

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ**  
Федеральное государственное бюджетное образовательное учреждение  
Высшего образования  
«Воронежский государственный технический университет»

УТВЕРЖДАЮ  
Декан факультета информационных  
технологий и компьютерной безопасности  
\_\_\_\_\_/П.Ю. Гусев/  
подпись  
«31» августа 2021 г.

**РАБОЧАЯ ПРОГРАММА**  
дисциплины

**«Анализ данных и машинное обучение»**

**Направление подготовки** 09.03.01 Информатика и вычислительная техника

**Профиль** Программное обеспечение автоматизированных систем

**Квалификация выпускника** бакалавр

**Нормативный период обучения** 4 года/4 года и 11м.

**Форма обучения** очная/заочная

**Год начала подготовки** 2021

Автор программы

\_\_\_\_\_/ Ветохин В.В. /

Заведующий кафедрой  
Компьютерных  
интеллектуальных  
технологий проектирования

\_\_\_\_\_/ Чижов М.И. /

Руководитель ОПОП

\_\_\_\_\_/ Чижов М.И. /

Воронеж 2021

## **1.ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ**

**1.1.Цели дисциплины** Целью освоения дисциплины «Анализ данных и машинное обучение» является овладение студентами моделями и методами анализа данных и основ машинного обучения в задачах поиска информации, обработки и анализа данных, а также приобретение навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

### **1.2.Задачи освоения дисциплины**

Изучение инструментов анализа данных.

Способы представления данных для анализа.

Математическое обеспечение анализа данных.

Изучение основных моделей и методов машинного обучения и разработки данных

Приобретение навыков адекватно применять указанные модели и методы, а также программные средства, в которых они реализованы

Приобрести навыки (приобрести опыт) анализа реальных данных с помощью изученных методов

## **2.МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП**

Дисциплина «Анализ данных и машинное обучение» относится к дисциплинам вариативной части блока Б1.

## **3.ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ**

Процесс изучения дисциплины «Анализ данных и машинное обучение» направлен на формирование следующих компетенций:

ПК-3 - Способен подготавливать данные для проведения аналитических работ.

ПК-4 - Способен разрабатывать и использовать техническую документацию в соответствии со спецификой образовательной программы.

| <b>Компетенция</b> | <b>Результаты обучения, характеризующие сформированность компетенции</b>                                                                                                                                                               |
|--------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ПК-3               | Знать:<br>- основные классы задач анализа данных;<br>- способов представления данных для анализа;<br>- основные модели и методы машинного обучения;<br>- программные средства, реализующие методы анализа данных и машинного обучения. |
|                    | Уметь:<br>- идентифицировать задачи анализа данных;                                                                                                                                                                                    |

|      |                                                                                                                                                                                                                                                                                                                                           |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|      | - обоснованно подбирать методы решения для практических задач анализа данных.                                                                                                                                                                                                                                                             |
|      | Владеть:<br>навыками использования специального программного обеспечения, реализующего методы анализа данных и машинного обучения.                                                                                                                                                                                                        |
| ПК-4 | Знать:<br>- методологию и стандарты данных;<br>- анализ данных на основе стандарта CRISP DM, этапы реализации стандарта;<br>- шкалы измерений и классификацию данных;<br>- виды статистических графиков (график плотности распределения, диаграмма рассеяния, диаграмма остатков), матрицу корреляций, box-plot, временные ряды и тренды. |
|      | Уметь осуществлять графический анализ данных.                                                                                                                                                                                                                                                                                             |
|      | Владеть инструментальными средствами, позволяющими реализовать первичный графический анализ данных.                                                                                                                                                                                                                                       |

#### 4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины «Анализ данных и машинное обучение» составляет 4 з.е.

Распределение трудоемкости дисциплины по видам занятий  
**очная форма обучения**

| Виды учебной работы                             | Всего часов | Семестры |
|-------------------------------------------------|-------------|----------|
|                                                 |             | 6        |
| <b>Аудиторные занятия (всего)</b>               | 64          | 64       |
| В том числе:                                    |             |          |
| Лекции                                          | 36          | 36       |
| Лабораторные работы (ЛР)                        | 36          | 36       |
| <b>Самостоятельная работа</b>                   | 82          | 82       |
| <b>Курсовой проект</b>                          | +           | +        |
| Часы на контроль                                |             |          |
| Виды промежуточной аттестации – зачет с оценкой | +           | +        |
| Общая трудоемкость:<br>академические часы       | 144         | 144      |
| зач.ед.                                         | 4           | 4        |

**заочная форма обучения**

| Виды учебной работы               | Всего часов | Семестры |
|-----------------------------------|-------------|----------|
|                                   |             | 8        |
| <b>Аудиторные занятия (всего)</b> | 24          | 24       |

|                                                 |            |            |
|-------------------------------------------------|------------|------------|
| В том числе:                                    |            |            |
| Лекции                                          | 8          | 8          |
| Лабораторные работы (ЛР)                        | 16         | 16         |
| <b>Самостоятельная работа</b>                   | <b>116</b> | <b>116</b> |
| <b>Курсовой проект</b>                          | <b>+</b>   | <b>+</b>   |
| Часы на контроль                                | 4          | 4          |
| Виды промежуточной аттестации – зачет с оценкой | +          | +          |
| Общая трудоемкость:                             |            |            |
| академические часы                              | 144        | 144        |
| зач.ед.                                         | 4          | 4          |

## 5.СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

### 5.1Содержание разделов дисциплины и распределение трудоемкости по видам занятий

#### Очная форма обучения

| № п/п | Наименование темы                                        | Содержание раздела                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Лекц | Лаб. зан. | СРС | Всего, час |
|-------|----------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|-----------|-----|------------|
| 1     | Введение в анализ данных и машинное обучение             | Методология и стандарты данных. Анализ данных на основе стандарта CRISP DM, этапы реализации стандарта. Обзор современных цифровых технологий обработки данных, их применение: аналитика больших данных, прогнозирование, классификация, кластеризация, машинное обучение. Способы описания объектов в задачах машинного обучения. Типы признаков объектов. Виды задач машинного обучения. Модели обучения. Алгоритм обучения. Проблема переобучения. Этапы решения задач машинного обучения | 6    | 4         | 10  | 20         |
| 2     | Подготовка данных и первичный графический анализ данных  | Качественные и категориальные признаки. Шкалы измерений и классификация данных. Трудности при сборе данных, важность корректных данных. Виды статистических графиков. Меры среднего и их применимость. Меры вариации. Инструменты графического анализа: график плотности распределения, диаграмма рассеяния, матрица корреляций, box-plot, временные ряды и тренды.                                                                                                                          | 8    | 8         | 6   | 22         |
| 3     | Классификация, деревья решений и метод ближайших соседей | Дерево решений. Как строится дерево решений. Критерии качества разбиения в задаче классификации. Энтропия. Неопределенность Джини. Как дерево решений работает с количественными признаками. Основные параметры дерева. Дерево решений в задаче регрессии. Метод ближайших соседей. Выбор параметров модели и                                                                                                                                                                                | 4    | 8         | 14  | 26         |

|              |                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |           |           |           |            |
|--------------|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-----------|-----------|------------|
|              |                                           | кросс-валидация. Примеры применения деревьев решений и метода ближайших соседей. Плюсы и минусы деревьев решений и метода ближайших соседей.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |           |           |           |            |
| 4            | Линейные модели классификации и регрессии | Линейная регрессия. Метод наименьших квадратов. Метод максимального правдоподобия. Логистическая регрессия. Логистическая регрессия как линейный классификатор. Регуляризация логистических потерь.                                                                                                                                                                                                                                                                                                                                                                                                                                     | 4         | 4         | 14        | 22         |
| 5            | Статистический анализ данных              | Введение в статистический анализ данных. Доверительный интервал для среднего, доверительный интервал для доли. Статистическое испытание гипотез. Испытание гипотезы о среднем генеральной совокупности. Испытание гипотезы по выборочной доле. Статистические испытания на независимость. Корреляция. Корреляционный анализ. Коэффициент корреляции Пирсона. Ранговый коэффициент корреляции Спирмена. Статистический вывод в регрессионном анализе. Множественная регрессия. Анализ временных рядов.                                                                                                                                   | 8         | 4         | 14        | 26         |
| 6            | Нейросетевые технологии анализа данных    | Понятие искусственной нейронной сети. Задачи, решаемые с использованием искусственных нейронных сетей. Структура и принципы функционирования искусственного нейрона. Классификация искусственных нейронных сетей. Однослойные и многослойные сети. Подходы к обучению искусственных нейронных сетей. Обучение с учителем. Обучение без учителя. Алгоритм обратного распространения ошибки. Принципы построения и функционирования свёрточных нейронных сетей. Этапы нейросетевого моделирования. Практическое использование нейросетевых технологий для решения задач обработки и анализа данных. Оценка качества нейросетевых моделей. | 6         | 8         | 14        | 28         |
| <b>Итого</b> |                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | <b>36</b> | <b>36</b> | <b>72</b> | <b>144</b> |

### Заочная форма обучения

| № п/п | Наименование темы                            | Содержание раздела                                                                                                                                                                                                                                                                                                                                                     | Лекц | Лаб. зан. | СРС | Всего, час |
|-------|----------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|-----------|-----|------------|
| 1     | Введение в анализ данных и машинное обучение | Методология и стандарты данных. Анализ данных на основе стандарта CRISP DM, этапы реализации стандарта. Обзор современных цифровых технологий обработки данных, их применение: аналитика больших данных, прогнозирование, классификация, кластеризация, машинное обучение. Способы описания объектов в задачах машинного обучения. Типы признаков объектов. Виды задач | 1    | 4         | 16  | 21         |

|   |                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |   |   |    |    |
|---|----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|---|----|----|
|   |                                                          | машинного обучения. Модели обучения. Алгоритм обучения. Проблема переобучения. Этапы решения задач машинного обучения                                                                                                                                                                                                                                                                                                                                                                                 |   |   |    |    |
| 2 | Подготовка данных и первичный графический анализ данных  | Качественные и категориальные признаки. Шкалы измерений и классификация данных. Трудности при сборе данных, важность корректных данных. Виды статистических графиков. Меры среднего и их применимость. Меры вариации. Инструменты графического анализа: график плотности распределения, диаграмма рассеяния, матрица корреляций, box-plot, временные ряды и тренды.                                                                                                                                   | 1 | 4 | 20 | 25 |
| 3 | Классификация, деревья решений и метод ближайших соседей | Дерево решений. Как строится дерево решений. Критерии качества разбиения в задаче классификации. Энтропия. Неопределенность Джини. Как дерево решений работает с количественными признаками. Основные параметры дерева. Дерево решений в задаче регрессии. Метод ближайших соседей. Выбор параметров модели и кросс-валидация. Примеры применения деревьев решений и метода ближайших соседей. Плюсы и минусы деревьев решений и метода ближайших соседей.                                            | 2 | 4 | 20 | 26 |
| 4 | Линейные модели классификации и регрессии                | Линейная регрессия. Метод наименьших квадратов. Метод максимального правдоподобия. Логистическая регрессия. Логистическая регрессия как линейный классификатор. Регуляризация логистических потерь.                                                                                                                                                                                                                                                                                                   | 1 | - | 20 | 21 |
| 5 | Статистический анализ данных                             | Введение в статистический анализ данных. Доверительный интервал для среднего, доверительный интервал для доли. Статистическое испытание гипотез. Испытание гипотезы о среднем генеральной совокупности. Испытание гипотезы по выборочной доле. Статистические испытания на независимость. Корреляция. Корреляционный анализ. Коэффициент корреляции Пирсона. Ранговый коэффициент корреляции Спирмена. Статистический вывод в регрессионном анализе. Множественная регрессия. Анализ временных рядов. | 1 | - | 20 | 21 |
| 6 | Нейросетевые технологии анализа данных                   | Понятие искусственной нейронной сети. Задачи, решаемые с использованием искусственных нейронных сетей. Структура и принципы функционирования искусственного нейрона. Классификация искусственных нейронных сетей. Однослойные и многослойные сети. Подходы к обучению искусственных нейронных сетей. Обучение с учителем. Обучение без учителя. Алгоритм обратного распространения ошибки.                                                                                                            | 2 | 4 | 20 | 26 |

|              |  |                                                                                                                                                                                                                                              |   |    |     |     |
|--------------|--|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|----|-----|-----|
|              |  | Принципы построения и функционирования свёрточных нейронных сетей. Этапы нейросетевого моделирования. Практическое использование нейросетевых технологий для решения задач обработки и анализа данных. Оценка качества нейросетевых моделей. |   |    |     |     |
| <b>Итого</b> |  |                                                                                                                                                                                                                                              | 8 | 16 | 116 | 140 |

## 5.2 Перечень лабораторных работ

1. Анализ данных с помощью Pandas.
2. Визуальный анализ данных с Python.
3. Деревья решений и метод ближайших соседей в задачах классификации. Метрики оценки качества обучения.
4. Улучшение показателей метрик обучения модели.
5. Создание моделей.
6. Нейросетевой анализ данных в Python на основе библиотек Keras.

## 6. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ ПРОЕКТОВ (РАБОТ) И КОНТРОЛЬНЫХ РАБОТ

В соответствии с учебным планом освоение дисциплины предусматривает выполнение курсовой проект в 6 семестре для очной формы обучения, в 8 семестре для заочной формы обучения.

Примерная тематика курсовой работы: «Разработка модели-классификатора набора тестовых данных».

Задачи, решаемые при выполнении курсовой работы:

- Сбор и интеллектуальный анализ исходных данных
- Визуальный анализ исходных данных
- Разработка модели-классификатора
- Подготовка отчета по проделанной работе.

## 7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

**7.1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания**

### 7.1.1 Этап текущего контроля

Результаты текущего контроля знаний и межсессионной аттестации оцениваются по следующей системе:

«аттестован»;

«неаттестован».

| Компетенция | Результаты обучения, характеризующие сформированность компетенции                                                               | Критерии оценивания           | Аттестован                                                    | Неаттестован                                                    |
|-------------|---------------------------------------------------------------------------------------------------------------------------------|-------------------------------|---------------------------------------------------------------|-----------------------------------------------------------------|
| ПК-3        | Знать:<br>- основные классы задач анализа данных;<br>- способов представления данных для анализа;<br>- основные модели и методы | Выполнение лабораторных работ | Выполнение работ в срок, предусмотренный в рабочих программах | Невыполнение работ в срок, предусмотренный в рабочих программах |

|      |                                                                                                                                                                                                                                                                                                                                           |                               |                                                               |                                                                 |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------|---------------------------------------------------------------|-----------------------------------------------------------------|
|      | машинного обучения;<br>- программные средства, реализующие методы анализа данных и машинного обучения.                                                                                                                                                                                                                                    |                               |                                                               |                                                                 |
|      | Уметь:<br>- идентифицировать задачи анализа данных;<br>- обоснованно подбирать методы решения для практических задач анализа данных.                                                                                                                                                                                                      | Выполнение лабораторных работ | Выполнение работ в срок, предусмотренный в рабочих программах | Невыполнение работ в срок, предусмотренный в рабочих программах |
|      | Владеть:<br>навыками использования специального программного обеспечения, реализующего методы анализа данных и машинного обучения.                                                                                                                                                                                                        | Выполнение лабораторных работ | Выполнение работ в срок, предусмотренный в рабочих программах | Невыполнение работ в срок, предусмотренный в рабочих программах |
| ПК-4 | Знать:<br>- методологию и стандарты данных;<br>- анализ данных на основе стандарта CRISP DM, этапы реализации стандарта;<br>- шкалы измерений и классификацию данных;<br>- виды статистических графиков (график плотности распределения, диаграмма рассеяния, диаграмма остатков), матрицу корреляций, box-plot, временные ряды и тренды. | Выполнение лабораторных работ | Выполнение работ в срок, предусмотренный в рабочих программах | Невыполнение работ в срок, предусмотренный в рабочих программах |
|      | Уметь<br>осуществлять графический анализ данных.                                                                                                                                                                                                                                                                                          | Выполнение лабораторных работ | Выполнение работ в срок, предусмотренный в рабочих программах | Невыполнение работ в срок, предусмотренный в рабочих программах |
|      | Владеть инструментальными средствами, позволяющими реализовать первичный графический анализ данных.                                                                                                                                                                                                                                       | Выполнение лабораторных работ | Выполнение работ в срок, предусмотренный в рабочих программах | Невыполнение работ в срок, предусмотренный в рабочих программах |

### 7.1.2 Этап промежуточного контроля знаний

Результаты промежуточного контроля знаний оцениваются в 6 семестре для очной формы обучения, 8 семестре для заочной формы обучения по четырех балльной системе:

«отлично»;

«хорошо»;

«удовлетворительно»;

«неудовлетворительно».

| Компетенция | Результаты обучения, характеризирующие сформированность компетенции                                                                       | Критерии и оценивание | Отлично                     | Хорошо                     | Удовл.                     | Неудовл.                             |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|-----------------------------|----------------------------|----------------------------|--------------------------------------|
| ПК-3        | Знать:<br>- основные классы задач анализа данных;<br>- способов представления данных для анализа;<br>- основные модели и методы машинного | Тест                  | Выполнение теста на 90-100% | Выполнение теста на 80-90% | Выполнение теста на 70-80% | В тесте менее 70% правильных ответов |

|      |                                                                                                                                                                                                                                                                                                                                           |                                                          |                                                        |                                                                                       |                                                          |                                      |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|--------------------------------------------------------|---------------------------------------------------------------------------------------|----------------------------------------------------------|--------------------------------------|
|      | обучения;<br>- программные средства, реализующие методы анализа данных и машинного обучения.                                                                                                                                                                                                                                              |                                                          |                                                        |                                                                                       |                                                          |                                      |
|      | Уметь:<br>- идентифицировать задачи анализа данных;<br>- обоснованно подбирать методы решения для практических задач анализа данных.                                                                                                                                                                                                      | Решение стандартных практических задач                   | Задачи решены в полном объеме и получены верные ответы | Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах | Продемонстрирован верный ход решения в большинстве задач | Задачи не решены                     |
|      | Владеть:<br>навыками использования специального программного обеспечения, реализующего методы анализа данных и машинного обучения.                                                                                                                                                                                                        | Решение прикладных задач в конкретной предметной области | Задачи решены в полном объеме и получены верные ответы | Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах | Продемонстрирован верный ход решения в большинстве задач | Задачи не решены                     |
| ПК-4 | Знать:<br>- методологию и стандарты данных;<br>- анализ данных на основе стандарта CRISP DM, этапы реализации стандарта;<br>- шкалы измерений и классификацию данных;<br>- виды статистических графиков (график плотности распределения, диаграмма рассеяния, диаграмма остатков), матрицу корреляций, box-plot, временные ряды и тренды. | Тест                                                     | Выполнение теста на 90-100%                            | Выполнение теста на 80-90%                                                            | Выполнение теста на 70-80%                               | В тесте менее 70% правильных ответов |
|      | Уметь осуществлять графический анализ данных.                                                                                                                                                                                                                                                                                             | Решение стандартных практических задач                   | Задачи решены в полном объеме и получены верные ответы | Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах | Продемонстрирован верный ход решения в большинстве задач | Задачи не решены                     |
|      | Владеть инструментальными средствами, позволяющими реализовать первичный графический анализ данных.                                                                                                                                                                                                                                       | Решение прикладных задач в конкретной предметной области | Задачи решены в полном объеме и получены верные ответы | Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах | Продемонстрирован верный ход решения в большинстве задач | Задачи не решены                     |

## 7.2 Примерный перечень оценочных средств (типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности)

### 7.2.1 Примерный перечень заданий для подготовки к тестированию

1. Какие из типов диаграмм подойдут для описания категориальных переменных

- Гистограммы
- Столбиковые диаграммы
- Кумулятивные частотные графики
- Секторные (круговые) диаграммы

1. Какое из средних наиболее восприимчиво к экстремальным значениям в наборе данных

- Мода
- Медиана
- Среднее арифметическое

2. В каком случае для вычисления среднего арифметического необходима средняя точка интервала

- Когда исходные данные представлены отдельными значениями
- Когда исходные данные представлены частотным распределением отдельных значений
- Исходные данные распределение сгруппированных значений

3. Статистическое испытание гипотез проводится

- По генеральной совокупности
- По выборочным данным
- Не имеет значения

4. Что из этого корректные названия типов признаков

- Номинальные (категориальные) признаки
- Бинарные признаки
- Нетривиальные признаки
- Устойчивые признаки
- Числовые (количественные) признаки

5. Задача сегментации изображения может это задача классификации кластеризации ни то, ни другое?

6. В приведенном списке отметьте задачи интеллектуального анализа данных, относящиеся к классу "обучение с учителем":

- классификация
- кластеризация
- регрессия
- поиск ассоциативных правил

7. В приведенном списке отметьте задачи интеллектуального анализа данных, относящиеся к классу "обучение без учителя":

- классификация
- регрессия
- кластеризация
- поиск ассоциативных правил

8. В приведенном списке выберите корректное утверждение относительно структур и моделей интеллектуального анализа

- структура содержит модели, которые используются для анализа ее данных
- модель содержит структуры, которые используются для анализа ее данных
- модели и структуры – это несвязанные типы объектов, ни один из них не может включать (содержать) другой.

9. Какие значения может принимать коэффициент корреляции Пирсона:

- от 0 до 1;
- от -10 до 10;
- от -1 до 1;
- от 0 до 10.

10. Что является результатом регрессионного анализа?

- расстояние регрессии;
- уравнение регрессии;
- матрица регрессии;
- коэффициент регрессии.

### **7.2.2 Примерный перечень заданий для решения стандартных задач**

1 Основное назначение систем интеллектуального анализа (Data Mining):

- 1) обнаружение в сырых данных скрытых знаний;
- 2) проведение статистического анализа;
- 3) решения задач математического программирования;
- 4) поиск агрегированных данных.

2.Какая из перечисленных задач не относится к задачам Data Mining:

- классификация.
- кластеризация.
- решение дифференциальных уравнений
- прогнозирование.

3. Перечислите правильную последовательность этапов Knowledge Discovery in Databases – процесса обнаружения знаний в базах данных:

- 1) трансформация, интерпретация результатов, выборка, очистка, построение моделей.
- 2) построение моделей, выборка, очистка, трансформация, интерпретация результатов.
- 3) построение моделей, выборка, очистка, трансформация,

интерпретация результатов.

4) выборка, очистка, трансформация, построение моделей, интерпретация результатов.

4. С чем может быть связано получение неверного результата при программной реализации вычислительного метода анализа данных? Укажите неправильный ответ

- 1) неверно выбран язык программирования
- 2) ошибка в программном коде
- 3) неверно выбраны параметры метода
- 4) неправильно построен алгоритм решения задачи.

5. Какая из перечисленных программных систем не может быть использована для анализа данных?

- 1 Matlab
- 2 EXCEL
- 3 3dMax
- 4 Statistica.

6. Объективными предпосылками появления ошибки модели регрессии являются:

- А) нерепрезентативность выборки
- Б) вероятность ошибочной интерпретации переменных
- В) вероятность ошибочного измерения переменных
- Г) сложность модели.

- 1 А, В
- 2 А, Б
- 3 Б, В
- 4 Б, Г

7. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

- А) должны быть качественно обоснованными
- Б) должны быть количественно измеримы
- В) не должны быть интеркоррелированы
- Г) должны быть функционально зависимыми.

- 1 Б, В
- 2 А, Б
- 3 Б, Г
- 4 В, Г

9. Статистическое испытание гипотез проводится

- по генеральной совокупности;
- по выборочным данным;
- не имеет значения.

10. Какое значение коэффициент корреляции не может принимать

- 0
- -1,28
- 0,95
- -0,13.

### 7.2.3 Примерный перечень заданий для решения прикладных задач

1. Найдите моду и медиану для данных, представленных частотным распределением

| x | f  |
|---|----|
| 1 | 7  |
| 2 | 9  |
| 3 | 18 |
| 4 | 12 |
| 5 | 6  |

2. Ниже перечислены рекомендуемые этапы проведения интеллектуального анализа данных. Расставьте их в правильной последовательности. Ответ укажите в виде последовательности чисел, например 654321

- (1) изучение данных
  - (2) подготовка данных
  - (3) постановка задачи
  - (4) развертывание и обновление моделей
  - (5) построение моделей
  - (6) исследование и проверка моделей
- Ответ: 321564

3. Выберите наиболее корректное продолжение фразы "При проведении интеллектуального анализа данных на этапе изучения данных ..."

- может проводиться поиск минимальных и максимальных значений параметров, анализ статистических характеристик, сравнение полученных результатов с представлениями о предметной области
- может проводиться удаление "выбросов", обработка отсутствующих значений параметров, численное преобразование (например, нормализация)
- может определяться, нужно ли будет делать прогнозы на основании модели интеллектуального анализа данных или просто найти содержательные закономерности и взаимосвязи.

4. Среди представленных в списке алгоритмов интеллектуального анализа выберите наиболее быстрый (требующий наименьшего количества вычислений):

- Microsoft Clustering
- Microsoft Naive Bayes
- Microsoft Neural Network.

5. В алгоритмах регрессии "регрессором" называется:

- независимая переменная (аргумент)
- зависимая переменная
- выявленная алгоритмом зависимость

6. Определение методами интеллектуального анализа значения непрерывного числового параметра на основании значений независимых переменных производится при решении задачи:

- классификации
- кластеризации
- регрессии.

7. Построение моделей Data Mining осуществляется с целью:

- исследования или изучения моделируемого объекта и получения новых знаний, необходимых для принятия решений;
- выбора наиболее быстродействующей модели;
- исследования всех возможных свойств и характеристик изучаемого объекта.

8. Рассчитайте корреляционную матрицу. Визуализируйте результат. Сделайте вывод.

9. Создание каких моделей Data Mining означает поиск правил, которые объясняют зависимость выходных параметров от входных?

- моделей классификации и прогнозирования
- моделей кластеризации и классификации
- моделей правил ассоциаций

10. Проверьте датасет на наличие выбросов. Используйте один из методов борьбы с ними (масштабирование, логарифмирование, нормализация в степени 1/3, бокс-трансформация). Выбор обоснуйте.

#### **7.2.4 Примерный перечень вопросов для подготовки к зачету**

1. Что понимается под интеллектуальным анализом данных?
2. Виды задач машинного обучения.
3. Что включает в себя наука о данных (Data Science)
5. Перечислите этапы реализации стандарта CRISP DM.
6. Стандарты данных и методология
7. Анализ данных на основе стандарта CRISP DM
8. Предварительная обработка данных
9. Каким образом количественные переменные можно перевести в категориальные?
10. Как строится дерево решений.
11. Критерии качества разбиения в задаче классификации.
12. Как дерево решений работает с количественными признаками.
13. Дерево решений в задаче регрессии.
14. Основные параметры дерева решений.

15. Метод ближайших соседей. Выбор параметров модели и кросс-валидация.
16. Линейная регрессия. Метод наименьших квадратов.
17. Линейная регрессия. Метод максимального правдоподобия.
18. Логистическая регрессия. Логистическая регрессия как линейный классификатор.
19. Описательная статистика: меры центральной тенденции, использование для анализа данных
20. Описательная статистика: меры разброса, использование для анализа данных
21. Какие основные способы существуют для графического представления количественных данных?
22. При каком распределении данных мода, медиана и среднее арифметическое совпадают?
23. Что такое статистический вывод?
24. Каким образом формулируются нулевая и альтернативная гипотезы.
25. Назовите четыре возможных результата при испытании гипотез.
26. Каким образом можно определить значимость статистической связи между двумя категориальными переменными?
27. Связи между какими признаками можно анализировать с помощью таблиц сопряженности.
28. На основании какого коэффициента можно говорить о возможном наличии линейных связей между количественными признаками.
29. С помощью каких тестов проверяются гипотезы о среднем арифметическом?
30. С помощью каких тестов проверяются гипотезы о связи между качественными признаками?
31. Что такое регрессия и корреляция?
32. Математически опишите линию регрессии по выборочным данным.
33. Понятие искусственной нейронной сети.
34. Задачи, решаемые с использованием искусственных нейронных сетей.
- 35 Структура и принципы функционирования искусственного нейрона.

#### **7.2.5Примерный перечень заданий для подготовки к экзамену.**

Не предусмотрено учебным планом.

#### **7.2.6.Методика выставления оценки при проведении промежуточной аттестации**

Зачёт проводится по билетам, каждый из которых содержит 7 теоретических вопроса и 1 задачу. Оценивается по принятой шкале («отлично», «хорошо», «удовлетворительно», «неудовлетворительно»).

Оценка «отлично» - за правильное выполнение всех заданий билета и правильный ответ на дополнительные вопросы по всей дисциплине.

Оценка «хорошо» - за правильное решение задачи и ответ на не менее 5

теоретические вопросов билета.

Оценка «удовлетворительно» - за правильное решение задачи и правильный ответ на 4 теоретических вопроса.

Оценка «неудовлетворительно» - если не решена задача или дано менее 4-х правильных ответов на теоретические вопросы.

### **7.2.7 Паспорт оценочных материалов**

| №п/п | Контролируемые разделы(темы) дисциплины                  | Код контролируемой компетенции | Наименование оценочного средства                                |
|------|----------------------------------------------------------|--------------------------------|-----------------------------------------------------------------|
| 1    | Введение в анализ данных и машинное обучение             | ПК-3, ПК-4                     | Тест, защита лабораторных работ, требования к курсовому проекту |
| 2    | Подготовка данных и первичный графический анализ данных  | ПК-3, ПК-4                     | Тест, защита лабораторных работ, требования к курсовому проекту |
| 3    | Классификация, деревья решений и метод ближайших соседей | ПК-3                           | Тест, защита лабораторных работ, требования к курсовому проекту |
| 4    | Линейные модели классификации и регрессии                | ПК-3                           | Тест, защита лабораторных работ, требования к курсовому проекту |
| 5    | Статистический анализ данных                             | ПК-3, ПК-4                     | Тест, защита лабораторных работ                                 |
| 6    | Нейросетевые технологии анализа данных                   | ПК-3                           | Тест, защита лабораторных работ, требования к курсовому проекту |

### **7.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности**

Тестирование осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экзаменатором и выставляется оценка согласно методики выставления оценки при проведении промежуточной аттестации.

Решение стандартных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 45 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

Решение прикладных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 45 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется

оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

## **8 УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ)**

### **8.1 Перечень учебной литературы, необходимой для освоения дисциплины**

1. Бессмертный, И.А. Искусственный интеллект: учебное пособие / И.А. Бессмертный. — Санкт-Петербург : НИУ ИТМО, 2010. — 132 с. — Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: <https://e.lanbook.com/book/43663>.

2. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; перевод с английского А. А. Слинкина. — 2-е изд. — Москва : ДМК Пресс, 2018. — 652 с. — ISBN 978-5-97060-618-6. — Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: <https://e.lanbook.com/book/107901>

3. Ларионова, И.А. Статистика: введение в регрессионный анализ : временные ряды : учебное пособие / И.А. Ларионова. — Москва : МИСИС, 2016. — 75 с. — ISBN 978-5-87623-936-5. — Текст : электронный // Электронно-библиотечная система «Лань»: [сайт]. — URL: <https://e.lanbook.com/book/93609>.

4. Минаева Ю.В. Методы статистического и интеллектуального анализа данных: учеб. пособие. – Воронеж: ВГТУ, 2017. 90 с.

5. Методические рекомендации по выполнению курсовых проектов (работ) по программам высшего образования – программам бакалавриата, специалитета, магистратуры / ФГБОУ ВО «Воронежский государственный технический университет»; сост. В.Н. Почечихина, И.Н. Крючкова, Е.И. Головина. Воронеж: Изд-во ВГТУ. 2020. 10 с.

6. Организация самостоятельной работы обучающихся : методические указания для студентов, осваивающих основные образовательные программы высшего образования – бакалавриата, специалитета, магистратуры: методические указания / сост. В.Н. Почечихина, И.Н. Крючкова, Е.И. Головина, В.Р. Демидов; ФГБОУ ВО «Воронежский государственный технический университет ». – Воронеж, 2020. – 14 с.

7. 47-2024 Интеллектуальный анализ данных [Электронный ресурс] : методические указания к выполнению курсового проекта для студентов направления 09.03.01 "Информатика и вычислительная техника" очной и заочной форм обучения / ФГБОУ ВО "Воронеж.гос. техн. ун-т", Каф. компьютерных интеллектуальных технологий проектирования; сост.: В.В. Ветохин - Воронеж: Воронежский государственный технический университет, 2024. - Электрон.текстовые и граф. данные (406 Кб).

### **8.2 Перечень информационных технологий, используемых при**

**осуществлении образовательного процесса по дисциплине, включая перечень лицензионного программного обеспечения, ресурсов информационно-телекоммуникационной сети «Интернет», современных профессиональных баз данных и информационных справочных систем:**

Лицензионное программное обеспечение:

- Среда разработки приложений Visual Studio
- Microsoft Office Word 2013/2007
- Microsoft Office Excel 2013/2007

Свободное программное обеспечение:

- OpenOffice

Python

Отечественное программное обеспечение:

- Яндекс.Браузер

Ресурсы информационно-телекоммуникационной сети «Интернет»:

- Образовательный портал ВГТУ

- <https://docs.microsoft.com/>

Информационные справочные системы:

- <http://window.edu.ru/>

- <https://wiki.cchgeu.ru>

Современные профессиональные базы данных:

- eLIBRARY.RU
- База ГОСТ docplan.ru

## **9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА**

- Проекционное оборудование;
- Сервер баз данных;
- «Лаборатория компьютерного моделирования и дизайна» или «Учебный центр ВГТУ, Академии Софтлайн, сетевой академии CISCO» (г. Воронеж, ул. Плехановская, д. 11).

## **10. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)**

По дисциплине «Анализ данных и машинное обучение» читаются лекции, проводятся лабораторные работы.

Основой изучения дисциплины являются лекции, на которых излагаются наиболее существенные и трудные вопросы, а также вопросы, ненашедшие отражения в учебной литературе.

Лабораторные работы выполняются на лабораторном оборудовании в соответствии с методиками, приведенными в указаниях к выполнению работ.

| Вид учебных занятий                   | Деятельность студента                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|---------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Лекция                                | Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначение вопросов, терминов, материала, которые вызывают трудности, поиск ответов в рекомендуемой литературе. Если самостоятельно не удастся разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или на практическом занятии. |
| Лабораторная работа                   | Лабораторные работы позволяют научиться применять теоретические знания, полученные на лекции при решении конкретных задач. Чтобы наиболее рационально и полно использовать все возможности лабораторных для подготовки к ним необходимо: следует разобрать лекцию по соответствующей теме, ознакомиться с соответствующим разделом учебника, проработать дополнительную литературу и источники, решить задачи и выполнить другие письменные задания.                                                                                                                             |
| Самостоятельная работа                | Самостоятельная работа студентов способствует глубокому усвоения учебного материала и развитию навыков самообразования. Самостоятельная работа предполагает следующие составляющие:<br>- работа с текстами: учебниками, справочниками, дополнительной литературой, а также проработка конспектов лекций;<br>- выполнение домашних заданий и расчетов;<br>- работа над темами для самостоятельного изучения;<br>- участие в работе студенческих научных конференций, олимпиад;<br>- подготовка к промежуточной аттестации.                                                        |
| Подготовка к промежуточной аттестации | Готовиться к промежуточной аттестации следует систематически, в течение всего семестра. Интенсивная подготовка должна начаться не позднее, чем за месяц-полтора до промежуточной аттестации. Данные перед экзаменом, экзаменом три дня эффективнее всего использовать для повторения и систематизации материала.                                                                                                                                                                                                                                                                 |