ПРИМЕНЕНИЕ ИНСТРУМЕНТАРИЯ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИНТЕРПОЛЯЦИИ ЗАГРЯЗНЕНИЯ ВОЗДУХА ГОРОДА ИРКУТСКА.

APPLICATION OF MACHINE LEARNING TOOLS FOR AIR POLLUTION INTERPOLATION IN THE IRKUTSK.

Адаменко Алексей Сергеевич
Кедрин Виктор Сергеевич

Ключевые слова: интерполяция; машинное обучение; нейронная сеть; анализ

Цель работы состоит в интерполяции данных, содержащих информацию о загрязнении воздуха города Иркутска. В файлах формата .csv находятся записи наблюдений.

Замеры совершались примерно каждые 20-30 минут. Оборудование отслеживало содержание в воздухе вредных веществ, температуру воздуха, а также дату и время. В рамках работы изучался вопрос зависимости наличия вредных веществ в воздухе в зависимости от даты и температуры.

Актуальность работы: современный инструментарий позволяет анализировать и интерпретировать большие объёмы данных. На основе этого возможно проектировать сложные модели машинного обучения. Полученные модели позволяют решать экологические вопросы и прогнозировать информацию, что не представлялось возможным с помощью классических методов.

На первом этапе производилась предобработка данных с помощью библиотеки Pandas [2, с. 3]. Данный этап позволил сформировать из исходных данных репрезентативную выборку информативных признаков, а также исключил некорректные данные, которые были получены из-за неисправности оборудования либо ошибок наблюдения.

Для получения статистической и визуальной оценки данных использовались функция describe из пакета Pandas и библиотека matplotlib [1, с. 3].

Для более глубокого визуального анализа применялась функция heatmap из библиотеки seaborn [3, с. 3]. С помощью неё были выведены графики попарных взаимодействий признаков. При первом взгляде на графики появилось предположение о почти линейной зависимости некоторых признаков.

На следующем шаге исследуем, как каждый признак в отдельности распределён в зависимости от времени. Практически в каждом из

графиков наблюдается сезонность: с повышением выбросов веществ в воздух в моменты похолоданий.

Рассмотрим средние значения каждого признака в каждом месяце. Для этого используем функцию groupby библиотеки Pandas. И изобразим графики полученных данных.

На полученных графиках подтверждается сезонность наблюдений у нескольких признаков. Отчётливо видно, что выбросы 'CO, мг/м3', 'NO, мг/м3','NO2, мг/м3', 'SO2, мг/м3' начинают возрастать в октябре, достигают своего пика зимой "декабрь-январь", и затем спадают вплоть до июня, где достигают своего минимума.

Далее проводится работа с дынными о содержании в воздухе 'NO, мг/м3'. Построим регрессионную модель для данного признака в зависимости от температуры, воспользовавшись функцией LinearRegression библиотеки sklearn [4, с. 3]. Данная модель демонстрирует отрицательную зависимость от роста температуры.

Основываясь на полученной информации, попробуем предсказать уровень содержания в воздухе 'NO, мг/м3'.

Для решения поставленной задачи использован инструментарий из библиотеки TensoFflow [5, с. 3]. При первой попытке обученная модель была очень плоха. Сделано предположение, что данные имеют большой разброс, для исправления нормализовали данные с помощью функции normalize библиотеки sklearn. Затем произведена вторая попытка обучить модель, но уже с преобразованными данными.

В результате чего получена приемлемая модель, которая достаточно точно предсказывает уровень содержания в воздухе 'NO, мг/м3' основываясь на месяце и температуре воздуха. Среднеквадратичная ошибка составляет примерно 0.19.

Литература

1. Matplotlib: [сайт]. — URL: https://matplotlib.org (дата обращения: 15.04.2022)
2. Pandas: [сайт]. — URL: https://pandas.pydata.org (дата обращения: 15.04.2022).
3. seaborn: [сайт]. — URL: https://seaborn.pydata.org (дата обращения: 15.04.2022).
4. sklearn: [сайт]. — URL: https://scikit-learn.org/stable (дата обращения: 16.04.2022)

5. TensorFlow: [сайт]. — URL: https://www.tensorflow.org/api_docs/python/tf (дата обращения: 16.04.2022)