СЕССИЯ 1

Исходные файлы:

1) Data.xlsx (Исходные данные)
2) Сессия 1.pdf (Инструкция к 1 сессии)

Результаты работы:

1) Data.zip (Предобработанные данные)
2) Report.html + Report.ipynb (Отчет о проделанной работе)
3) Readme.txt (Дополнительные комментарии)

ВВЕДЕНИЕ

На этом чемпионате 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 и приложение-виджет для туристов, желающих отправиться отдыхать заграницу. Система будет предлагать туристам, желающим посетить страну уровень заболеваемости в виде анимированного «светофора» (пример на рисунке 1) и прогнозировать дальнейшее развитие эпидемиологической ситуации, также отображаемое визуально (пример на рисунке 1).

Данные по ежедневной статистике в странах мира с 31.12.2019 представлена в файле data.xlsx

В настоящее время на многих сайтах туроператоров есть инфорграфика, демонстрирующая пользователю, насколько безопасно ехать в ту или иную страну, однако зачастую она не интерактивная, обновляется достаточно редко и не отображает прогноз ситуации на ближайшее время, что является ключевым фактором для туристов при выборе направления.

В рамках всего конкурсного задания вам потребуется предобработать данные, выполнить анализ данных и выявить ключевые зависимости, построить необходимые модели машинного обучения, разработать интерактивный виджет, который можно встраивать на сайты для отражения рекомендаций на текущий момент.

Виджет будет представлен в виде светофора, позволит выбрать страну и будет отображать один из трёх цветов (зеленый – посещать страну безопасно, жёлтый – посещать страну можно, но не рекомендуется, красный – посещать страну небезопасно). В виджете для сайта также будет предусмотрена возможность выбрать предстоящую дату поездки на основе прогноза, реализуемого на основе разработанных моделей машинного обучения.

На этой сессии необходимо только подготовить набор данных и произвести его предобработку для дальнейшего исследования и построения моделей обучения.

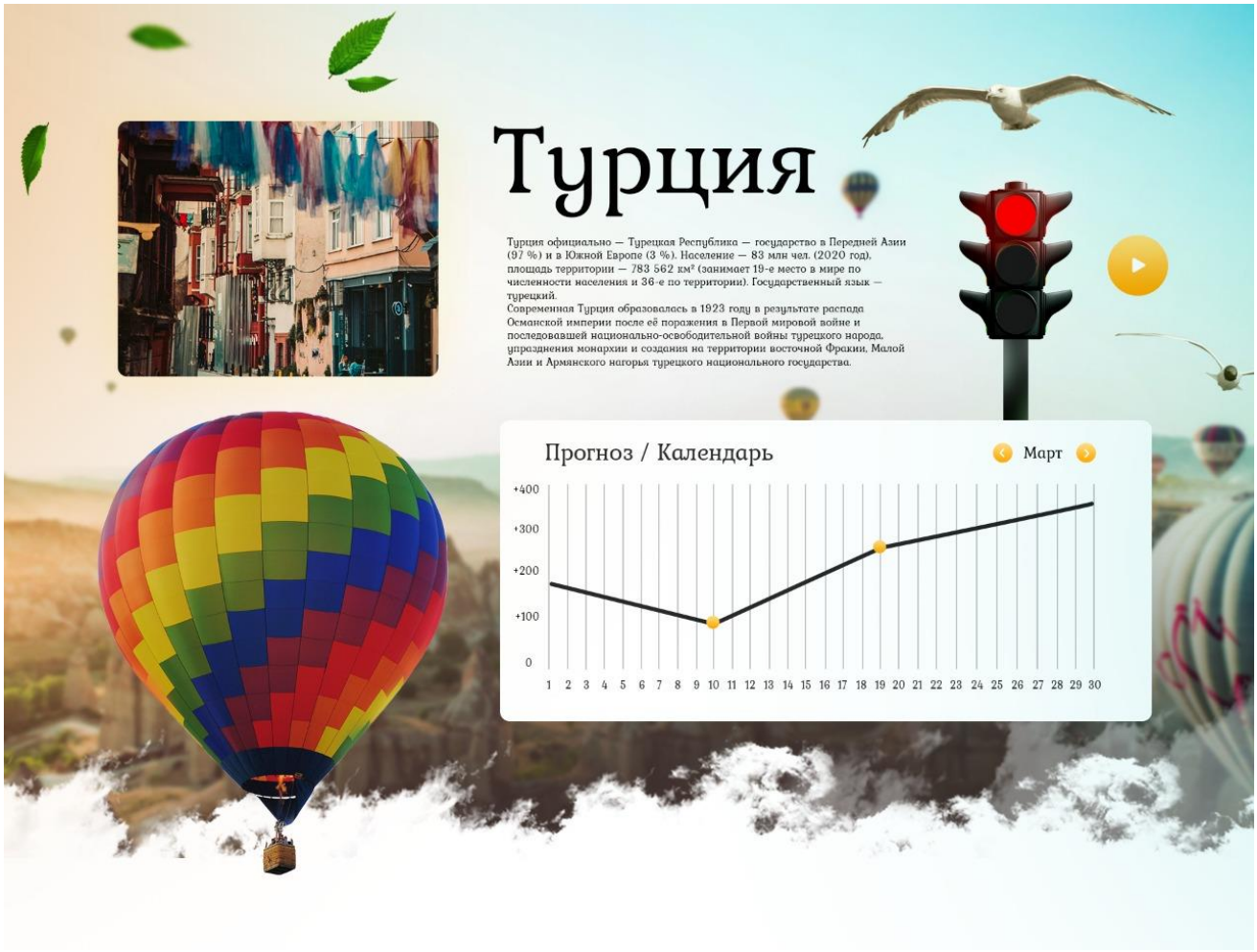

Рисунок 1. Пример визуальной реализации виджета

ЗАДАНИЕ

1.1 Парсинг данных
На основании файла 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 (кластера) – 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 (кластеров), и оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора документов.

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 Кластеризация набора данных

Выберите модель кластеризации данных. Задача кластеризации – определить уровни опасности для туристов и дать им наименования. В результате кластеризации может получиться несколько уровней опасности для одной страны в разные периоды времени. Приведите обоснование выбора модели.

1.6 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