УДК 504.054

ОСОБЕННОСТИ ФОРМИРОВАНИЯ ДАТАСЕТА ДЛЯ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ ЗАГРЯЗНЕНИЯ АТМОСФЕРЫ

Баталов А.И. (ГБОУ лицей №408), **Бойцова А.Д.** (Университет ИТМО)
Научный руководитель – ассистент ФЭТ, Фахртдинова С.З.

Введение. В современном обществе проблема загрязнения атмосферы является одной из наиболее актуальных и значимых. Это подтверждают такие международные мероприятия как Парижская конвенция Организации Объединенный наций (ООН) 2015 года, утверждение 17 целей устойчивого развития ООН, отечественный Национальный проект «Экология», включающий в себя Федеральный проект «Чистый воздух». Также наблюдается расширение промышленного производства, сопряженного с ростом негативного воздействия на окружающую среду. Перспективным направлением решения проблем в области охраны окружающей среды является использование искусственных нейронных сетей (ИНС). Они часто применяются в физических исследованиях, таких как распознавание речи и изображений, а также в химических и биологических исследованиях. В этом контексте разработка системы предсказания изменений показателей, характеризующих загрязнение атмосферы, представляет собой важное направление научно-исследовательской деятельности. Однако на этапе создания датасета для моделей машинного обучения возникают проблемы в сборе, валидации, верификации и оформления данных. Данная работа призвана осветить аспекты формирования датасета для модели машинного обучения в рамках работы над системой предиктивного анализа загрязнения атмосферного воздуха.

Основная часть. Для определения загрязнения атмосферного воздуха проводится расчет трех взаимосвязанных показателей состояния атмосферного воздуха, отражающих степень как максимального кратковременного, так, и длительного загрязнения атмосферного воздуха в городе: стандартный индекс (СИ), наибольшая повторяемость (НП), индекс загрязнения атмосферного воздуха (ИЗА), представляющий собой сумму среднегодовых значений концентраций пяти загрязняющих веществ, которые вносят наибольший вклад в загрязнение атмосферного воздуха города [1]. Сама же система требует сбора большого количества данных для обучения модели.

Изначально сбор данных производится из различных источников, таких как официальная статистика Росстата, мировая статистика WorldBankData и открытые отчеты локальных предприятий, чтобы получить достаточный объем информации, необходимой для обучения нейронной сети. После сбора данных проводится их обработка для подготовки к использованию в нейронной сети. Это включает объединение данных в единую таблицу и сохранение их в формате .csv. Для числовых признаков используются те же значения, преобразованные в единый формат, но для категориальных признаков проводится преобразование в числовые значения, чтобы нейронная сеть могла работать с ними. Также производится удаление данных, которые могут негативно повлиять на точность предсказаний, например, выбросы или отсутствующие значения.

На этапе формирования датасета возникают такие проблемы как валидация и верификация данных. Верификация проводится всегда и выполняется методом сличения характеристик данных с заданными требованиями. Результатом верификации выступает вывод о соответствии информации заданным требованиям. Валидация же выполняется методом анализа заданных условий применения и оценки соответствия характеристик данных этим требованиям. Результат валидации – оценка возможности применения сформированного датасета для конкретных задач [2]. В области охраны окружающей среды проблема верификации и валидации данных основана на нерегулярности сбора и предоставлении информации, а также постоянно меняющихся значений предельно допустимых концентраций (ПДК) загрязняющих веществ, обоснованных изменениями в документах контролирующих

организаций и органов. Поэтому создание датасета учитывает множество факторов для получения достоверных данных.

Выводы. В работе были рассмотрены аспекты сбора, обработки, валидации и верификации данных для ИНС. Также были выявлены проблемы формирования датасета, решение которых заключается как в сравнении различных источников, так и в экспертной оценке данных. Результатом работы стал сформированный датасет для модели машинного обучения в рамках проекта системы предиктивного анализа загрязнения атмосферного воздуха.

Список используемых источников:

1. Об утверждении методики определения высокого и очень высокого загрязнения атмосферного воздуха [Текст]: приказ Минприроды России от 17 февраля 2022 г. №106.

2. Калгина Е.А., Полякова Л.В. Отличие валидации от верификации // Е.А. Калгина // Успехи в химии и химической технологии. – 2018. – №8 (204). [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/otlichie-validatsii-ot-verifikatsii (дата обращения: 12.01.2024).

Баталов А.И. (автор) Подпись

Фахртдинова С.З. (научный руководитель) Подпись