**Модуль А. Парсинг и предобработка данных**

Время на выполнение модуля – 3 часа

**Задания:**

|  |
| --- |
| 1.1 Загрузка данных |
| Загрузить данные с ресурсов. Необходимо оставить не менее 40% от исходного количества записей. В наборах данных для последующего обучения должны быть включены записи по каждому месяцу, присутствующему в исходных наборах, в пропорциональном количестве. |

|  |
| --- |
| 1.2 Предобработка данных и выделение значимых атрибутов |
| Для каждого набора данных (по годам) предполагается применить методы кластерного анализа для создания групп сходных объектов. 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 количество записей в каждом наборе (по годам), удовлетворяющих следующим требованиям:   1. Районы, из которых чаще всего производилась посадка пассажиров, где чаевых было оставлено не менее 15% от суммы оплаты за поездку (длина поездки не должна превышать 2 км); 2. Стоимость на километр поездки не превышала среднюю стоимость всех поездок на километр по такому же тарифу; 3. Среднее количество пассажиров на поездку, пользующихся услугами такси с самыми популярными тарифами.   Сравните полученные численные значения и выявите динамику изменений по годам. |

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