

Разработка системы анализа медицинских изображений для эпидемиологического мониторинга COVID-19

Мартынов Артем
Студент группы М24-525

1. Архитектура решения

Цель проекта:

Разработать аналитическую систему для эпидемиологического мониторинга COVID-19 на основе метаданных рентгеновских снимков, используя стек PySpark

Датасет:

COVID-19 Chest X-Ray Dataset (метаданные: 'patientid', 'age', 'sex', 'finding', 'view', 'date')

Архитектура :

- Схема пайплайна:
 1. Загрузка и предобработка данных
 - Обработка пропусков, аномальных значений
 - Удаление дубликатов
 2. SQL-аналитика
 3. Создание фильтров (UDF), категоризация
 4. Визуализация (Matplotlib, Seaborn)
- Инструменты: Python, PySpark, Spark SQL
- Оптимизация: партиционирование Parquet

2. Ключевые статистики

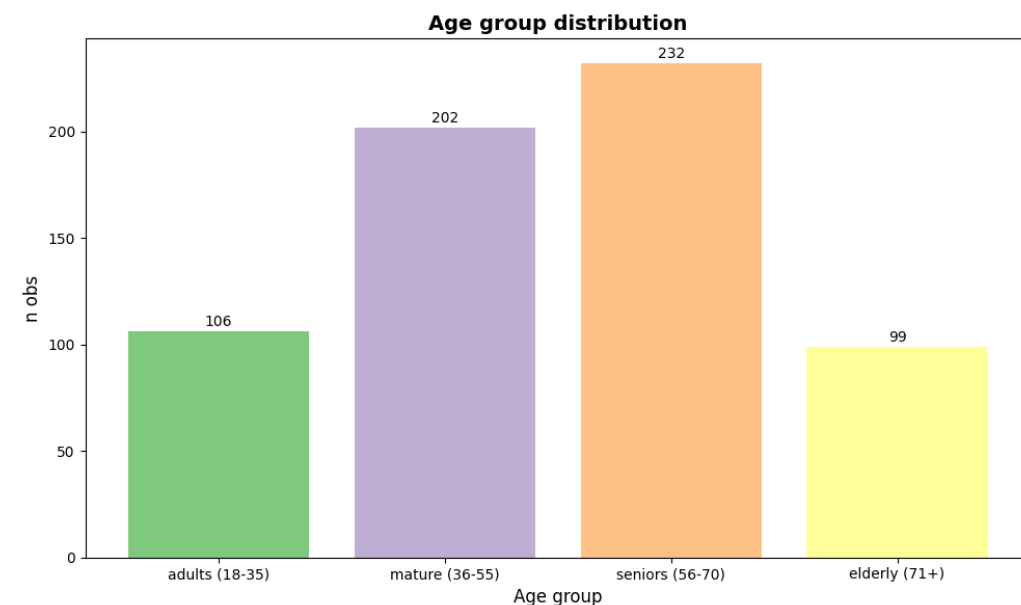
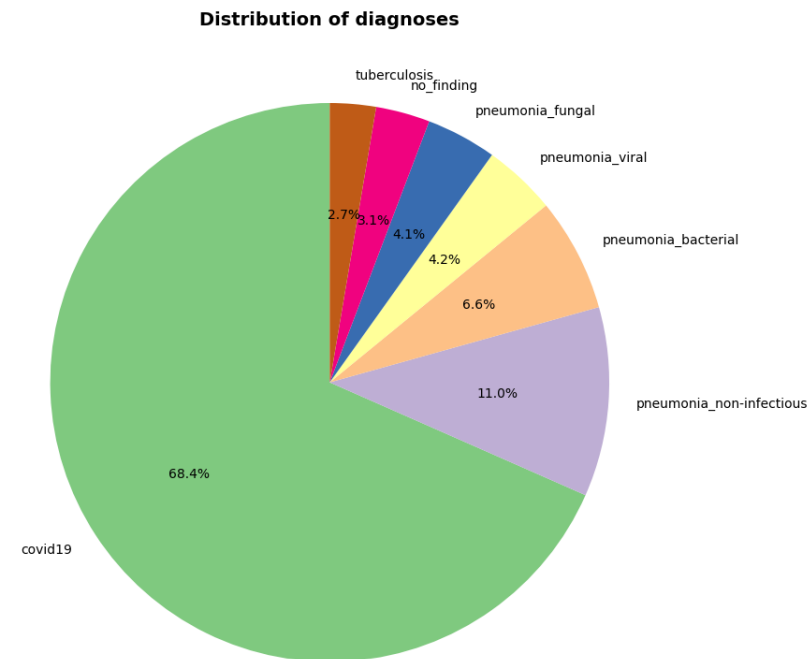
- 950 записей (по количеству снимков, пациенты повторяются)
- Пропуски до обработки: age (24.95%), sex (8.42%), date (30.42%)
 - Удалено 84 записи (866 осталось)
- Медианный возраст: 54
 - min: 18, max: 94 (после обработки)
- Данные за 2003-2020. Распределение кол-ва наблюдений по годам:
 - 2015 (21.36%)
 - 2020 (68.36%)
 - Остальные (10.28%)
- Дубликаты: удалено 227 записей (639 осталось)
- Распределение по полу: M – 429 (67.14%), F – 210 (32.86%)
 - При этом для COVID-19: M – 308 (70.48%), F – 129 (29.52%)
- Распределение диагнозов: COVID-19 (68.4%), прочие пневмонии (31.6%)
- Распределение взятия снимков по месяцам: январь (76.06%), март (18.31%), остальные (4.54%)
- Пик взятия снимков: 2020 год (69.48%)
- Основные проекции снимков: PA, AP, AP Supine

3. Визуализации и выводы

Согласно проведенному анализу, пик заболеваемости легочными болезнями пришелся на 2020 год и связан с COVID-19

Наибольшая частотность заболевания COVID-19 наблюдается у возрастной группы 56-70 лет. Обычно эта группа не является самой частотной в демографической пирамиде, поэтому можно с достаточно высокой вероятностью сказать, что у более возрастных групп легочные заболевания наблюдаются чаще. Группа 71+ может быть слабо представлена в этой выборке ввиду своей численности.

Тем не менее, для более строгих выводов необходимо знать структуру демографических пирамид стран(-ы), участвовавших в выборке



3. Визуализации и выводы

Больше всего снимков делается в проекции PA, вне зависимости от болезни. Но для COVID-19 их было сделано больше всего

Для COVID-19 наиболее характерны снимки в проекции PA, AP и AP Supine, каждый из которых составляет 25-34% от общего числа снимков для данного заболевания

