Big Data Engineering and Analytics

Leo Steel

Word count: 1982

Task 1: Data analytics

Introduction

Background and problem definition

I aim to develop a system that predicts the air quality index for different regions. Air pollution is a significant issue that impacts both human health and the environment. Still, the data surrounding this issue is vast and complex due to its continually changing variables. Every city collects pollution readings from sensors several times a day and considers factors such as weather, traffic, and geography. The biggest hurdle is that this data is spread, irregular and hard to analyse. My project aims to collect and clean a large dataset and then use it to forecast air quality in advance. This could help people prepare for when pollution levels are likely to be high and inform healthcare professionals of spikes in admissions.

Objectives

Objective 1: to pragmatically acquire and store data across a multi-year regional span using DAQI data from UK AIR DEFRA databases

Objective 2: to use parallel processing to handle multiple regions of the UK at the same time, significantly reducing the time spent processing data compared to standard methods

Objective 3: explore the data using visualisation tools like heatmaps to identify pollution levels in regions

Objectives 4: to create a regression model that can look at past air quality trends and patterns, providing a reliable forecast for future levels

Objective 5: Examining how sensitive data should be handled, with a focus on Privacy, ethics and data management

Evaluation of data characteristics

Volume: The dataset is sourced from the UK's Department for Environment, Food and Rural Affairs (DEFRA) and uses the daily air quality index to represent 16 regions across the UK from 2022 to 2025, totalling almost 1,400 days and 22,000 data points. Efficient and careful data engineering is necessary to process all these historical data points, ensuring they accurately reflect the model's performance.

Velocity: Although this dataset is historical, it provides near-real-time information, as DEFRA updates it rapidly. A functioning system must quickly ingest and sort daily updates to enable accurate air quality forecasting.

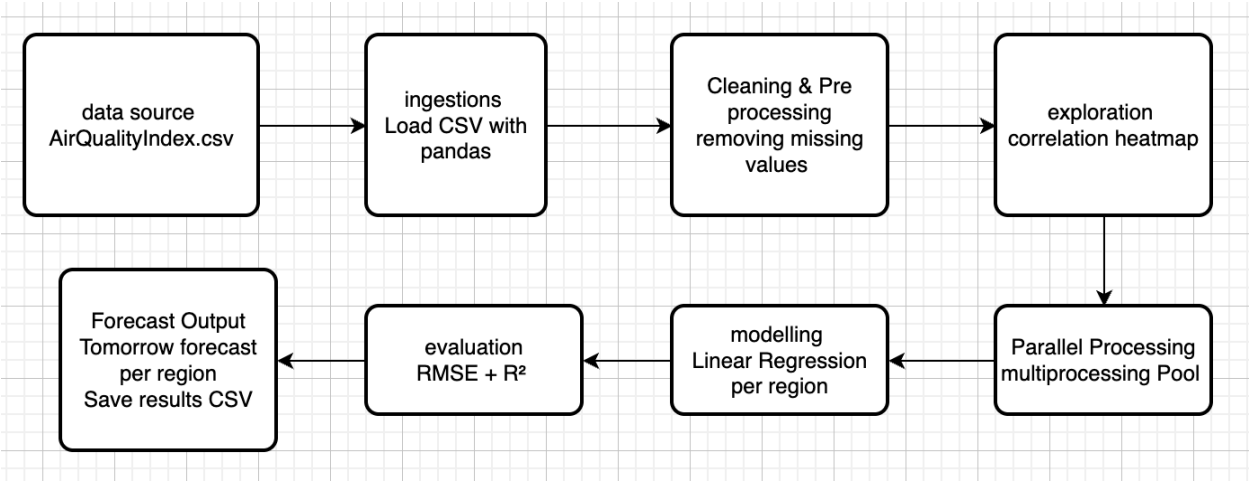
Variety: The dataset is fairly spread out, spanning from Scotland's Highlands, London and Wales, providing a contrast in air quality data ranges. To improve accuracy, it could be paired with external data sources, such as traffic data, to create a multi-source dataset.

Veracity: This dataset includes official government data from DEFRA sensors, ensuring high trustworthiness and consistency. However, preprocessing is necessary to clean the data by removing inconsistencies and stray commas.

Value: Predicting air quality enables at-risk individuals to plan outdoor activities with foresight, supports hospital preparedness, and ultimately helps save lives, thereby adding significant value to public health.

Proposed approach

Pipeline



The figure above shows the end-to-end data pipeline from data acquisition, pre-processing, data visualisation, pareelasation and machine learning

Parallel processing strategy

I understand that processing the 22,000 data points can be completed on laptops, but this strategy is not scalable. I have decided to switch to parallel processing to complete this task, which will allow it to handle multiple jobs simultaneously. For example, my dataset includes 16 regions as features. Without parallel processing, this would be done consecutively, all on the same CPU. However, with parallel processing, we can divide the task into 16 parts, allowing 4 regions to run on different cores at a time. This enables us to significantly speed up the process.

Conceptual correctness

Linear regression is the ideal choice for analysing the DAQI dataset because it contains continuous numerical variables. Regression models are specifically designed for forecasting continuous values rather than categorical ones. Additionally, linear regression enables me to assess prediction error using metrics such as RMSE and R^2, which provide valuable information about the reliability of the predictions.

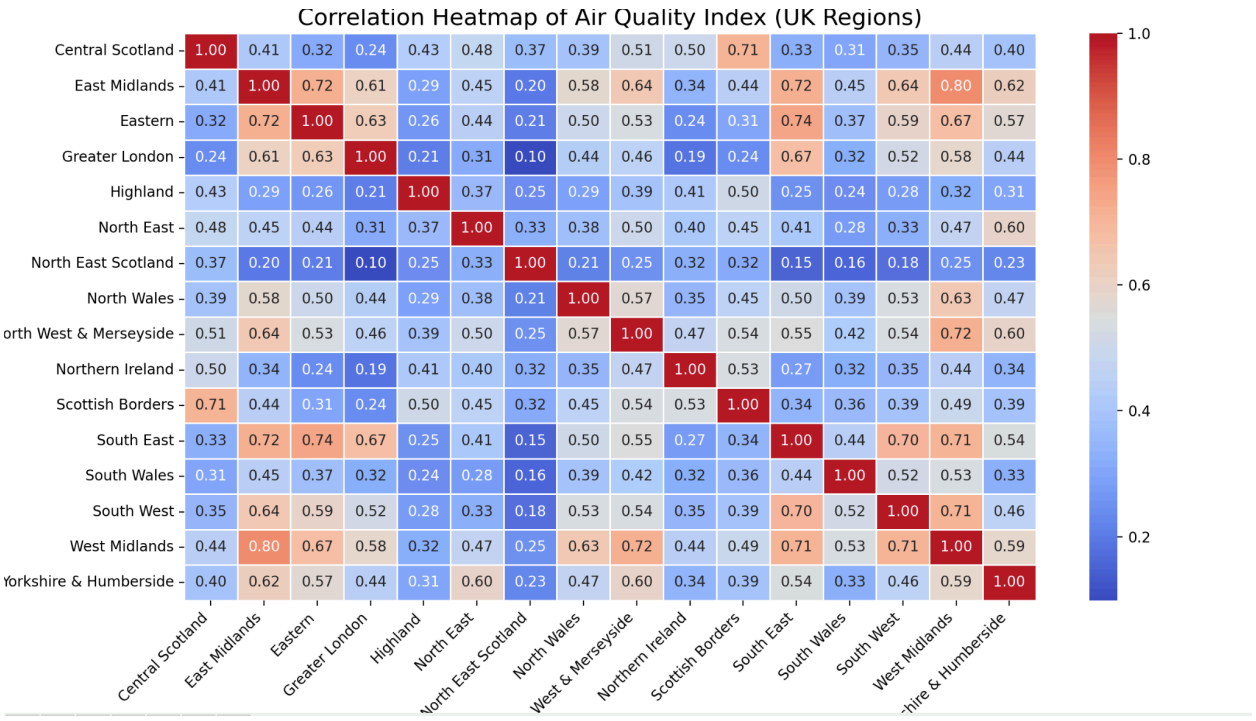
Application of tools and technology

Implmnetation

To address the DAQI dataset, I'll use tools from the multiprocessing library to help with parallel processing. Additionally, I will use visualisation tools to present the results clearly, using libraries such as pandas, seaborn, and matplotlib. For predictive modelling, I will include libraries from sklearn, including linear regression, mean_squared_error, and r2_score. Together, these tools will enable the collection and processing of data to derive significant insights that can be effectively used. This approach could, for example, provide warnings for individuals with respiratory illnesses or assist hospitals in preparing for an influx of patients requiring breathing treatments.

The insights gained can be directly applied to real-world situations, such as issuing warnings to the public with repository conditions or assisting healthcare professionals in better preparing for increased demand during adverse weather. This not only saves money but also saves lives.

Data exploration via data visualisation



The image above is a correlation heat map that reveals a strong positive correlation between geographically nearby regions, such as Greater London and the Southeast, suggesting that air pollution is a regional phenomenon rather than isolated to a particular place.

Knowledge extraction via machine learning

region	predicted result	RMSE	baseline	R^2
North East Scotland	2.2	0.83	0.78	-0.102
North Wales	2.58	0.72	0.61	-0.17
North West & Merseyside	2.73	0.78	0.72	-0.67
Northern Ireland	2.38	0.61	0.52	-0.076
Scottish Borders	2.57	0.69	0.59	-0.078
South East	2.86	0.89	0.8	-0.029
South Wales	2.79	0.77	0.74	-0.048
South West	2.81	0.77	0.67	-0.15
West Midlands	2.58	0.87	0.68	-0.218
Yorkshire & Humberside	2.69	0.95	0.82	-0.135

The figure above presents the key data from my linear regression model, which predicts daily air quality indices for 16 regions across the UK. The model was trained using time-based data, with 80% allocated for training and 20% reserved for testing.

To evaluate the model's performance, I measured the Root Mean Square Error (RMSE) and R^2, comparing them to a baseline in which tomorrow's results are assumed to be equal to today's. The result and evaluation data were generated independently for each region using parallel processing.

Evaluation

Appraisal of project objectives

objective number	objectives	☑ acheived	explanation
1	acquiring DAQI data from UK AIR DEFRA databases	☑	Data was successfully loaded from CSV and cleaned by dropping the top row and any stray commas
2	parallel processing to handle multiple regions of the UK	☑	This was accomplished using Python's multiprocessing library handling 16 different regions, excuted in batchs of 4 in parrel
3	data visualisations	☑	A correlation heatmap was produced by dropping the date column and skiiping stray comma values
4	create a regression model	☑	A linear regression model was implemented, trained on a time-based structtree with RMSE and R^2 to measure the predicition accuracy
5	Examining how sensitive data should be handled, with a focus on Privacy, ethics and data management	☑	Sensitive data were examined and reviewed in two papers that focused on ethical considerations and the responsible use of the data.

Overall, the core project was met. We focused particularly on several key sections: the data was successfully loaded, cleaned, and processed. Parallel processing was implemented correctly, with a setup designed for easy scalability, allowing regions from the DAQI to be processed in batches of four. The forecast model was executed, taking into account the time-based data structure split. However, using metrics like RMSE and R^2, I found that the model was underfitting, indicating it underperformed the baseline measurements.

The outcomes do not invalidate the system; rather, we should focus on the key takeaways that the pipeline and parallel processing strategy worked as intended. However, the complexity of the DAQI data should be highlighted, as there are vast amounts of variables to accommodate that a simple linear regression cannot effectively address. Variables such as day-to-day changes, weather, geography, and traffic make the data too noisy.

Lessons learned and recommendations.

The key takeaways from this project include the significance of using a time-based train/test split for time-series datasets, as random splitting can lead to untruthful results. By combining RMSE and R^2 with baseline performance.

Additionally, while linear regression is somewhat simple to implement and analyse, it struggles with noisy data, particularly when predicting small changes in the Daily Air Quality Index (DAQI). Given the number of variables evaluated for the National Air Quality Index (AQI), depending on a single feature produced insufficient results. However, parallel processing offers opportunities for scalability and faster implementation.

If I were to undertake this task again, I would focus on gathering datasets with multiple variables, such as weather conditions, to enhance the predictive model and improve forecast accuracy. The next logical step would be to implement the SARIMA model. This statistical time series model takes into account trends and seasonality, utilising factors such as past errors and repeating seasonal cycles to refine its results, which are more aligned with DAQI datasets

Task 2: Privacy, ethics and data management of common data stores

Introduction

Paper identification and description

I have chosen two separate papers from world-leading organisations. This was an intended move to put data security and ethics at the forefront, as both organisations are under intense regulatory and public scrutiny. The NHS paper focuses on protecting patient privacy in federated learning, while Apple's paper discusses using user data for emoji prediction in federated learning. This selection will hopefully provide two different views: one that uses data for consumer-related applications and the other that addresses clinical and public health objectives

Paper 1: Federated Learning for Emoji Prediction in a Mobile Keyboard

Approach and dataset

This paper describes how user-entered text can be utilised to improve and suggest emojis. The approach aims to use federated learning, a method that separates data into silos to enable mutual learning. In this case, Apple keeps the data on user devices, and only specific datasets are updated centrally.

Data stores and integration

Instead of storing personal raw data in a centralised Apple data warehouse, the model is brought to the data stored locally on the user's phone. This means that the model can operate while the user is engaged on their device, and whenever the model updates, the changes are sent to Apple's central servers. As a result, raw data is spread across many users' devices. This approach is a clever shift from conventional machine learning, which typically moves the data to

the model. Reducing support on large, centralised storage systems lowers infrastructure costs while still allowing continued model training.

Privacy and ethical considerations

From an ethical standpoint, storing private user data locally instead of in a single central location reduces the risk of breaches. This approach ensures that users' data remains close to them, increasing trust and providing reassurance when handling consumers' personal information.

Paper 2: Privacy preservation for federated learning in health care

Approach and dataset

This paper discusses the application of federated learning (FL) to train AI models across hospitals without sharing patients' personal data. The motivation for implementing FL arose from the need for healthcare systems to access large, noisy datasets while adhering to strict privacy laws governing patient data.

Data stores and integration

However, as the paper continues, it highlights that FL models may not be as private as previously thought. Although raw private data isn't shared, there remains a risk of data leaks when model updates are sent to a central aggregation point, which can be exploited to steal model parameters or infer sensitive details.

Privacy and ethical considerations

To address this problem, the authors have discussed using privacy-preserving techniques such as encryption, confidential computing, and differential privacy, which add noise to the model in a controlled way to reduce the likelihood that attackers can tap into sensitive information.

Critical appraisal

Both privacy and ethical considerations were addressed, providing a comprehensive view of how these papers aim to protect customer data while also utilising it for machine learning. However, the depth of the discussion differs between the two papers.

In Apple's paper on emoji prediction using federated learning, the focus is primarily on the benefits of keeping data on local devices. However, it does not provide sufficient detail on how it mitigates threats to the model updates being sent to the central aggregator hub. In contrast, the NHS article offers comprehensive insights into privacy risks, identifying potential attacks and outlining mitigation strategies such as encryption and differential privacy.

The two papers also vary in the personal data they use. Apple's approach uses user text data for emoji prediction, and consent for this could be embedded in user agreements and privacy policies. On the other hand, healthcare data is highly sensitive and needs strict accountability

and regulatory compliance. This may affect the model's accuracy due to security measures for model updates, such as differential privacy. The NHS focuses on patient confidentiality rather than predictive performance.

As a result, the NHS app offers a clearer, more transparent approach to privacy protection. However, Apple’s protection measures could be strengthened by documenting the steps taken to secure model updates and by reducing user data leakage, even if they follow a structural approach similar to that of the healthcare model.

References

Privacy preservation for federated learning in health care. (n.d.). *Patterns*, *5*(7). https://doi.org/10.1016/j.patter.2024.100974

Ramaswamy, S., Mathews, R., Rao, K., & Beaufays, F. (2019, June 11). *Federated learning for emoji prediction in a mobile keyboard*. arXiv.Org. https://arxiv.org/abs/1906.04329