План работ по проекту

(Предварительная версия)

Этап 1. Поиск, сбор и первичный анализ данных

1. Анализ основных датасетов на предмет пригодности для решения поставленной задачи. Анализ структуры данных.

2. Дополнительная задача. Поиск и анализ альтернативных и дополнительных датасетов. Анализ структуры данных.

   Потенциально полезные данные для дополнительных датасетов:
   - погода
   - промышленные предприятия
   - транспорт
   - сельское хозяйство
   - урбанизация
   - география

3. Исследовательский анализ данных. Оценка доли пропусков, неполных и аномальных данных. Агрегация и визуализация данных по загрязнителям, странам, годам и др.

4. Предобработка и очистка данных. Выбор наиболее оптимального способа корректировки проблемных данных. Исключение проблемных данных из датасета (в случае необходимости).

5. Формирование и сохранение второго (промежуточного) датасета для дальнейшей работы. Выбор страны (или нескольких стран) с наибольшим количеством и наиболее качественными данными. Выведение выбранных данных в отдельный (третий) датасет.

6. Статистический анализ данных. Построение простой предсказательной модели (на основе формул, экспертного мнения), оценка метрик. Исключение выбросов из датасета (в случае необходимости).

7. Разделение данных полученной выборки случайным образом на две части в соотношении 70:30, 80:20, 90:10 (в зависимости от размера полученного датасета) на обучающую выборку, и выборку для тестирования и валидации моделей (датасеты №4 и №5, соответственно).

Этап 2. Решение ML задач

1. Выбор наиболее оптимальных ML-моделей исходя из полученного размера и качества итоговой выборки.

2. Разработка и апробация нескольких ML-моделей для предсказания индекса качества воздуха.

3. Сравнительная оценка предсказывающей способности разных ML-моделей. Выбор наилучшей модели.

4. Использование ML-методов для поиска и удаления аномалий.
5. Использование ML-методов для кластеризации стран по составу воздуха.
6. Дополнительная задача. Оценка влияния размера обучающей выборки на предсказывающую способность модели. Обучение модели на новой выборке (в случае необходимости).

Этап 3. Решение DL задач

1. Использование DL-моделей для получения прогнозов.
2. Адаптивная селекция и/или композиция моделей.

Этап 4. Разработка финального отчета