Анализ Данных о Рентгенографии Грудной Клетки при COVID-19 с Использованием PySpark

Анализ Данных о Рентгенографии Грудной Клетки при COVID-19 с Использованием PySpark

Цель проекта

Разработка аналитической системы для эпидемиологического мониторинга респираторных заболеваний на основе метаданных рентгеновских снимков.

Инструменты и данные

- PySpark
- Pandas
- Matplotlib
- Seaborn

Репозиторий:

ieee8023/covid-chestxray-dataset

Ключевые этапы

- настройка среды
- очистка данных
- SQL-анализ
- визуализация
- экспорт

Архитектура Аналитической Системы

Источник данных: COVID-19 Chest X-Ray Dataset (метаданные снимков, включая patientid, age, sex, finding, view, date, filename; 950 записей из репозитория ieee8023).

- **Обработка и анализ**: Выполняются с использованием PySpark для распределенной обработки больших данных, интеграции с SQL и визуализациями.
- **Основные этапы системы**:
  - Загрузка и изучение схемы данных (CSV с 30 колонками, первичный осмотр и схема).
  - Очистка и стандартизация признаков (возраст: заполнение медианой, фильтрация 0-120 лет; пол: унификация к "male/female/unknown"; диагноз: стандартизация к "COVID-19/Pneumonia/Normal/Other"; удаление дубликатов).
  - SQL-аналитика и агрегации (запросы по диагнозам, гендеру, возрастам, времени, проекциям).
  - Визуализация результатов (круговые, столбчатые, линейные графики, тепловые карты).
  - Сохранение очищенного набора в формате Parquet для быстрого доступа и ML.
- **Цель**: Получение объективных статистических показателей для эпидемиологического мониторинга респираторных заболеваний, выявление паттернов и поддержки ML-классификации.

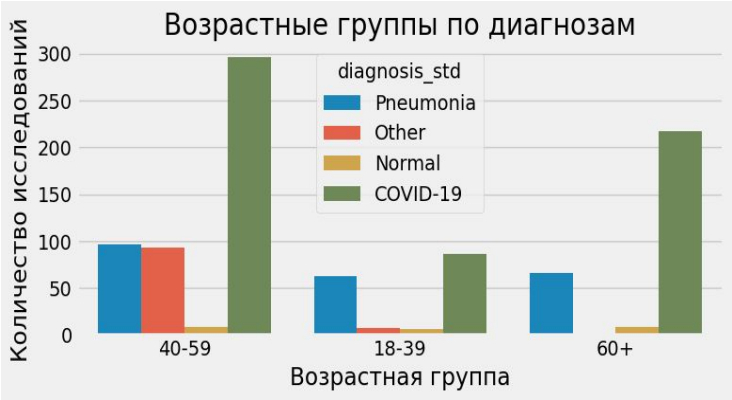
Основные Результаты Анализа

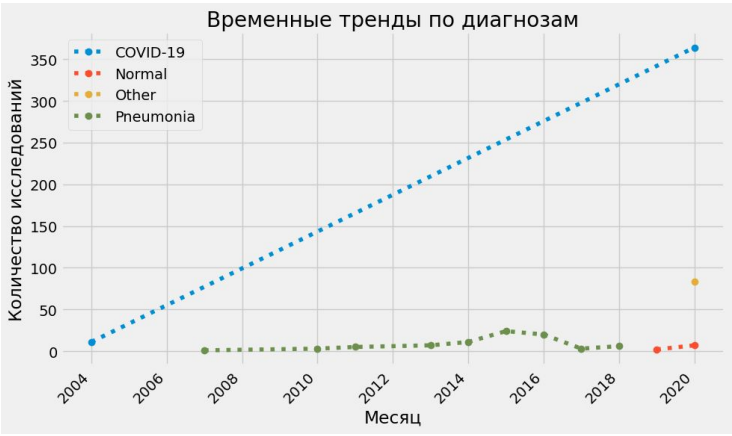
Всего после очистки: Около 950 снимков (после удаления дубликатов и фильтрации).

- Распределение диагнозов:
  - COVID-19: ≈ 63% (600 случаев).
  - Pneumonia: ≈ 24% (226 случаев).
  - Other: ≈ 11% (102 случая).
  - Normal: ≈ 2% (22 случая).
- Средний возраст пациентов: ≈ 54 года, диапазон от 18 до 94 лет.
- Возрастные группы: Преимущественно 40–59 (492 случая) и 60+ (174 случая); группы: 18-29 (66), 30-49 (218).
- Пик исследований: 2020 год (364 COVID-19), связанный с пандемией; редкие случаи до 2020 (пневмония 2004-2019).
- Демография: Преобладание мужчин с COVID-19 (351); старшие возраста в тяжелых диагнозах.

Визуальные Зависимости

- Распределение диагнозов по возрастным группам (столбчатый график: пики в 40-59 и 60+ для COVID-19 и Pneumonia).
- Временные тренды исследований (линейный график: резкий рост COVID-19 в 2020).
- Тепловая карта «диагноз × проекция снимка» (высокие значения для PA-COVID-19 (204), AP SUPINE-COVID-19 (136)).





Итоговые Выводы

- COVID-19 чаще фиксируется в проекциях PA и AP SUPINE, что характерно для серьёзных состояний, когда пациент не может стоять.
- Основные пациенты — лица старше 40 лет, что подтверждает статистику по рискам заболеваний дыхательной системы.
- Сформированный очищенный набор данных сохранён в Parquet, готов к использованию в ML-моделях и дальнейшей исследовательской работе.