**Модуль А. Парсинг и предобработка данных**

**Задания:**

|  |
| --- |
| 1.1 Загрузка данных |
| Создать базу данных, содержащую структуру, согласно исходным данным. Загрузить данные из предоставленных ресурсов. Данные необходимо объединить в один набор. Загрузить содержимое каждого набора в базу данных, предусмотреть обработку исключений при загрузке. В наборах данных для последующего обучения должны быть включены все записи, присутствующие в исходных наборах. Базу данных необходимо поддерживать в актуальном состоянии на протяжении всей работы. Для решения задачи должны использоваться также дополнительные источники. |

|  |
| --- |
| 1.2 Предобработка данных и выделение значимых атрибутов |
| Для представленного набора данных предполагается применить методы кластерного анализа для создания групп, схожих по характеристикам. Такие кластеры в будущем помогут выявить динамику зависимости различных параметров для оптимизации ресурсов предприятия пассажирских перевозок.  Необходимо определить, какие атрибуты имеют наибольшее влияние на определение таких групп объектов (кластеров), и оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора. |

|  |
| --- |
| 1.3 Описание структуры набора данных |
| Для каждого атрибута подготовьте описание, содержащее текстовое представление (расшифровка, перевод, назначение). Выполнить проверку нормальности распределения значений в атрибутах. Для определения отклонений от нормы провести графическую и статистические оценки. Обосновать выбор методов оценки. На основе полученных результатов сделать вывод о типе распределения и мере скошенности. |

|  |
| --- |
| 1.4 Определение численных характеристик наборов данных |
| Определите характеристики, удовлетворяющие следующим требованиям:   1. Общая загруженность станции 2. Пропускная способность в процентах по отношению к средней пропускной способности 3. Количественные характеристики входного и выходного оборудования 4. Временные интервалы поездки 5. Продолжительность поездки |

|  |
| --- |
| 1.5 Формирование итогового набора данных |
| Задача кластеризации предполагает определение наиболее предпочтительных условий для поездки. Поэтому необходимо переформировать набор данных таким образом, чтобы в наборе отражались все возможные маршруты с указанием всех возможных характеристик, определённых выше, а также других характеристик, полученных из внешних источников |

|  |
| --- |
| 1.5 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам модуля, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов |

**Модуль Б. Разведочный анализ данных**

**Задания:**

|  |
| --- |
| 2.1 Построение аналитической системы |
| Для своевременного принятия решения на основе данных, минимизации ошибок и быстрого достижения целей, необходимо построить аналитический дашборд. Интерактивная информационная панель должна наглядно представлять, визуализировать и анализировать данные. Для этого необходимо подключить к системе источник данных, т.е. подключиться к базе данных, настроить регламент обновления и уровни доступа.  Аналитическая система должна позволить пользователю:  – Автоматизировать работу с данными;  – Анализировать все в реальном времени;  – Видеть все главные метрики на одном экране;  – Легко считывать информацию и делать выводы. |

|  |
| --- |
| 2.2 Разработка функционала аналитической системы |
| Разработанная информационная панель должна позволить пользователю проводить анализ загруженности станций в течение дня.  Необходимо вычислить такие характеристики, как загруженность станции в процентах относительно максимальной загруженности в заданный интервал времени, реальную пропускную способность станции в заданный интервал времени, топ самых загруженных станций, среднее количество пассажиров на станциях в течение заданного интервала времени, количество станций без инфраструктуры для маломобильных граждан. Дашборд должен позволять выбирать станции и задавать временные интервалы.  Визуализируйте динамику полученных метрик во времени. В дашборде должны быть отражены ключевые сводные показатели. |

**После выполнения пунктов 2.1 и 2.2 необходимо продемонстрировать выполненную работу экспертам**

|  |
| --- |
| 2.3 Кластеризация набора данных |
| Выберите модель кластеризации данных. Задача кластеризации – сегментация поездок, определение схожих групп объектов в каждом наборе данных. В результате должно быть выявлено 3 кластера: в первый кластер попадают все маршруты, для которых поездка в заданное время оптимальна, во второй кластер попадают маршруты с нежелательным временем поездки, в третий кластер попадают маршруты, от поездок по которым следует воздержаться в заданный период времени. |

|  |
| --- |
| 2.4 Вычисление показателей производительности кластеризации |
| Выберите метрику оценки качества кластеризации. Обоснуйте выбор методов и приемов. Выполните визуальный анализ кластерных структур и оценки качества кластеризации. Определите лучший алгоритм кластеризации на основе выбранной метрики. |

|  |
| --- |
| 2.5 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам модуля в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов |

2. СПЕЦИАЛЬНЫЕ ПРАВИЛА КОМПЕТЕНЦИИ

2.1. Личный инструмент конкурсанта

### Участники могут использовать защиту для ушей.

### Участники могут принести с собой свои клавиатуры, мышки и коврики для мышек. Все принесенные клавиатуры, мышки и коврики должны быть предварительно сданы на проверку Техническому эксперту. Запрещено использование Клавиатур и мышек с подключением по беспроводным каналам. Устройства ввода не должны быть программируемыми.

2.2.Материалы, оборудование и инструменты, запрещенные на площадке

1. Участники не должны приносить:

* Дополнительные программы.
* Мобильные телефоны.
* Портативные электронные устройства (планшеты, и т.п.).
* Устройства для хранения информации (флэш-накопители, диски, и т.п.).

1. Оборудование не должно иметь доступ к внутренним локальным сетевым ресурсам (например, общие папки, FTP). Запрещается использовать личные репозитории, требующие авторизации для просмотра