____ /20____ учебный год
без изменений / с изменениями, дополнениями

Протокол № _____ заседания кафедры от «__» _____ 20____ г.

Заведующий кафедрой _____
подпись, ФИО

Директор института _____
подпись, ФИО