Контрольное задание

Компетенция «Машинное обучение и большие данные»

**Общее описание**

В эти три для мы с вами познакомимся с задачами машинного обучения, с которыми могут столкнуться специалисты в реальной жизни: от загрузки и предобработки данных и до вопросов комментирования кода.

**Модуль 2. Загрузка, предобработка и подготовка данных. Анализ данных, машинное обучение. Кластерный анализ**

В данном модуле участники должны загрузить данные, выполнить предобработку, подготовить данные для последующего использования. Построить модели машинного обучения, а также провести кластерный анализ и подготовить отчет о проделанной работе.

Входные данные:

1. В архиве Stud\_IT\_input находятся следующие файлы: **matrix.png**, **Life\_expectancy\_dataset.csv**, **Aristotle.txt** и каталог **Energy\_consumption**.

Подробное описание задания:

1. Загрузите из папки **training\_datasets** файл **clock.mp4**.
   1. Для загруженного файла проведите предобработку. Снизьте объем исходных данных так, чтобы из начальных ~26.5\*109 (12\*60\*60\*640\*320\*3) точек датасета осталось менее 1% исходных точек, но при этом была сохранена полна ошение продемонстрируйте (кол-во исходных точек / кол-во точек после предобработки)
   2. Создайте разметку. Для каждого кадра назначьте три таргета: показания часов, минут и секунд.
   3. Разбейте датасет на обучающую и тестовую выборки
   4. Создайте 3 модели машинного обучения, способные по данным с кадра предсказать текущее время (часы, минуты, секунды). При создании модели не забудьте создать собственную метрику, которая оценивала бы отклонение предсказаний от фактических величин в секундах (отклонение на 1 час в 3600 раз хуже, чем отклонение на 1 секунду). Проведите оценку качества работы алгоритмов при помощи своей новой метрики. По результатам сделайте выводы.
2. Загрузите из папки **training\_datasets** файл **audio.wav**.
   1. Для загруженного файла постройте спектрограмму. Выведите ее на экран.
   2. Для загруженного файла выполните преобразование Фурье для перехода в амплитудно-частотный диапазон. Постройте амплитудно-частотную характеристику для всего набора данных, а также для первых 600000 точек и последних 600000 точек. Сделайте выводы .
   3. Для исходного аудио-сигнала выделите аггрегирующие характеристики, приведя его к частоте дискретизации 1с. Выделите не менее 5 аггрегирующих характеристик.
   4. Для аггрегированых характеристик выполните сглаживание не менее чем тремя способами. Результат визуализируйте
3. Загрузите из папки **training\_datasets** файл **higgs.csv.**

**В данном датасете приведены высокоуровневые переменные, полученные в результате обработки экспериментов по изучению бозонов Хиггса. Вам предлагается научиться определять, в каком из экспериментов был получен бозон, а в каком - нет**

* 1. Выберите метрику для оценки качества кластеризации данных. Выбор обоснуйте.
  2. Для представленного набора данных проведите кластерный анализ.
  3. Постройте график изменения метрики кластеризации в зависимости от числа кластеров (не менее 5 вариантов числа кластеров). На этом же полотне для каждого случая отобразите на графиках помимо значения метрики еще и время, затрачиваемое на кластеризацию с данным количеством кластеров.  
     Определите оптимальное число кластеров для разделения данных.
  4. Сделайте визуализацию отображения номера кластера на исходный массив данных на одном графике (цвет точек должен определяться номером кластера). Сделайте предположение о наличии бозона Хиггса в экспериментах того или иного кластера.
  5. Сделайте общий вывод о применимости методов кластеризации при определении наличия бозонов Хиггса в экспериментах.

1. Подготовить отчет по модулю (отчет должен быть в формате HTML, отчет должен быть структурирован, описывать ход работы, содержать комментарии).

Выходные данные по модулю:

1. Отчет в формате HTML.
2. Файл .ipynb, **полностью исполняемый и без выпадающих ошибок**.

Отчет должен иметь строго определенные название:

M2\_{lastname}\_{first name}

(например, M2\_Ivanov\_Ivan)