ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**«ЛЕНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ИМЕНИ А.С. ПУШКИНА»**

УТВЕРЖДАЮ

Проректор по учебно-методической

работе

\_\_\_\_\_\_\_\_\_\_\_\_ С.Н.Большаков

РАБОЧАЯ ПРОГРАММА

дисциплины

**Б1.В.03.ДВ.01.01 ТЕХНОЛОГИЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ**

Направление подготовки **09.04.03 Прикладная информатика**

Направленность (профиль) Прикладная информатика в цифровой экономике

(год начала подготовки – 2022)

Санкт-Петербург

2022

**1. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ:**

Процесс изучения дисциплины направлен на формирование следующих компетенций:

|  |  |  |
| --- | --- | --- |
| Индекс компетенции | Содержание компетенции  (или ее части) | Индикаторы компетенций (код и содержание) |
| УК-1 | Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий | ИУК-1.1. Знает процедуры критического анализа, методики анализа результатов исследования и разработки стратегий проведения исследований, организации процесса принятия решения.  ИУК-1.2. Умеет принимать конкретные решения для повышения эффективности процедур анализа проблем, принятия решений и разработки стратегий.  ИУК-1.3. Владеет методами установления причинно-следственных связей и определения наиболее значимых среди них; методиками постановки цели и определения способов ее достижения; методиками разработки стратегий действий при проблемных ситуациях. |
| ПК-3 | Способен проектировать информационные процессы и системы с использованием инновационных инструментальных средств | ИПК-3.1. Знает инновационные инструментальные средства проектирования ИС.  ИПК-3.2. Умеет проектировать информационные процессы и системы, адаптировать современные ИКТ.  ИПК-3.3. Владеет способностью проектировать информационные процессы и системы с использованием инновационных инструментальных средств, адаптировать современные ИКТ к задачам прикладных ИС. |
| ПК-4 | Способен принимать эффективные проектные решения в условиях неопределенности и риска | ИПК-4.1. Знает условия неопределенности и риска проектных решений.  ИПК-4.2. Умеет принимать эффективные проектные решения в условиях неопределенности.  ИПК-4.3. Владеет способностью принимать эффективные проектные решения в условиях неопределенности и риска. |
| ПК-6 | Способен управлять информационными ресурсами и ИС | ИПК-6.1. Знает структуру деятельности по управлению сложными объектами, суть понятий информационные ресурсы и информационные системы.  ИПК-6.2. Умеет формулировать управляющие воздействия на информационные ресурсы и информационные системы.  ИПК-6.3. Владеет одним из средств компьютерного мониторинга состояния информационных ресурсов и информационных систем. |
| ПК-8 | Способен использовать и развивать методы научных исследований и инструментария в области проектирования и управления информационными системами в прикладных областях | ИПК-8-1. Знает методы научных исследований и инструментария в области проектирования и управления ИС.  ИПК-8-2. Умеет выбирать и использовать методы научных исследований и инструментария в области проектирования и управления ИС.  ИПК-8.3. Владеет способами применения методов научных исследований и инструментария в области проектирования и управления ИС в прикладных областях. |

**2. Место дисциплины в структуре ОП:**

Цель дисциплины: знакомство обучающихся с основными технологиями решения задач обработки больших данных, формирование умений применять методы анализа больших данных, реализовывать приложения для аналитики больших данных.

Задачи дисциплины:

* изучение базовых понятий технологии Big Data;
* изучение базовых понятий прогнозирования;
* определение массивов больших данных и проведение анализа кластеров больших данных;
* изучение современных технологий создания и обслуживания больших данных;
* овладение методикой прогнозирования;
* построение различными способами прогнозов развития экономических процессов.

Дисциплина относится к дисциплинам по выбору части, формируемой участниками образовательных отношений, модуль Современные технологии представления и передачи данных.

Освоение дисциплины и сформированные при этом компетенции необходимы в последующей деятельности.

**3. Объем дисциплины и виды учебной работы:**

Общая трудоемкость освоения дисциплины составляет 3 зачетные единицы, 108 академических часов *(1 зачетная единица соответствует 36 академическим часам).*

Очная форма обучения

|  |  |  |
| --- | --- | --- |
| Вид учебной работы | Трудоемкость в акад.час | |
|  |  | Практическая подготовка |
| **Контактная работа (аудиторные занятия) (всего):** | 20 | |
| в том числе: |  | |
| Лекции | 10 | - |
| Лабораторные работы / Практические занятия (в т.ч. зачет) | 10/- | 4/- |
| **Самостоятельная работа (всего)** | 88 | |
| **Вид промежуточной аттестации (экзамен):** | - | |
| контактная работа | - | |
| самостоятельная работа по подготовке к экзамену | - | |
| **Общая трудоемкость дисциплины (в час. /з.е.)** | 108/3 | |

Заочная форма обучения

|  |  |  |
| --- | --- | --- |
| Вид учебной работы | Трудоемкость в акад.час | |
|  |  | Практическая подготовка |
| **Контактная работа (аудиторные занятия) (всего):** | 16 | |
| в том числе: |  | |
| Лекции | 4 | - |
| Лабораторные работы/ Практические занятия | 12/- | 4/- |
| **Самостоятельная работа (всего)** | 88 | |
| **Вид промежуточной аттестации (зачет):** | 4 | |
| контактная работа | 0,25 | |
| самостоятельная работа по подготовке к зачету | 3,75 | |
| **Вид промежуточной аттестации (экзамен):** | - | |
| контактная работа | - | |
| самостоятельная работа по подготовке к экзамену | - | |
| **Общая трудоемкость дисциплины (в час. /з.е.)** | 108/3 | |

**4.СОДЕРЖАНИЕ ДИСЦИПЛИНЫ:**

При проведении учебных занятий обеспечивается развитие у обучающихся навыков командной работы, межличностной коммуникации, принятия решений, лидерских качеств (включая при необходимости проведение интерактивных лекций, групповых дискуссий, ролевых игр, тренингов, анализ ситуаций и имитационных моделей, составленных на основе результатов научных исследований, проводимых организацией, в том числе с учетом региональных особенностей профессиональной деятельности выпускников и потребностей работодателей**).**

**4.1 Блоки (разделы) дисциплины.**

|  |  |
| --- | --- |
| № | Наименование блока (раздела) дисциплины |
| 1 | Введение в науку и технологию больших данных. |
| 2 | Структуры и методы представления больших данных. |
| 3 | Программно-аппаратная инфраструктура систем на основе больших данных. |
| 4 | Методы поиска, сбора, обработки и представления больших данных. |
| 5 | Исследовательский анализ больших данных и построение моделей. |

**4.2. Примерная тематика курсовых работ (проектов):**

Курсовая работа по дисциплине не предусмотрена учебным планом.

**4.3. Перечень занятий, проводимых в активной и интерактивной формах, обеспечивающих развитие у обучающихся навыков командной работы, межличностной коммуникации, принятия решений, лидерских качеств. Практическая подготовка\*.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№ п/п** | **Наименование блока (раздела) дисциплины** | **Занятия, проводимые в активной и интерактивной формах** | | **Практическая подготовка\*** |
| **Форма проведения занятия** | **Наименование видов занятий** |
| 1. | Введение в науку и технологию больших данных. | лекционное занятие | лекция-дискуссия |  |
| 2. | Структуры и методы представления больших данных. | лекционное занятие | лекция-дискуссия |  |
| 3. | Программно-аппаратная инфраструктура систем на основе больших данных. | лекционное занятие | лекция-дискуссия |  |
| 4. | Методы поиска, сбора, обработки и представления больших данных. | лекционное занятие  лабораторное занятие | лекция-дискуссия  работа в группах |  |
| 5. | Исследовательский анализ больших данных и построение моделей. | лекционное занятие  лабораторное занятие | лекция-дискуссия  работа в группах | лингвистическое резюмирование результатов |

**\***Практическая подготовка при реализации учебных предметов, курсов, дисциплин (модулей) организуется путем проведения практических занятий, практикумов, лабораторных работ и иных аналогичных видов учебной деятельности, **предусматривающих участие обучающихся в выполнении отдельных элементов работ, связанных с будущей профессиональной деятельностью.**

**5. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине:**

**5.1 Темы конспектов**

Тема 1. Введение в науку и технологию больших данных.

Тема 2. Структуры и методы представления больших данных.

Тема 3. Программно-аппаратная инфраструктура систем на основе больших данных.

Тема 4. Методы поиска, сбора, обработки и представления больших данных.

Тема 5. Исследовательский анализ больших данных и построение моделей.

**5.2 Темы для творческой самостоятельной работы обучающегося**

Темы для творческой самостоятельной работы студента формулируются обучающимся самостоятельно, исходя из перечня тем занятий текущего семестра.

**5.3. Темы рефератов**

* 1. Подходы к структурированию и типы данных в DataScience.
  2. Машинные данные и данные на естественном языке.
  3. Этапы в технологии представления больших данных.
  4. Экосистемы больших данных
  5. Распределенные файловые системы
  6. Научные проблемы в области больших данных
  7. Применение технологии больших данных в прогнозировании и предвидении в социально-политических процессах
  8. Понятие машинного обучения, его цели и сферы применения.
  9. Типы машинного обучения
  10. Базы данных в DataScience
  11. Многомерные статистические методы в экономике, управлении и финансах
  12. Реализация статистических методов в пакетах прикладных программ
  13. Математические основы многомерных статистических методов
  14. Методы множественного корреляционно-регрессионного анализа.
  15. Кластерный анализ
  16. кластерного анализа
  17. Постановка задач снижения размерности
  18. Факторный анализ
  19. Введение в теорию качественных признаков и нечисловой информации
  20. Интеграция данных из различных источников.

**5.4. Вопросы для подготовки к практическим занятиям (лабораторным работам):**

*Лабораторная работа №1. Изучение опыта применения методов кластеризации данных*

1. В чем состоит цель кластеризации? Приведите формальную постановку задачи кластеризации.
2. Приведите перечень и особенности методов кластеризации.
3. Какие метрики применяют в кластеризации?
4. Приведите математические характеристики кластеров и меры качества результатов кластеризации.
5. Охарактеризуйте этапы кластерного анализа выбранного метода.

*Лабораторная работа №2. Прогнозирование на основе статистического подхода.*

1. Постановка задачи, основные задачи анализа временных рядов. Критерии качества моделей. Стационарные и нестационарные временные ряды.
2. Какие основные классы методов анализа временных рядов? Data-driven и model‐driven методы анализа. Проблемы прогнозирования.
3. Принципы прогнозирования в статистическом подходе к анализу временных рядов.
4. Декомпозиция временных рядов, типы паттернов.
5. Модели тренда временных рядов (на основе функций от времени).

**6 Оценочные средства для текущего контроля успеваемости:**

**6.1. Текущий контроль**

|  |  |  |
| --- | --- | --- |
| №п/п | № и наименование блока (раздела) дисциплины | Форма текущего контроля |
| 1 | Разделы 1-5 | Проверка наличия конспектов.  Устный опрос, тест |
| 2 | Разделы 4-5 | Защита отчета по результатам выполнения лабораторных работ |

**6.2. Примеры оценочных средств для текущего контроля успеваемости**

***Темы конспектов.***

Представлены в разделе 5.1.

***Перечень практических заданий.***

Представлены в п. 5.4.

***Примеры тестовых заданий.***

1. Появление термина «большие данные» связывают ...

*Выберите один правильный вариант ответа*

* с началом в 2002 году «цифровой эры», когда мировой объём цифровых данных превысил количество аналоговых
* с выходом в 2008 году специального выпуска журнала Nature, посвящённого феномену взрывного роста объёмов и многообразия обрабатываемых данных
* с появлением тренда «Big Data» в 2011 году на Цикле зрелости (хайпа) новых технологий компании Gartner
* с проведением 28-30 октября 2014 года в г. Пекине (КНР) I-ой Международной конференции по использованию больших данных в официальной статистике

2. Выберите вариант ответа, в котором перечислены **четыре** из основных характеристик понятия "Big Data":

*Выберите один правильный вариант ответа*

* Virtualization, Volume, Vehicle, Variability
* Variety, Velocity, Veracity, Value
* Vacation, Volume, Velocity, Visualization
* Value, Variety, Volume, Video

3. Отметьте НЕВЕРНОЕ понимание термина Velocity в контексте характеристик Big Data:

*Выберите один правильный вариант ответа*

* Необходимости высокоскоростной обработки данных и получения результата в реальном масштабе времени
* Входные данные поступают через высокоскоростную потоковую передачу
* Значительное увеличение количества источников данных
* Частота обновления информации в базах данных
* Высокая скорость генерирования новых данных

4. Один из нижеперечисленных факторов не имеет отношения к появлению тренда больших данных. Укажите его.

*Выберите один правильный вариант ответа*

* Рост числа мобильных устройств
* Рост числа интернет-пользователей
* Увеличение количества «умных» вещей
* Снижение издержек на хранение данных
* Появление технологии вычислений в оперативной памяти

5. Укажите вариант представления информации, в котором данные являются структурированными.

*Выберите один правильный вариант ответа*

* Данные о продажах компании за год, представленные в виде ежемесячных отчётов, созданных с помощью текстового процессора MS Word
* Файл с фильмом, записанным в формате MPEG-4
* Записи мониторинга физической активности спортсмена за год в текстовом файле формата CSV
* Отчёт о проделанной работе, представленный в файле формата PDF

6. Среди нижеперечисленных высказываний о «больших данных» выберите **некорректное**.

*Выберите один правильный вариант ответа*

* Увеличение производительности телекоммуникационных каналов привело к росту объёмов передаваемой информации
* Появление методов интеллектуального анализа данных привело к росту рынка больших данных
* Традиционные системы управления базами данных могут эффективно использовать только структурированные источники поступления информации
* Увеличение объёмов данных и разнообразия их источников улучшает как достоверность самих данных, так и принятых решений на основе этих данных

7. Среди нижеперечисленных высказываний о «больших данных» выберите **некорректное**.

*Выберите один правильный вариант ответа*

* Технологии на основе больших данных позволяют увеличить эффективность маркетинговых кампаний
* Технологии на основе больших данных используются для поддержки принятия решений
* Технологии на основе больших данных используются для сокращения числа поломок производственного оборудования
* Технологии на основе больших данных позволяют повысить уровень информационной безопасности

8. Укажите завершение предложения «Применение технологий больших данных позволяет компаниям…», которое делает это утверждение **некорректным**:

*Выберите один правильный вариант ответа*

* увеличить показатели доходности и прибыльности
* оптимизировать логистические процессы
* оценить уровень удовлетворённости клиентов после взаимодействия с компанией
* увеличить численность производственного персонала

9. Среди нижеперечисленных задач обработки данных одна является типичной для использования технологий «больших данных». Укажите её.

*Выберите один правильный вариант ответа*

* Анализ прибыли компании за год
* Анализ отношения потребителей к бренду
* Анализ результатов социологического опроса
* Анализ квартальной бухгалтерской отчётности организации

10. Укажите все корректные высказывания о технологии больших данных.

*Выберите все правильные варианты ответа*

* Степень изменения масштабов сбора информации в обществе оказывает глубокое воздействие на наше понимание экономики
* Применение технологий интеллектуального анализа данных способно сыграть ключевую роль в противодействии мошенничеству
* Наиболее массовое направление применения больших данных — это построение предсказательных моделей
* Применение технологий больших данных в промышленности является одним из наиболее эффективных способов повысить производительность и снизить расходы
* Большие данные содержат ценную аналитическую информацию о клиентах и об особенностях их мотивации, помогают анализировать привычки клиентов

11. Укажите корректное определение понятию «Машинное обучение» среди нижеперечисленных.

*Выберите один правильный вариант ответа*

* Наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению
* Методы построения компьютерных алгоритмов, способных обучаться решению поставленной задачи
* Задача автоматического извлечения структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов
* Компьютерное моделирование видов человеческой деятельности, традиционно считающихся интеллектуальными
* Класс методов анализа данных, концентрирующийся на прогнозировании будущего поведения объектов и субъектов с целью принятия оптимальных решений

12. Укажите корректное определение понятию «Предсказательная аналитика» среди нижеперечисленных.

*Выберите один правильный вариант ответа*

* Наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению
* Методы построения компьютерных алгоритмов, способных обучаться решению поставленной задачи
* Задача автоматического извлечения структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов
* Компьютерное моделирование видов человеческой деятельности, традиционно считающихся интеллектуальными
* Класс методов анализа данных, концентрирующийся на прогнозировании будущего поведения объектов и субъектов с целью принятия оптимальных решений

13. Укажите корректное определение понятию «Data Mining» среди нижеперечисленных.

*Выберите один правильный вариант ответа*

* Наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению
* Методология и процесс обнаружения знаний в больших массивах данных
* Задача автоматического извлечения структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов
* Компьютерное моделирование видов человеческой деятельности, традиционно считающихся интеллектуальными
* Класс методов анализа данных, концентрирующийся на прогнозировании будущего поведения объектов и субъектов с целью принятия оптимальных решений

14. Среди представленных определений термина «Hadoop» одно является некорректным. Укажите его.

*Выберите один правильный вариант ответа*

* Набор утилит и программный каркас для выполнения распределённых программ, работающих на вычислительных кластерах
* Распределённая операционная система, позволяющая автоматически распределять обработку данных по узлам вычислительных кластеров
* Фреймворк для программирования распределённых вычислений в рамках парадигмы MapReduce
* Распределённая файловая система, предназначенная для хранения файлов большого объёма

15. Коммерческая клиника желает установить структуру своих клиентов с точки зрения вклада в доход клиники. К какому типу относится эта задача анализа данных?

*Выберите один правильный вариант ответа*

* Прогнозирование
* Цензурирование
* Кластеризация
* Регрессия

**7. ПЕРЕЧЕНЬ УЧЕБНОЙ ЛИТЕРАТУРЫ:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № п/п | Наименование | Авторы | Место издания | Год издания | Наличие | |
| печатные издания | ЭБС (адрес в сети Интернет) |
| 1. | Наука о данных: базовый курс: учебное пособие | Д. Келлехер, Б. Тирни | М.: Альпина Паблишер | 2020 |  | <https://biblioclub.ru> |
| 2. | Базы данных в высокопроизводительных информационных системах: учебное пособие | Николаев Е.И. | Ставрополь: СКФУ | 2016 |  | <https://biblioclub.ru> |
| 3. | Математика на языке R: учебник | Зададаев С. А. | М.: Прометей | 2018 |  | <https://biblioclub.ru> |
| 4. | Модели баз данных: учебное пособие | О.Е. Аврунев, В.М. Стасышин | Новосибирск: НГТУ | 2018 |  | <https://biblioclub.ru> |
| 5. | Анализ данных: учебник для вузов | Мхитарян В.С. | М.: Издательство Юрайт | 2022 |  | <https://urait.ru/bcode/489100> |
| 6. | Введение в анализ данных: учебник и практикум | Миркин Б. Г. | М.: Издательство Юрайт | 2022 |  | <https://urait.ru/bcode/469306> |

**8. Ресурсы информационно-телекоммуникационной сети «Интернет»:**

1. «НЭБ». Национальная электронная библиотека. – Режим доступа: [http://нэб.рф/](http://www.biblioclub.ru/)

2. «eLibrary». Научная электронная библиотека. – Режим доступа: [https://elibrary.ru](https://elibrary.ru/)

3. «КиберЛенинка». Научная электронная библиотека. – Режим доступа: <https://cyberleninka.ru/>

4. ЭБС «Университетская библиотека онлайн». – Режим доступа: [http://www.biblioclub.ru/](http://www.knigafund.ru/)

5. Российская государственная библиотека. – Режим доступа: <http://www.rsl.ru/>

**9. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, ИСПОЛЬЗУЕМЫЕ ПРИ ОСУЩЕСТВЛЕНИИ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ:**

В ходе осуществления образовательного процесса используются следующие информационные технологии:

- средства визуального отображения и представления информации (LibreOffice) для создания визуальных презентаций как преподавателем (при проведении занятий) так и обучаемым при подготовке докладов для семинарского занятия.

- средства телекоммуникационного общения (электронная почта и т.п.) преподавателя и обучаемого.

- использование обучаемым возможностей информационно-телекоммуникационной сети «Интернет» при осуществлении самостоятельной работы.

**9.1. Требования к программному обеспечению учебного процесса**

Для успешного освоения дисциплины, обучающийся использует следующие программные средства:

* Windows 10 x64
* MicrosoftOffice 2016
* LibreOffice
* Firefox
* GIMP
* Свободно-распространяемый набор утилит, библиотек и Фреймворк Hadoop (Apache Software Foundation)
* QlikSenseDesktop (свободно-распространяемое ПО)

**9.2. Информационно-справочные системы (при необходимости):**

Не используются

**10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ**

Для проведения занятий лекционного типа предлагаются наборы демонстрационного оборудования и учебно-наглядных пособий.

Для изучения дисциплины используется следующее оборудование: аудитория, укомплектованная мебелью для обучающихся и преподавателя, доской, ПК с выходом в интернет, мультимедийным проектором и экраном.

Для самостоятельной работы обучающихся используется аудитория, укомплектованная специализированной мебелью и техническими средствами (ПК с выходом в интернет и обеспечением доступа в электронно-информационно-образовательную среду организации).