1. **Паспорт Образовательной программы**

**«Большие данные»**

|  |  |
| --- | --- |
| **Версия программы** | **1**.2. |
| **Дата Версии** | 12.10.2020 г. |

1. **Сведения о Провайдере**

|  |  |  |
| --- | --- | --- |
| 1.1 | Провайдер | **Московский открытый институт** |
| 1.2 | Логотип образовательной организации |  |
| 1.3 | Провайдер ИНН | 7708142686 |
| 1.4 | Ответственный за программу ФИО | Драгунова Алина Владимировна |
| 1.5 | Ответственный должность | **Руководитель отдела разработки** |
| 1.6 | Ответственный Телефон | **89994567936** |
| 1.7 | Ответственный Е-mail | **ADragunova\_fip@mail.ru** |

1. **Основные Данные**

|  |  |  |
| --- | --- | --- |
| **№** | **Название** | **Описание** |
| 2.1 | Название программы | **Большие данные** |
| 2.2 | Ссылка на страницу программы | <https://docs.google.com/document/d/1Y9VXPa1H5s2IxrPaEIjENFy0GX4hJUy8MRWhSMzwyDI/edit> |
| 2.3 | Формат обучения | Онлайн |
|  | Подтверждение от ОО наличия возможности реализации образовательной программы с применением электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа | В «Московском открытом институте» сформирована возможность передачи данных в форме следующих элементов цифрового следа: детальное описание программы обучения, перечень модулей, описание входной и итоговой диагностик участников – измерение соответствующих программе компетенций участников перед началом и по завершения обучения по образовательной программе, оценки участников образовательной программы, оценка результатов деятельности участников образовательной программы, рефлексию участников образовательной программы. |
| 2.4 | Уровень сложности | Базовый |
| 2.5 | Количество академических часов | **72** |
|  | Практикоориентированный характер образовательной программы: не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы (кол-во академических часов) | Образовательная программа «Большие данные» является практикоориентированной, так как 70% трудоемкости учебной деятельности отведено на практические занятия, что составляет 50 академических часов. Также в рамках самостоятельной работы предусмотрены выполнения практических заданий обучающимся, которые затем проверяются преподавателем и дается подробный комментарий обучающемуся. |
| 2.6 | Стоимость обучения одного обучающегося по образовательной программе, а также предоставление ссылок на 3 (три) аналогичные образовательные программы иных организаций, осуществляющих обучение, для оценки объективности стоимости или обоснование уникальности представленной образовательной программы в случае отсутствия аналогичных образовательных программ на рынке образовательных услуг | **25000 руб.**  **Аналогичные образовательные программы:**  Курс «Data‌ ‌Scientist‌»  Skillbox..29400 руб. 6 модулей - <https://skillbox.ru/course/profession-data-scientist>  Курс «Архитектура Модели Данных». Bigdataschool. 54 000 руб.5 модулей. - <https://www.bigdataschool.ru/bigdata/data-model-architecture.html>  Курс «Аналитика Больших Данных для Руководителей : BDAM». Сетевая академия ЛАНИТ.54000 руб.9 модулей. - <https://academy.ru/catalog/big-data-science/BD-AM.html>  Курс «СПЕЦИАЛИСТ ПО БОЛЬШИМ ДАННЫМ 13.0». New pro lab/ 90 000 руб. 3 модуля. <https://newprolab.com/ru/bigdata> |
| 2.7 | Минимальное количество человек на курсе | **15** |
| 2.8 | Максимальное количество человек на курсе | **60** |
| 2.9 | Данные о количестве слушателей, ранее успешно прошедших обучение по образовательной программе | 125 |
| 2.10 | Формы аттестации | По итогам окончания образовательной программы обучающийся сдает зачет в форме выполнения практического задания (решения кейса на основе реальной профессиональной ситуации) |
|  | Указание на область реализации компетенций цифровой экономики, к которой в большей степени относится образовательная программа, в соответствии с Перечнем областей | Большие данные |

1. **Аннотация программы**

Образовательная программа «Большие данные» ориентирована на развитие у обучающихся навыков применения современных методов сбора, переработки и хранения большого объема данных.

В ходе изучения студенты познакомятся с многофункциональными платформами, и научатся использовать аналитические инструменты, работающие по технологии больших данных в исследовательской деятельности и практической работе.

***Цель изучения программы:***

Формирование знаний и навыков, необходимых для работы с большими данными организации.

***Задачи изучения программы***:

* приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;
* применение статистических и математических методов для анализа больших объемов информации;
* приобретение практических навыков работы с программой R-Studio;
* формирование уровня знаний, умений, опыта деятельности в рамках программы подготовки кадров к Цифровой Экономике, построенных на основе Программы «Цифровая экономика России».

Подготовка по образовательной программе «Большие данные» обеспечит выпускнику востребованность на рынке труда практически в любой области. Высококвалифицированные специалисты, подготовленные по программе «Большие данные», способны применять современные методы машинного обучения (machine learning) и анализа больших данных для извлечения наибольшей ценности из огромного массива данных процессинга.

**Негосударственное образовательное частное**

**учреждение высшего образования**

**«Московский открытый институт»**

**Дополнительная профессиональная образовательная программа повышения квалификации**

**«Большие данные»**

72 час.

Утверждено \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Исполнительный директор Лаврентьева И.Ю.

**ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ**

1. **Цель программы**

Формирование знаний и навыков, необходимых человеку в цифровой среде для использования различных цифровых средств, позволяющих во взаимодействии с другими людьми достигать поставленных целей: анализ больших данных и машинное обучение.

**2. Планируемые результаты обучения:**

Слушатель, успешно освоивший программу, должен **знать:**

* основные термины и понятия из области управления данными;
* основы баз данных;
* базовых понятий прогнозирования
* основы технологий прогнозирования
* методы системного анализа и математической статистики
* методы информационного обслуживания и основных требований информационной безопасности
* о современных программных средствах анализа больших объемов информации
* о базах данных и их информационном обслуживании

Слушатель, успешно освоивший программу, должен **уметь**:

* Определять массивы больших данных
* Строить прогнозы развития социально-политических процессов
* Пользоваться программными средствами анализа больших объемов информации
* Визуализировать исходную информацию и аналитические данные
* Использовать методы системного анализа и математической статистики для решения социально-экономические задачи
* Проводить сравнительный анализ и выбор современных информационно-коммуникационных технологий для решения прикладных задач
* Анализировать и выбирать оптимальные программные средства для анализа данных
* Осуществлять ведение базы данных, обработку и анализ данных

Слушатель, успешно освоивший программу, должен **иметь практический опыт**:

* Сбора и хранения больших данных;
* Применения современных программных средства анализа больших объемов информации.

**3. Категория слушателей**

* 1. Образование: среднее профессиональное образование; общее среднее образование
  2. Наличие опыта профессиональной деятельности: не требуется
  3. Предварительное освоение иных дисциплин/курсов /модулей: Информационные системы.

1. **Учебный план программы «Большие данные»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль** | **Всего, час** | **Виды учебных занятий** | | |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| **1** | Введение в анализ больших данных | **12** | 3 | 3 | 6 |
| **2** | Технологии хранения и обработки больших данных | **10** | 2 | 3 | 5 |
| **3** | Статистические методы анализа данных | **8** | 1 | 3 | 4 |
| **4** | Современные программные средства анализа | **10** | 2 | 3 | 5 |
| **5** | Сбор и хранение больших данных | **10** | 2 | 3 | 5 |
| **6** | Методы обработки и анализа больших данных | **10** | 2 | 3 | 5 |
| **7** | Визуализация исходной информации и данных | **10** | 2 | 3 | **5** |
| **Итоговая аттестация** | | **2** | **зачет** | | |
| **ИТОГО:** | | **72** | **14** | **21** | **35** |

1. **Календарный план-график реализации образовательной** программы

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Наименование учебных модулей** | **Трудоёмкость (час)** | **Сроки обучения** |
| **1** | Введение в анализ больших данных | **12** | **02.11.20 – 03.11.20** |
| **2** | Технологии хранения и обработки больших данных | **10** | **05.11.20 – 07.11.20** |
| **3** | Статистические методы анализа данных | **8** | **08.11.20 – 10.11.20** |
| **4** | Современные программные средства анализа | **10** | **11.11.20 -13.11.20** |
| **5** | Сбор и хранение больших данных | **10** | **14.11.20 – 17.11.20** |
| **6** | Методы обработки и анализа больших данных | **10** | **18.11.20 – 20.11.20** |
| **7** | Визуализация исходной информации и данных | **10** | **21.11.20** |
| **8** | Итоговая аттестация | **2** | **25.11.20** |
| **Всего:** | | **72** | **02.11.20 – 25.11.20** |

1. **Учебно-тематический план программы «Интернет-маркетинг»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль / Тема** | **Всего, час** | **Виды учебных занятий** | | | **Формы контроля** |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| **1** | Введение в анализ больших данных | **12** | 3 | 3 | 6 | Лабораторный практикум №1 |
| **2** | Технологии хранения и обработки больших данных | **10** | 2 | 3 | 5 |
| **3** | Статистические методы анализа данных | **8** | 1 | 3 | 4 | Ситуационный практикум № 2 |
| **4** | Современные программные средства анализа | **10** | 2 | 3 | 5 | Лабораторный № 3 |
| **5** | Сбор и хранение больших данных | **10** | 2 | 3 | 5 | Лабораторный практикум № 4 |
| **6** | Методы обработки и анализа больших данных | **10** | 2 | 3 | 5 |
| **7** | Визуализация исходной информации и данных | **10** | 2 | 3 | **5** | Лабораторный практикум № 5 |

**7. Учебная (рабочая) программа повышения квалификации «Интернет-маркетинг»**

***Модуль 1. Ведение в анализ больших данных (12 ак.часов)***

Основные определения. Задачи анализа больших данных. Вопросы безопасности.Понятие Data Mining. Когнитивный анализ данных. Обзор источников информации для Big Data (открытые источники информации, статистические сборники, опубликованные отчеты и результаты исследований; доступ к закрытой информации).Методики сбора данных.

***Модуль  2. Технологии хранения и обработки больших данных (10 ак. часов)***

Конфигурации и профили в Java ME .Среды разработки, версии Java ME, профили MIDL 1 и 2, их отличительные особенности, требования к аппаратной среде.  Мидлеты. Жизненный цикл мидлета. События высокого и низкого уровня; компоненты пользовательского интерфейса. Сборка и запуск сложных мидлетов. Компиляция, верификация, загрузка мидлета, отладка; сетевое взаимодействие, тонкие клиенты;

***Модуль  3. Статистические методы анализа (8 ак. часа)***

Статистика. Параметрический метод. Непараметрический метод. Номинальные метод.

***Модуль  4. Современные программные средства анализа (10 ак. часов)***

Обзор современных популярных программных средства анализа данных: Statistica, SPSS, Excel, R-Studio и другие;

***Модуль 5. Сбор и хранение больших данных (10 ак. часов)***

Поиск источников информации в сети Интернет: открытые и закрытые источники данных Портал открытых данных РФ. Сохранение данных в программе Excel. Преобразование и первичная обработка данных.

***Модуль 6. Методы обработки и анализа больших данных (10 ак. часов)***

Представление исходных данных в программе R-Studio (векторы, массивы, матрицы, списки, таблицы). Статистическая обработка данных в программах Excel и R-Studio: подсчет описательных статистик, графическое представление данных. Группировка данных, обнаружение значимых корреляций, зависимостей и тенденций в результате анализа имеющейся информации, выявления отношений между данными различного типа. Применение различных методов выделения, извлечения и группировки данных, которые позволяют выявить систематизированные структуры данных и вывести из них правила для принятия решений и прогнозирования их последствий (регрессионный, дисперсионный, кластерный, дискриминантный, факторный анализы).

***Модуль 7. Визуализация исходной информации и данных (10 ак. часов)***

Возможности графического представления информации в программе R-Studio: графические функции отображения одномерных и многомерных данных, графический вывод с использованием графических параметров. Формирование команды. Стили управления.

**Описание практико-ориентированных заданий и кейсов**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Номер темы/модуля** | **Наименование практического занятия** | **Описание** |
| **1.1** | **1, 2** | Лабораторный практикум 1 | Практическое задание, в рамках которого путем проведения экспериментов происходит углубление и закрепление теоретических знаний в интересах профессиональной подготовки. |
| **1.2** | **3** | Лабораторный практикум 2 | Практическое задание, в рамках которого путем проведения экспериментов происходит углубление и закрепление теоретических знаний в интересах профессиональной подготовки. |
| **1.3** | **4** | Лабораторный практикум 3 | Практическое задание, в рамках которого путем проведения экспериментов происходит углубление и закрепление теоретических знаний в интересах профессиональной подготовки. |
| **1.4** | **5, 6** | Лабораторный практикум 4 | Практическое задание, в рамках которого путем проведения экспериментов происходит углубление и закрепление теоретических знаний в интересах профессиональной подготовки. |
| **1.5** | **7** | Лабораторный практикум 5 | Практическое задание, в рамках которого путем проведения экспериментов происходит углубление и закрепление теоретических знаний в интересах профессиональной подготовки. |

**8.Оценочные материалы по образовательной программе**

**8.1. Вопросы тестирования по модулям**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ модуля** | **Вопросы входного тестирования** | **Вопросы промежуточного тестирования** | **Вопросы итогового тестирования** |
| **1** | Что такое «большие данные»? А. **Термин, исползуемый для описания огромного объема данных, который растет со временем в геометрической прогрессии** B. Совокупность данных объективного характера, описывающая некий социум в пространствее и во времени С. Большие данные – это огромная PR-акция крупных вендоров и не более того D. Большие данные – это явление, когда цифровые данные  наиболее полно представляют изучаемый объект.  Какие существуют цели и задачи Data Science? **А)Распознавание видео** Б)Технологии обработки таких данных В)Управление качеством данныхГ)Построение рекомендательных моделей **Д)Сегментация** | Какие существуют методы анализа больших данных? А. Статистический анализ, регрессионный, кластерный B. Факторный анализ, корреляционный, статистический, многомерный. C. **Статистический анализ, регрессионный, факторный, кластерный, корреляционный** D. Регрессионный анализ, факторный, корреляционный.  Какие существуют цели и задачи Big Data? А)**Технологии хранения больших объемов структурированных и не структурированных данных** Б)Технологии предоставления данных потребителю В)Кластеризация Г)**Распознавание текстов** | Какая модель баз данных не существует? А)Сетевая модель данных. **Б)HTML подход к организации баз данных.** В)Сетевая модель данных.  Г)Реляционная модель данных. Д)Объектно-ориентированная модель. |
| **2** | Что такое Озёра данных? А) Фреймворк, который позволяет обеспечивать работу распределённых программ на кластерах из сотен и тысяч узлов. Б) Многофункциональный набор инструментов для визуализации данных. В) **Неструктурированные хранилища для большого количества «сырых» данных, не подвергающихся каким-либо изменениям перед сохранением.** | Какая технология используются для сбора и хранения данных? А)Apache Spark **Б)Apache Hadoop** В)Scikit-learn. Г)Google Chart | Какие технологии используются для Анализа данных? **А)Apache Spark** Б)Apache Hadoop **В)Scikit-learn.** Г)Apache Ranger. |
| **3** | Что не относится к методам статистического анализа? А) Кросстабуляция (сопряжение) Б)Анализ соответствий. А В) Факторный анализ. **Г) Статическое наблюдение** | Сущность кластерного анализа? **А)Статистический метод классификации объектов по группам за счет выявления наперед не известных общих признаков** Б)Методика сбора данных из большого количества источников. В) В основе методик оценки настроений потребителей лежат технологии распознавания естественного языка человека  Какие формы может иметь стастического наблюдение?  **А)Отчетность Б)Регистр В) Специально организованное наблюдение** Г)Периодический учет | Данный вид анализа предназначен для анализа зависимостей между двумя списками признаков (независимых переменных), характеризующих объекты. А)Частотный анализ **Б) Канонический анализ.** В) Корреляционный анализ Г)Регрессионный анализ  Перепись населения каждой страны должна отвечать единым принципам: **1) всеобщность**; 2) разнообразие программ;3) цикличность проведения; **4) одномоментность.** |
| **4** | Что не является преимуществом работы с данными в программе R-Studio? А)бесплатная и кроссплатформенная;  Б)богатый арсенал стат. методов;  В)качественная векторная графика; **Г)большой объем информации на русском языке** | При какой модели вначале используются наиболее простые алгоритмы, а часть данных, которые можно обработать при помощи таких алгоритмов и которые бессмысленно обрабатывать с использованием более сложных методов, анализируется и исключается из дальнейшей обработки. Оставшиеся данные передаются на следующий этап обработки, где используются более сложные алгоритмы, и так далее по цепочке. На последнем узле сценария обработки применяются самые сложные алгоритмы, но объем анализируемых данных во много раз меньше первоначальной выборки.? **А) Комбинирование моделей** Б) Параллельная обработка В) Репрезентативные выборки | Около 70 % ассортимента составляют товары группы Z. Еще около 25 % - товары группы Y и только примерно 5 % - товары группы X. Таким образом, для скольких % актуально построение и применение сложных модулей? А) 25% **Б)30%** В) 50% г)70% |
| **5** | Что позволяет делать описательный анализ? А) позволяет пользователям оценить вероятность тех или иных событий в будущем. В качестве примеров можно привести системы прогнозирования, заблаговременных предупреждений и обнаружения мошенничества, а также приложения для профилактического обслуживания. Б) **помогает пользователям ответить на вопрос: «Что произошло и почему?» В качестве примера можно привести традиционную среду для запросов и отчетов с панелями управления и системами оценок.** В) формирует для пользователя определенные рекомендации (предписания). Он помогает ответить на вопрос: «Что делать, если произойдет событие Х?» | От чего может зависеть показатель дедупликации для конкретных задач?  А)тип данных  Б)степень возможной В)изменяемости данных  Г)политика резервного копирования  **Д)Все вышеперечисленное** | Какие виды источников данных относятся к Структурированному типу?  **А) Базы данных ERP, CRM-си- стем**  **Б)Базы данных прикладных систем (банковские, торговые, страховые)**  В)Веб-данные (веблоги)  Г)Контент корпоративного портала |
| **6** | Что не является программным средством анализа данных?- Statistica, -SPSS,- Excel; **-Word** | Продолжите алгоритм анализа данных: Подготовка данных - Стадия Map - Стадия Reduce - ...: А)**Вывод результатов** Б)Запуск скрипта В)Загрузка данных Г)Прогноз | Чего нет в цепочке Фрагментного параллелизма? А)Фрагментация Б)Обработка запроса В)Слияние **Г)Изолирование**  Что такое A/B тестирование? **А)Методика, в которой контрольная выборка поочередно сравнивается с другими** Б) Статистический метод классификации объектов по группам за счет выявления наперед не известных общих признаков. В) В этом методе задействуется множество предикативных моделей за счет чего повышается качество сделанных прогнозов г) Набор методик, которые позволяют создать математическую модель наперед заданного вероятного сценария развития событий |
| **7** | Что отличает визуализацию данных от графического интерфейса? **А) Краткость Б) Масштабируемость** В)Контр интуитивность Г)Точность | Какие данные нужны для удачного способа визуализировать Диаграммы производительности? А)Комбинированные данные Б)Структурная информация **В)Данные об аспектах работы** Г)Количественная информация | Что относится к основным критериям влияющим на принятие решений на основе данных? Количество предоставленной информации А)Наличие альтернативных вариантов Б)Временные рамки В)Альтернативные точки зрения Г)Все вышеперечисленное |

**8.2.**  **описание показателей и критериев оценивания, шкалы оценивания**

За каждый ситуационный практикум, обучающийся может получить максимально 20 баллов.

20-16 баллов – работа выполнена в срок, самостоятельно, правильно поняты и использованы соответствующие формулы, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, правильно выбраны совместимые комплектующие, сделаны необходимые выводы, хорошо аргументированы, даны исчерпывающие ответы на все поставленные вопросы;

15-11 баллов – работа выполнена в срок, самостоятельно, правильно поняты и использованы соответствующие формулы, правильно определены соответствующие спецификации, использована требуемая информация, правильно выполнены требуемые расчеты, правильно выбраны совместимые комплектующие, необходимые выводы сделаны частично, хорошо аргументированы, даны ответы на все поставленные вопросы;

10-6 баллов – работа выполнена в срок, в основном самостоятельно, использованы соответствующие формулы; определены соответствующие спецификации, имеются ошибки в расчетах; выбраны совместимые комплектующие необходимые, выводы сделаны частично, слабо аргументированы, даны ответы не на все вопросы;

5-1 баллов – обучающийся подготовил работу несамостоятельно или не завершил в срок, описание спецификации содержит незначительные ошибки, выводы и ответы на вопросы отсутствуют.

**8.3.**  **примеры контрольных заданий по модулям или всей образовательной программе**

Вопрос 1. Какая модель баз данных не существует?

А)Сетевая модель данных.

Б)HTML подход к организации баз данных.

В)Сетевая модель данных.

Г)Реляционная модель данных.

Д)Объектно-ориентированная модель.

Вопрос 2. Какие технологии используются для Анализа данных?

А)Apache Spark

Б)Apache Hadoop

В)Scikit-learn.

Г)Apache Ranger.

Вопрос 3.Данный вид анализа предназначен для анализа зависимостей между двумя списками признаков (независимых переменных), характеризующих объекты.

А)Частотный анализ

Б) Канонический анализ.

В) Корреляционный анализ

Г)Регрессионный анализ

Вопрос 4. Перепись населения каждой страны должна отвечать единым принципам:

А) всеобщность;

Б) разнообразие программ;

В) цикличность проведения;

Г) одномоментность.

Вопрос 5. Чего нет в цепочке Фрагментного параллелизма?

А)Фрагментация

Б)Обработка запроса

В)Слияние

Г)Изолирование

Вопрос 6. Что такое A/B тестирование?

А)Методика, в которой контрольная выборка поочередно сравнивается с другими

Б) Статистический метод классификации объектов по группам за счет выявления наперед не известных общих признаков.

В) В этом методе задействуется множество предикативных моделей за счет чего повышается качество сделанных прогнозов

Г) Набор методик, которые позволяют создать математическую модель наперед заданного вероятного сценария развития событий

Что относится к основным критериям влияющим на принятие решений на основе данных?

Вопрос 7. Колличество предоставленной информации

А)Наличие альтернативных вариантов

Б)Временные рамки

В)Альтернативные точки зрения

Г)Все вышеперечисленное

Вопрос 8. Около 70 % ассортимента составляют товары группы Z. Еще около 25 % - товары группы Y и только примерно 5 % - товары группы X. Таким образом, для скольких % актуально построение и применение сложных модулей?

А) 25%

Б)30%

В) 50%

г)70%

Вопрос 9. Какие виды источников данных относятся к Структурированному типу?

А) Базы данных ERP, CRM-си- стем

Б)Базы данных прикладных систем (банковские, торговые, страховые)

В)Веб-данные (веблоги)

Г)Контент корпоративного портала

***Задание 1.***

имеется csv-лог рекламной системы вида:

<user\_id>,<country>,<city>,<campaign\_id>,<creative\_id>,<payment></p> 11111,RU,Moscow,2,4,0.3 22222,RU,Voronezh,2,3,0.2 13413,UA,Kiev,4,11,0.7 …

Необходимо рассчитать среднюю стоимость показа рекламы по городам России.

***Задание 2.***

скачайте архив из случайных новостей с сайта [lenta.ru](http://lenta.ru/). Скачать архив можно по [ссылке](https://www.dropbox.com/s/opp5psid1x3jt41/lenta_articles.tar.gz).

Имеется набор документов. Необходимо для каждого слова, встречающегося в наборе документов, посчитать, сколько раз встречается слово в наборе.

***Задание 3.***

имеется набор текстовых документов, необходимо посчитать, сколько слов встретилось от 1 до 1000 раз в наборе, сколько слов от 1001 до 2000, сколько от 2001 до 3000 и так далее.

***Задание 4.***

Вам представлены данные Бостонского полицейского управления (2015-2018гг.), содержащий записи отчетов о происшествиях, датах и месте совершении преступлений. Вам необходимо выявить наиболее распространенные виды преступлений, а также построить рейтинг районов, где совершается большее количество преступлений. Проанализируйте полученные результаты и сформулируйте авторскую позицию о проблемных районах города, с предложением по нормализации преступности в рамках городской политики обеспечения безопасности. Пример датасета из открытого репозитория Kaggle: https://www.kaggle.com/AnalyzeBoston/ crimes-in-boston

***Задание 5.***

Вам представлены данные электорального цикла Президентских выборов США 2016г. (включая первичные результаты). Датасет содержит демографические данные по округам из переписи населения и результаты голосования. Вам необходимо определить по округам: долю голосов, полученных победившим в данном округе кандидатом; демографические показатели голосовавших в процентах (возраст - датасет позволяет определить старше/младше 25 лет; образование; расу - датасет позволяет определить трассовую принадлежность). После этого построить рейтинг кандидатов (топ 5) с процентным распределением голосов. Проанализируйте полученные результаты и сформулируйте авторскую позицию по распределению голосов за конкретных кандидатов в зависимости от уровня образования, возраста и расовой принадлежности. Пример датасета из открытого репозитория Kaggle: https://www.kaggle.com/benhamner/2016- us-election/kernels (USA), схожий датасет относитесь электорального цикла Венгрии: https:// www.kaggle.com/brezniczky/hungarian-parliamentary-elections-2018-dataset

***Задание 6.***

У предприятия, осуществляющего свою деятельность на рынке массовых розничных услуг, стоит задача формирования и анализа клиентской базы с целью определения кластеров покупателей и подготовки отдельный предложений для каждого кластера. По каким критериям возможно формирование кластеров? Какие данные можно использовать в этих целях? Предложите свой вариант кластерной сегментации на любом розничном рынке товаров или услуг

***Задание 7.***

На рынке кредитования физических лиц в России осуществляют свою деятельность банки и микрофинансовые организации. Существует ли разница в скоринговых моделях этих организаций? По каким ключевым критериям формируются данные скоринговые модели? Предложите свой вариант для основы скоринговой модели для микрофинансовой организации.

**8.4.**  **тесты и обучающие задачи (кейсы), иные практикоориентированные формы заданий**

***Лабораторный практикум №1.***

**Задание 1.** Используя компьютерные технологии закрепите навык работы с данными и приготовьте их в текстовом редакторе. Освоийте команды для чтения данных из текстового файла.

***Лабораторный практикум №2.***

**Задание 1.**

Используя компьютерные технологии потренируйтесь вычислять общих характеристик выборки данных. Закрепите навык оценивать общую тенденцию, находить ошибки в данных и пропуски, выполнять одномерные статистические тесты.

Изучите графические команды и научитесь строить графики. Построить графики динамики курса иностранной валюты за последние две недели.

Документ с результатами сохраните в рабочей папке под названием ЛП№2.

**Задание 2.**

1. Сгенерировать большой массив данных и записать в один файл. Установить пакет purrr. Записать массив данных по частям в несколько файлов . Сформировать репрезентативная выборку ограниченного размера.
2. Выполнить загрузку данных с использованием различных стратегий. Сделать выводы. Установить пакеты data.table, sqldf, ff.
3. Установить и загрузить библиотеки sqldf и nycflights13.Ознакомиться со структурой набора данных flights. Вычислить количество наблюдений для всех перевозчиков carrier в таблице flights. Отобразить в консоли значения полейdep\_time,dep\_delay,arr\_time, carrier, tailnum из таблицы flights (первые и последние 5 строк). Вычислить среднее время задержки прибытия (mean\_arr\_delay) и отправления (mean\_dep\_delay) для различных перевозчиков (carrier) .
4. Сгенерировать data.frame с тремя столбцами и 100 строками. Преобразовать данные из широкого в длинный формат.

***Лабораторный практикум №3.***

**Задание 1.**

1. Сгенерировать вектор (массив, таблица данных) и добавить в него элементы NA. Очистить данные с использованием функцииis.na() [1].

2. Сгенерировать таблицу данных с числовыми и текстовые столбцами. Очистить данные с функции complete.cases() [1].

2. Сгенерировать числовую таблицу данных с пропусками. С использованием функции preProcess из пакета caret заполнить пропуски предсказанными значениями (среднее, медиана).

3. Сгенерировать два числовых набора данных, добавить в них выбросы. С использованием функции boxplot обнаружить выбросы и удалить их.

4. Сгенерируйте таблицу данных, в которой дублируются строки. Удалите строки с использованием функций unique(), duplicated(). Сравните результаты .

5. Обработать пропуски в данных с использованием пакета mice.

6. Разобрать пример с мультиколлинеарностью .

**Задание 2.**

1. Установить пакет CARET, выполнить команду names(getModelInfo()), ознакомиться со списком доступных методов выбора признаков. Выполните графический разведочный анализ данных с использование функции featurePlot() для набора данных из справочного файла пакета CARET: x <- matrix(rnorm(50\*5),ncol=5) y <- factor(rep(c("A", "B"), 25)) Сохранить полученные графики в \*.jpg файлы. Сделать выводы.

2. С использование функций из пакета Fselector определить важность признаков для решения задачи классификации. Использовать набор data(iris). Сделать выводы.

3. Установите пакет Boruta и проведите выбор признаков для набора данных data("Ozone") . Построить график boxplot, сделать выводы

**Задание 3.**

1. Выполните классификацию k-ближайших соседей с использованием функции knn() из пакета class на наборе данных iris. Проведите нормализацию данных, разделите выборку на обучающую и тестовую. Оцените построенную модель с использованием функции CrossTable() из пакета gmodels. Постройте матрицу ошибок и диагональную оценку качества прогноза (diagonal mark quality prediction).

2. Рассмотрите пример реализации метода опорных векторов с использованием функции svm() из пакета e1071. Постройте линейный классификатор для прогнозирования. Для подбора параметров модели выполните перекрестную проверку с делением исходной выборки на 10 равных частей (cross=10) .

3. Выполните расчет главных компонент с использованием пакета vegan() и его функции rda(). Постройте ординационную диаграмму методом PCA и сделайте выводы.

**Задание 4.**

1. Установите пакет sparklyr, установите Java Virtual Machine (JVM). Подключитесь к локальному Spark-кластеру. Загрузите таблицу flights из пакета nycflights13 в Sparkкластер . Выполните запросы (задание 6, Лабораторная работа. Сравните результаты, сделайте выводы.

2. Настройте для использования Hadoop , подсчитайте количество слов в файле \*.txt c использованием HDFS . Файл сгенерировать самостоятельно.

3. Установите MongoDB . Подключите библиотеку mongolite. Выполните пример для набора iris c использование функции mongo() из видеоролика. Сохраните код и сделайте выводы.

**Задание 5.**

Выберите два любых кейса . входные данные: стык моделей и технологий, которые можно использовать для решения выбранных кейсов. Используя компьютерные технологии приведите иллюстративные примеры с использованием R. Прикрепите свои выводы в документ под названием ЛП№3 в рабочей папке.

***Лабораторный практикум №4.***

**Задание 1.**

Познакомиться со средой R-Studio (установка и настройка)

**Задание 2.**

Используя компьютерные технологии закрепите работу с матрицами списками и таблицами данных в R-Studio.

**Задание 3.**

Используя компьютерные технологии проведите анализ временного ряда на предмет прогнозирования будущего с помощью RStudio. Тренд и период колебаний, построение временного ряда, прогнозирование. Выполнить на примере колебания курса иностранной валюты.

**Задание 4.**

Используя компьютерные технологии проведите анализ связей двух выборок данных с помощью R-Studio. Научиться выполнять анализ двух таблиц, строить таблицы сопряженности, выполнять анализ корреляций, выполнять регрессионный анализ. На примере анализа данных цены на нефть и курса доллара США.

***Лабораторный практикум №5.***

**Задание 1.**

Изучить графические команды и используя компьютерные технологии научиться строить графики. Построить графики динамики курса иностранной валюты за последние две недели.

Сохраните документ с именем ЛП№5 в рабочей папке.

***Типовые задания для проведения итоговой аттестации обучающихся***

Итоговая аттестация по образовательной программе «Большие данные» проводится в форме зачета.

**8.5. описание процедуры оценивания результатов обучения**

Текущий контроль и промежуточная аттестации по образовательной программе «Интернет-маркетинг»ведется в соответствии Положением о текущем контроле успеваемости и промежуточной аттестации в «Московском открытом институте».

**Текущий контроль успеваемости**

Для повышения эффективности текущего контроля и последующей промежуточной аттестации слушателей осуществляется структурирование каждой дисциплины на темы, проводится регулярная оценка знаний, умений и компетенций слушателей.

Основными видами контроля знаний, умений и навыков в течение каждой темы являются защиты домашних заданий, контрольные работы, результатов ситуационных практикумов.

Текущий контроль по теме осуществляется по графику учебного процесса. Сроки контрольных мероприятий (КМ) и сроки подведения итогов по теме отображаются в рабочих учебных планах. Слушатель должен выполнить все контрольные мероприятия, предусмотренные в теме к указанному сроку, после чего преподаватель проставляет балльные оценки, набранные студентом по результатам текущего контроля.

Контрольное мероприятие считается выполненным, если за него слушатель получил оценку в баллах, не ниже минимальной оценки, установленной программой дисциплины по данному мероприятию.

Слушатели, не сдавшие контрольное мероприятие в установленный срок, продолжают работать над ним в соответствие с порядком, принятым Университетом.

**Промежуточная аттестация**

Формой промежуточной аттестации является зачет.

Формы промежуточной аттестации, порядок начисления баллов и фонды оценочных средств для промежуточной аттестации разрабатываются кафедрой, исходя из специфики образовательной программы.

**Зачет**

В рамках системы контроля успеваемости слушателей, зачет по образовательной программе формируется набором предусмотренной в образовательной программе суммы баллов, при выполнении им всех контрольных мероприятий.

Вопросы для контроля ключевых, базовых знаний, умений и навыков по образовательной программе формулируются ведущими преподавателями, утверждаются заведующим кафедрой и вносятся в программу дисциплины. Контрольные вопросы для проверки ключевых результатов обучения по образовательной программе обеспечивают возможность объективной независимой оценки знаний, умений и навыков, приобретенных студентом.

**9. Организационно-педагогические условия реализации программы**

**9.1. Кадровое обеспечение программы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **Фамилия, имя, отчество (при наличии)** | **Место основной работы и должность, ученая степень и ученое звание (при наличии)** | **Ссылки на веб-страницы с портфолио (при наличии)** | **Фото в формате jpeg** | **Отметка о полученном согласии на обработку персональных данных** |
| **1** | | Масягин Игорь | R&D-разработчик, **Lamoda**  Преподаватель очных и онлайн [программ по Data Science и Data Engineer.](https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwi3-KPVmbXsAhWI43cKHXkYAEIYABABGgJlZg&ae=2&ohost=www.google.com&cid=CAESQOD2Wqydz-EB1oN5iEhPqStL0RnHs-pUiXCkLZ_C9ZMVcIgXR__t0xbWdpuAP-s-HgwILI73zS0ni4tOA8ahlzQ&sig=AOD64_0G3I0-VLqpHnImIXsamG-BdopfNg&q&adurl&ved=2ahUKEwjJnJrVmbXsAhVB_CoKHVE4DRkQ0Qx6BAgOEAE&dct=1) | **-** |  | **Согласен** |
| **2** | Сапрыкин Артур Игоревич | Преподаватель и методист курса по «Data Science» в Нетология групп. | **-** | https://xn--47-6kceulec0czar2ay.xn--p1ai/img/speakers/card/saprykin-435f9132f8.jpg | **Согласен** |

**9.2. Учебно-методическое обеспечение и информационное сопровождение**

|  |  |
| --- | --- |
| **Учебно-методические материалы** | |
| Методы, формы и технологии | Методические разработки,  материалы курса, учебная литература |
| Лекции, практические занятия | 1. Прогнозирование и планирование в условиях рынка : учеб. пособие / Т.Н. Бабич, И.А. Козьева, Ю.В. Вертакова, Э.Н. Кузьбожев. ? М. : ИНФРА-М, 2017. ? 336 с. ? (Высшее образование: Бакалавриат) - Режим доступа: http://znanium.com/catalog/product/851194 |
| 2. Социально-экономическое прогнозирование: Учебное пособие / Герасимов А.Н., Громов Е.И., Скрипниченко Ю.С. - М.:СтГАУ - 'Агрус', 2017. - 144 с.: Режим доступа: http://znanium.com/catalog/product/975933 |
| 3. Методы хранения и обработки данных: Учебник / Дадян Э.Г. - М.:НИЦ ИНФРА-М, 2018: - Режим доступа: http://znanium.com/catalog/product/989190 |
| 4.Современные базы данных. Основы. Часть 1: Учебное пособие / Дадян Э.Г. - М.:НИЦ ИНФРА-М, 2017. - 88 с.: 60x90 1/16 ISBN 978-5-16-106526-6 (online) - Режим доступа: http://znanium.com/catalog/product/959289 |
| 5.Проектирование современных баз данных. Практикум: Учебно-методическое пособие / Дадян Э.Г. - М.:НИЦ ИНФРА-М, 2017. - 84 с.: 60x90 1/16 ISBN 978-5-16-106528-0 (online) - Режим доступа: http://znanium.com/catalog/product/959294 |

|  |  |
| --- | --- |
| **Информационное сопровождение** | |
| Электронные  образовательные ресурсы | Электронные  информационные ресурсы |
| Электронная библиотечная система «Университетская библиотека онлайн» - https://biblioclub.ru | Большие данные –  https://habr.com/ru/hub/bigdata/ |
| Федеральный портал «Российское образование». Каталог образовательных ресурсов - <http://www.edu.ru> | Информационно-аналитический портал по компьютерной тематике - <https://habr.com/ru/> |
| Открытый образовательный видеопортал UniverTV.ru - http://univertv.ru |  |

**9.3. Материально-технические условия реализации программы**

|  |  |
| --- | --- |
| Вид занятий | Наименование оборудования,  программного обеспечения |
| Лекция | Компьютер, экран, мышь, клавиатура  Microsoft Windows 7 pro и выше  Программное обеспечение Microsoft Office Professional  Сеть:  • скорость соединения от 2 Мб/с.  Оборудование:  • наушники — чтобы не появлялось эффекта эха.  К платформе можно подключиться с помощью персонального компьютера, мобильного устройства или планшета. |
| Практические занятия | Компьютер, экран, мышь, клавиатура  Microsoft Windows 7 pro и выше  Программное обеспечение Microsoft Office Professional  Сеть:  • скорость соединения от 2 Мб/с.  Оборудование:  • наушники — чтобы не появлялось эффекта эха.  К платформе можно подключиться с помощью персонального компьютера, мобильного устройства или планшета. |

# Сценарии профессиональной траектории граждан

|  |  |
| --- | --- |
| **Цели получения персонального цифрового сертификата** | |
| **текущий статус** | **цель** |
| **Трудоустройство** | |
| состоящий на учете в Центре занятости | трудоустроенный |
| безработный |
| безработный по состоянию здоровья |
| **Развитие компетенций в текущей сфере занятости** | |
| работающий по найму в организации, на предприятии | развитие профессиональных качеств |
| работающий по найму в организации, на предприятии | смена работы без изменения сферы профессиональной деятельности |
| временно отсутствующий на рабочем месте (декрет, отпуск по уходу за ребенком и др.) | повышение уровня дохода |
| **Переход в новую сферу занятости** | |
| освоение новой сферы занятости | самозанятый, ИП/бизнесмен |
| освоение смежных профессиональных областей | повышение уровня дохода, расширение профессиональной деятельности |

**ПАСПОРТ КОМПЕТЕНЦИЙ**

|  |
| --- |
| Интернет-маркетинг |
| (наименование дополнительной профессиональной образовательной программы повышения квалификации) |
| Негосударственное образовательное частное учреждение высшего образования «Московский открытый институт» |
| (наименование организации, реализующей дополнительную профессиональную образовательную программу повышения квалификации) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. | **Наименование компетенции** | | Управление информацией и данными | |
| 2. | **Указание типа компетенции** | общекультурная/  универсальная | - | |
| общепрофессиональная | - | |
| профессиональная | профессиональная | |
| профессионально­‑ специализированная | - | |
| 3. | **Определение, содержание и основные  сущностные характеристики компетенции** | | Под компетенцией понимается способность генерировать новые идеи для решения задач цифровой экономики, абстрагироваться от стандартных моделей: перестраивать сложившиеся способы решения задач, выдвигать альтернативные варианты действий с целью выработки новых оптимальных алгоритмов.  Слушатель должен:  **знать:**   * основные термины и понятия из области управления данными; * основы баз данных; * базовых понятий прогнозирования * основы технологий прогнозирования * методы системного анализа и математической статистики * методы информационного обслуживания и основных требований информационной безопасности * о современных программных средствах анализа больших объемов информации * о базах данных и их информационном обслуживании   **уметь**:   * Определять массивы больших данных * Строить прогнозы развития социально-политических процессов * Пользоваться программными средствами анализа больших объемов информации * Визуализировать исходную информацию и аналитические данные * Использовать методы системного анализа и математической статистики для решения социально-экономические задачи * Проводить сравнительный анализ и выбор современных информационно-коммуникационных технологий для решения прикладных задач * Анализировать и выбирать оптимальные программные средства для анализа данных * Осуществлять ведение базы данных, обработку и анализ данных   Владеть:   * Навыками сбора и хранения больших данных; * Навыками применения современных программных средства анализа больших объемов информации. | |
| 4. | **Дескриптор знаний, умений и навыков по уровням** | | **Уровни  сформированности компетенции обучающегося** | **Индикаторы** |
|  | | **Начальный  уровень**  (Компетенция недостаточно развита. Частично проявляет навыки, входящие в состав компетенции. Пытается, стремится проявлять нужные навыки, понимает их необходимость, но у него не всегда получается.) | Знает: Основы построения и использования систем больших данных  Умеет: самостоятельно работать с различными источниками информации (интернет, печатные издания)  Владеет: основными методами и способами получения и хранения информации |
|  | | **Базовый уровень**    (Уверенно владеет навыками, способен, проявлять соответствующие навыки в ситуациях с элементами неопределённости, сложности.) | Знает: особенности работа с большими данными  Умеет: владеть способами и средствами переработки информации;  Владеет: навыками обобщения имеющейся информации |
|  |  | | **Продвинутый**  (Владеет сложными навыками, способен активно влиять на происходящее, проявлять соответствующие навыки в ситуациях повышенной сложности.) | Знает: основные способы и средства переработки информации  Умеет: представлять результаты своих исследований.  Владеет: навыками работы с аппаратно-программными средствами обработки больших данных; |
|  |  | | **Профессиональный**  (Владеет сложными навыками, создает новые решения для сложных проблем со многими взаимодействующими факторами, предлагает новые идеи и процессы, способен активно влиять на происходящее, проявлять соответствующие навыки в ситуациях повышенной сложности.) | Знает: Основные технологии прогнозирования  Умеет: Анализировать кластеры больших данных  Владеет: Современными технологиями создания и обслуживания больших данных |
| 5. | **Характеристика взаимосвязи данной компетенции с другими компетенциями/ необходимость владения другими компетенциями для формирования данной компетенции** | | Данная компетенция тесно взаимосвязана со следующими компетенциями цифровой экономики: управление информацией и данными, критическое мышление в цифровой среде.  Для успешного формирования данной компетенции необходимо владение компетенциями цифровой грамотности. | |
| 6. | **Средства и технологии оценки** | | Тесты, ситуационные практикумы (проблемные задания по тематике образовательной программы, в которых слушателям предлагают осмыслить реальную профессионально-ориентированную ситуацию, необходимую для решения проблемы, рассмотренной в теме). | |