# Тестовое задание Astral AI для аналитика по данным

## Задача

Тестовое задание состоит в проведении Exploratory Data Analysis (EDA). Его результатом является чистый, работающий, воспроизводимый Jupyter Notebook, в котором считаются статистики, отрисовываются графики и делаются какие-то выводы (если удобно, можно ещё отдельно приготовить PDF, Google Doc или Markdown с отчётом).

В хорошем EDA по датасету обычно есть следующая информация:

* Количество данных - сколько всего данных, для скольки есть разметка, какой баланс классов, какую часть датасета ML-инженер сможет использовать в работе, а какую - нет
* Примеры исходных данных и разметки
* Аномалии в данных и разметке
* Любые другие интересные наблюдения - на что стоит обратить внимание человеку, который потом будет работать с этим датасетом.
* Бонус: любые гипотезы, которые можно было бы использовать в работе с этим датасетом.

Хорошо оформленный Jupyter Notebook - это:

* Импорты библиотек сверху в одной ячейке, используются заголовки разных уровней
* Для текста используются markdown cells
* Картинки хорошо различимы

## Данные

Недавно мы участвовали в одном хакатоне и работали со следующими данными.

В датасете представлены рентгенограммы органов грудной клетки (подвыборка из датасета ​NIH Chest X-rays​), содержащие как патологию, так и «норму», и наборы разметок. В наборах представлена разметка врача-эксперта (`Expert`) и трёх ML-алгоритмов (`sample\_1`, `sample\_2`, `sample\_3`).

Также врач-эксперт оценил корректность работы ML-алгоритмов по 5-ти бальной шкале (OpenPart.csv), где 5 – полностью согласен с разметкой интеллектуального алгоритма, 1 – полностью не согласен. Часть оценок была скрыта.

**Структура данных:**

* Корневая папка — Dataset
  + DX\_TEST\_RESULT\_FULL.csv – описание находок
  + Origin – входные изображения
  + Expert — разметка врача–эксперта (Ground Truth)
  + sample\_1 — первая разметка для сравнения
  + sample\_2 — вторая разметка для сравнения
  + sample\_3 — третья разметка для сравнения
  + OpenPart.csv – оценка корректности работы интеллектуальных алгоритмов

**Разметка предоставляется в двух форматах для простоты использования:**

* Описание находок в виде набора прямоугольников и эллипсов. Описание для прямоугольников и эллипсов (DX\_TEST\_RESULT\_FULL.csv):
  + Тип объекта Shape, бывает Rectangle — прямоугольник и Circle — эллипс.
  + Эллипс имеет ориентацию параллельно осям.
  + Координаты центра x, y: xcenter, ycenter.
  + Половина ширины и половина высоты области w/2, h/2: rhorizontal, rvertical.
* Сегментационная карта находок (`Expert`, `sample\_1`, `sample\_2`, `sample\_3`). Отрисованные значения файла DX\_TEST\_RESULT\_FULL.csv. Файлы в формате png. При этом если значение пикселя 0 — область не аномальная, если значение 255 — аномальная.