Вариант № 2

Задание для демонстрационного экзамена по комплекту оценочной документации

|  |  |
| --- | --- |
| **Номер компетенции** | F5 |
| **Название компетенции** | Машинное обучение и большие данные |
| **Номер КОД** | 1.1 |

# Описание задания.

# Описание модуля 1:

**Исходные файлы:**

1. [REGIONS.](https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv)zip (Архив с исходными данными)

**Результаты работы:**

1. Data.zip (Предобработанные данные)
2. Report.html + Report.ipynb (Отчет о проделанной работе)

**ВВЕДЕНИЕ**

Вам предстоит разработать систему по выявлению опасных дорожных участков и предотвращению будущих дорожно-транспортных происшествий на территории регионов Российской Федерации. Данные по ДТП доступны для загрузки на странице <https://dtp-stat.ru/opendata/> и представлены в формате .geojson по всем регионам. Каждый файл имеет следующую структуру:

{

"id": 384094, # идентификатор

"tags": ["Дорожно-транспортные происшествия"], # показатели с официального сайта ГИБДД

"light": "Светлое время суток", # время суток

"point": {"lat": 50.6039, "long": 36.5578}, # координаты

"nearby": [ "Нерегулируемый перекрёсток неравнозначных улиц (дорог)", "Жилые дома индивидуальной застройки"], # координаты

"region": "Белгород", # город/район

"address": "г Белгород, ул Сумская, 30", # адрес

"weather": ["Ясно"], # погода

"category": "Столкновение", # тип ДТП

"datetime": "2017-08-05 13:06:00", # дата и время

"severity": "Легкий", # тяжесть ДТП/вред здоровью

"vehicles": [ # участники – транспортные средства

{

"year": 2010, # год производства транспортного средства

"brand": "ВАЗ", # марка транспортного средства

"color": "Иные цвета", # цвет транспортного средства

"model": "Priora", # модель транспортного средства

"category": "С-класс (малый средний, компактный) до 4,3 м", # категория транспортного средства

"participants": [ # участники внутри транспортных средств

{

"role": "Водитель", # роль участника

"gender": "Женский", # пол участника

"violations": [], # нарушения правил участником

"health\_status": "Раненый, находящийся...", # состояние здоровья участника

"years\_of\_driving\_experience": 11 # стаж вождения участника (только у водителей)

}

]

},

],

"dead\_count": 0, # кол-во погибших в ДТП

"participants": [], # участники без транспортных средств (описание, как у участников внутри транспортных средств)

"injured\_count": 2, # кол-во раненых в ДТП

"parent\_region": "Белгородская область", # регион

"road\_conditions": ["Сухое"], # состояние дорожного покрытия

"participants\_count": 3, # кол-во участников ДТП

"participant\_categories": ["Все участники", "Дети"] # категории участников

}

В рамках Модуля 1 необходимо подготовить набор данных и произвести его предобработку для дальнейшего исследования и построения моделей обучения.

**ЗАДАНИЕ**

|  |
| --- |
| 1.1 Парсинг данных |
| На основании файлов .geojson размещенных на странице <https://dtp-stat.ru/opendata/>, содержащего данные по дорожно-транспортным происшествиям в каждом регионе, необходимо построить исходный набор данных. Набор данных должен быть преобразован в единый файл формата .csv. Можно дополнить набор какими-либо другими данными, если они могут быть полезны для дальнейшего исследования. |

|  |
| --- |
| 1.2 Предобработка данных и выделение значимых атрибутов |
| Предобработанный набор данных должен содержать только уникальные случаи дорожно-транспортных происшествий и не содержать в своих атрибутах перечислений. Задача определения опасности дорожного участка или перекрёстка заключается в определении класса (кластера) – уровня опасности. Уровни опасности определяются произвольно таким образом, чтобы в группу наиболее опасных дорожных участков, требующих срочных решений, попали не более 10% от имеющихся в наборе данных по каждому региону. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на определение таких классов (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) и статистику распределения данных (плотность, наличие пустых значений). |

|  |
| --- |
| 1.4 Формирование дополнительных атрибутов |
| [В](https://gogov.ru/articles/covid-rt) представленных данных по некоторым адресам происходило несколько дорожно-транспортных происшествий. Для каждого такого адреса необходимо вычислить индекс, отражающий зависимость от количества происшествий, частоты их происхождения, тяжести последствий, доли по отношению ко всем происшествиям в регионе. Проанализируйте возможность определения наиболее опасных дорожных участков на основе вычисленного индекса. |

|  |
| --- |
| 1.5 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам Модуля 1, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip |

# Описание модуля 2:

**Исходные файлы:**

1. Data.zip (Предобработанные данные)

**Результаты работы:**

1. Data.zip (Результаты модуля)
2. Report.html + Report.ipynb (Отчет о проделанной работе)

**ВВЕДЕНИЕ**

Вы продолжаете работать с данными, подготовленными в предыдущем Модуле. Предобработанные данные предстоит классифицироватьи визуализировать зависимости данных. Какая-либо работа, обусловленная задачами предыдущего модуля, выполненная в ходе текущего, оцениваться не будет, поэтому проделывайте её только в случае необходимости.

**ЗАДАНИЕ**

|  |
| --- |
| 2.1 Разбиение набора данных |
| Разделите исходный набор данных на обучающую и тестирующую выборки оптимальным образом. Приведите обоснование разбиения |

|  |
| --- |
| 2.2 Визуализация зависимостей данных |
| Используя программные средства, визуализируйте зависимости атрибутов в наборе данных. Визуализация должна отражать влияние атрибутов на определение уровня опасности дорожной ситуации на конкретных участках дороги или перекрёстках. Произведите расчеты зависимостей по выбранным алгоритмам. Приведите интерпретацию полученным результатам. |

|  |
| --- |
| 2.3 Классификация исходных компетенций |
| Выберите модель классификации данных по уровням опасности дорожной ситуации на конкретных участках дороги или перекрёстках. Приведите обоснование выбора модели. |

|  |
| --- |
| 2.4 Обучение |
| Проведите обучение выбранной модели на обучающей выборке. Протестируйте работу обученной модели на тестовой выборке. Определите показатели точности работы выбранной модели, сравните с остальными рассматриваемыми моделями. |

|  |
| --- |
| 2.5 Feature Engineering |
| Путём преобразования набора данных, добейтесь более точной работы выбранной модели. Опишите приемы генерации новых данных и результаты, к которым они привели, рассматривая все ранее определенные показатели точности |

|  |
| --- |
| 2.6 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам Модуля 2, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip |