Вариант № 1

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. [owid-covid-data.csv](https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv) (Исходные данные)

**Результаты работы:**

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

**ВВЕДЕНИЕ**

Вам предстоит разработать систему диагностики эпидемиологической ситуации в странах мира, связанной с распространением коронавирусной инфекции 2019-nCoV. Данная система диагностики будет включать исследование имеющихся открытых статистических данных портала <https://github.com/owid>. Система будет определять для туристов, желающих посетить страну, уровень заболеваемости.

Данные по ежедневной статистике в странах мира с 31.12.2019 г. представлены в файле [owid-covid-data.csv](https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv).

В рамках всего задания вам потребуется предобработать данные, выполнить анализ данных и выявить ключевые зависимости, построить необходимые модели машинного обучения.

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

**ЗАДАНИЕ**

| 1.1 Парсинг данных |
| --- |
| На основании файла [owid-covid-data.csv](https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv) репозитория <https://github.com/owid>, содержащего статистические данные об эпидемиологической ситуации в различных странах, необходимо построить исходный набор данных. Набор данных должен быть загружен непосредственно из репозитория и включать все необходимые атрибуты. Можно дополнить набор какими-либо другими данными, если они могут быть полезны для дальнейшего исследования. |

| 1.2 Предобработка данных и выделение значимых атрибутов |
| --- |
| Задача диагностики эпидемиологической ситуации заключается в определении класса – уровня опасности для туриста. Уровень опасности определяется тремя уровнями: зеленый — безопасно, желтый — средний уровень опасности и красный — посещение опасно. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на определение таких классов, 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 |
| --- |
| Сформируйте отдельный атрибут, в которых будет содержаться анализ распространения вируса с помощью коэффициента распространения инфекции (Rt = число зарегистрированных заболеваний за последние 4 дня / число зарегистрированных заболеваний за предыдущие 4 дня). Пример вычисления коэффициента RT приведен на портале: <https://gogov.ru/articles/covid-rt>. Проанализируйте возможность определения изменения эпидемиологической ситуации, используя сформированный атрибут. |

| 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**

В этой сессии вы продолжаете работать с данными, подготовленными в предыдущей сессии. Предобработанные данные предстоит классифицироватьи визуализировать зависимости данных. Какая-либо работа, обусловленная задачами предыдущей сессии, выполненная в ходе текущей, 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 классов – уровня опасности посещения страны для туриста. Произведите расчеты зависимостей по выбранным алгоритмам. Приведите интерпретацию полученным результатам |

| 2.3 Классификация исходных компетенций |
| --- |
| Выберите модель классификации данных по уровням опасности посещения для туристов. 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 |