Enhancing Accuracy in London's Air Quality Data Analysis: Addressing Bias through A Comprehensive Framework

Ejaz Hussain

Senior Data Scientist (UK Civil Services)

M.Sc., MBCS, FHEA, MCSE, PhD Researcher (DS)

Why Data Science is Crucial for Climate Change Crisis?

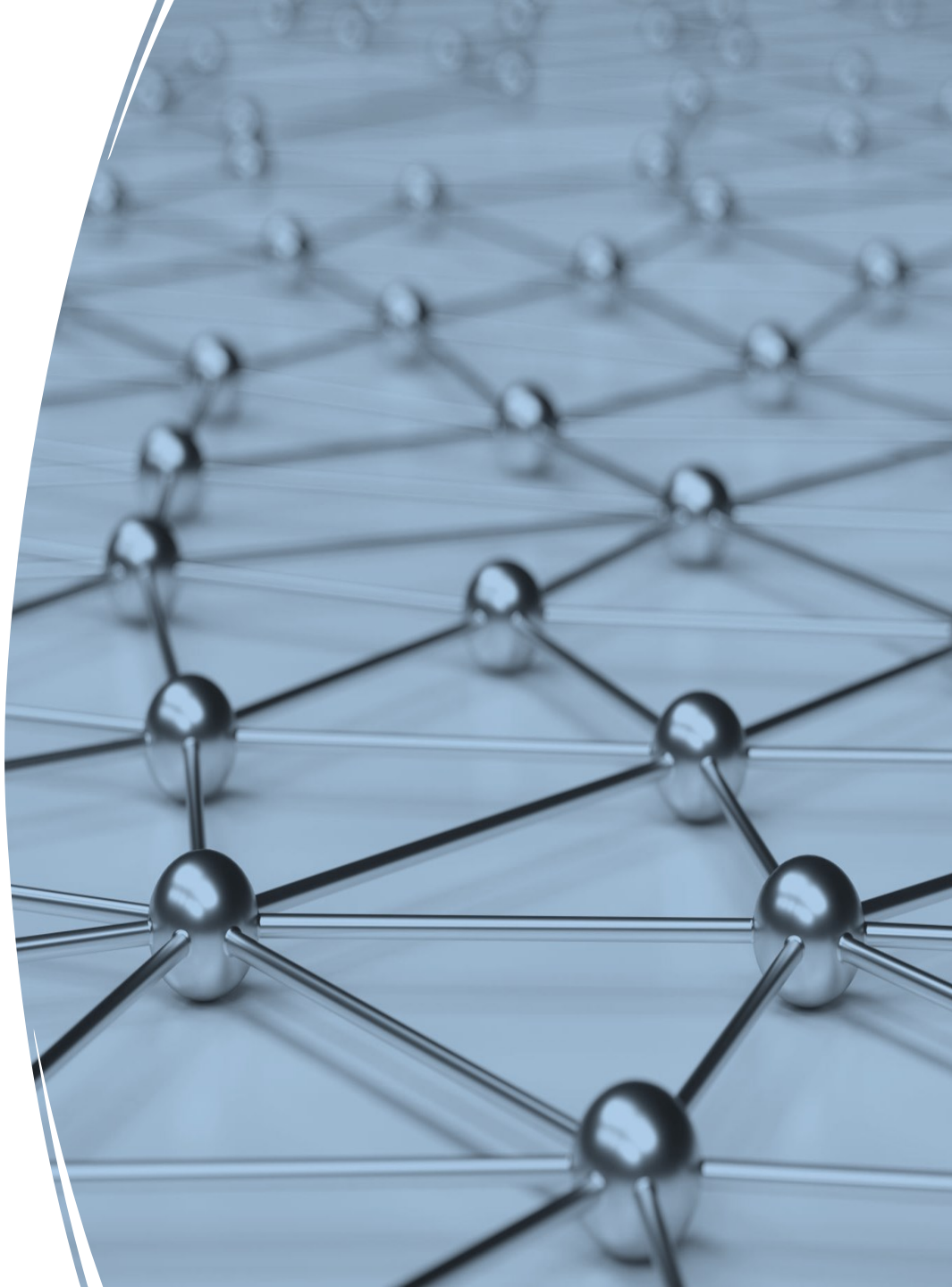
- Predictive Modelling And Forecasting
- Data-driven Policy Making
- Monitoring And Tracking Climate Changes
- Measuring Carbon Footprint
- Public Awareness And Education

"Data science and AI represent two of our most powerful assets in the fight against climate change. Now is the time to re-imagine the way we conduct climate science research and address the crisis head-on"

Dr. Alden Conner et al. The Alan Turing Institute [1]

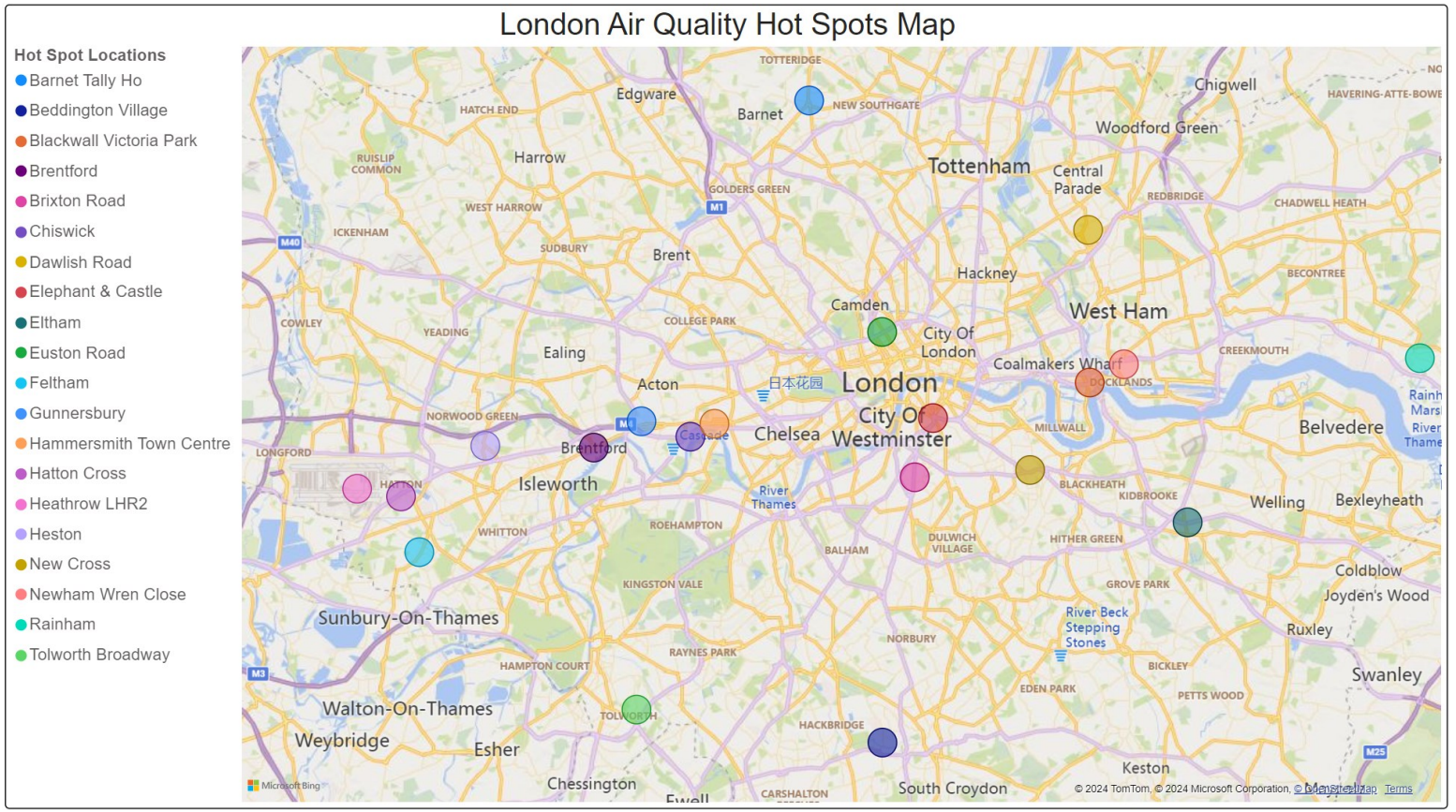
The Research Problem

- **Identify How and What to Statistically Measure?** For i.e. merging multiple air quality datasets with incorrect pollutant scales or replacing too many outliers for machine learning models.
- **Evaluate Existing Frameworks:** like, IBM AI Fairness 360 [2]
- **Bridge BIAS:** Does existing frameworks include data bias? What are those gaps?
- **A Scoring Ladder:** Apply a novel statistical approach to reduce BIAS in air quality outcomes
- **Re-Evaluate** the Difference



The Research Scope

London, United Kingdom

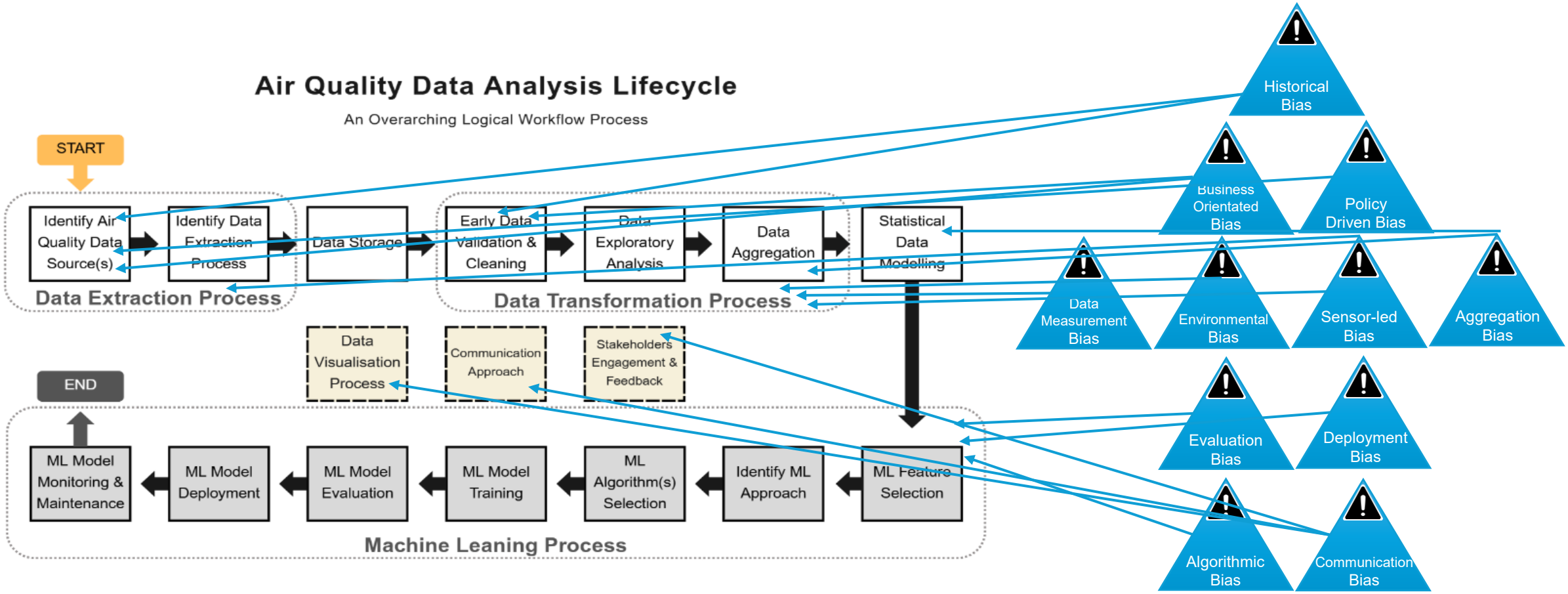


The Research Lab & Methodology

Air Quality Data Analysis 8 Key Stages

- Awareness on Air Quality Variables and Pollutant Scales – S1
- Data Collection Methods and Techniques – S2
- Data Cleaning and Preparation – S3
- Exploratory Data Analysis – S4
- Statistical Data Analysis – S5
- Time-Series Predictive ML Modelling – S6
- Data Validation – S7
- Deployment & Communication – S8

An example of a typical air quality monitoring station [3]



The Research Lab & Methodology

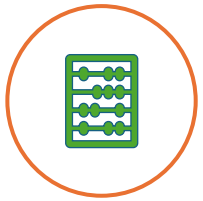
Air Quality Data Analysis Life-Cycle

A Scoring Ladder (Algorithm)

A Contribution to Research

A step-by-step statistical method to detect and score BIAS in Air Quality Data Analysis

Bias Check



Bias Scoring

A Scoring Ladder (Algorithm)
A Contribution to Research

A step-by-step statistical method to detect and score BIAS in Air Quality Data Analysis

Bias Index	Classification of Bias	Data Anaysis Stage	Known AQ Bias Risks	Checklist for AQ Bias and Scoring	Supported References	Scoring Weight
1	Historical Bias	S1 | S2	1) Under-representation of air quality pollutant 2) Unreasonable timeline selection for a Dataset 3) Unjust Air Quality Data Monitoring Sites (Hot Spot) Selections	1) To evaluate AQ pollutant(s) Representation Analysis in S1,S4 2) To determine sufficient timeline for data extraction in S2 3) To examine AQ Monitoring Sites in S1,S2	Check Overleaf Chapter 3 for Supported Citations	1
2	Business-Orientated Bias	S1 | S2	1) Business own Aims and Objectives for desired Outcomes 2) Data Selection Preferred Criteria 3) Monitoring Site Preference for Targeted Outcomes 4) Business Preferred Data Samples (Closed Datasets rather than Open Datasets) 5) Urban vs Rural Area's Reprsentations	1) Question and Reasoning on Business Context and Objectives 2) Fair Selction of Data Samples 3) To examine AQ Monitoring Sites in S1,S2		0.5
3	Policy-driven Bias	S1	1) Inappropraite use of Supplied Sampling Data 2) Negative Policy-driven Illusions & Opinions 3) Blind Trust	1) AQ Policy Awareness 2) Contribution of Local AQ Factors 3) AQ Broader-Context Policy Awareness		0.5
4	Environmental Bias	S1 | S4 | S5	1) Humidity Factor 2) Temperature Factor 3) Weather Conditions 4) Physical Obstacles 5) Unstable Power Supply	1) AQ Sensor(s) Sensitivity Analysis 2) Awareness on Physical Location and any Known Obstacles 3) Statistical Data Consistency Checks		1
5	Data Measurement Bias	S1 | S2 | S3 | S4 | S5	1) Unfair Feature(s) Selection 2) Lack of Awarness & Subject Matter Expertise 3) Air Pollutant(s) Incorrect Measuring UNITS 4) Preferred Selection Criteria on Data 5) Feature Rep. Not not meeting RWD Interests	1) Domain Expertise and Awarness 2) Fair Feature(s) Selection Process 3) Statisical Based Test for Feature(s) Correlational Sudy		1
6	Algorithmic Bias	S3 | S4 | S5	1) Unfair Feature(s) Selection 2) Biased Test and Evaluation Data sets 3) Flawed AQ Training Data set 4) Flawed AQ Selection Criteria 5) Prejudicated Assumptions 6) Preferred Decision Making Outcomes	1) Data and Model Transparency 2) Fair Feature(s) Selection Process 3) Algorithm Accountability		1
7	Aggregation Bias	S3 | S4 | S5	1) Unfair AQ Data Aggregation(s) 2) AQ Time Period Miscalculations 3) Flawed Relationships b/w Air Pollutant's (Variables) 4) Judgmental Assumptions 5) Inappropriate Data Practices 6) Outcomes Focused	1) Sensitivity Analysis for Correlational Patterns b/w Air Pollutant's 2) Fair Use of Aggregations in Data Analysis 3) Effective use of Data Visualisations (Relationships b/w Pollutant's, Variable Representation in EDA and then SDA Stages)		1
8	Sensor-led Bias	S2 | S3 | S4 | S5	1) Noisy Factors 2) Poor Sensor Sensitivity Strength 3) Poor Data Accuracy & Reliability 4) Data Interruptions 5) Data Corruption & Inconsistencies 6) Data Processing & Extraction Conflicts	1) AQ Sensor(s) Sensitivity Analysis 2) AQ Sensor Network(s) Awareness 3) Statistical based Data Consistency Checks		1
9	Evaluation Bias	S5 | S6 | S7	1) Imbalanced Test Data 2) Inappropriate use of Metrics 3) Domain Specific Negligence 4) Outdated Benchmarking	1) Environmental Specific Awareness & Expertise 2) Use of Balanced Test Dataset 3) Use of Appropriate Statistical Practice		1
10	Deployment Bias	S8	1) Infrastructure Incapabilities 2) Real-World Data Challenges 3) Development Environment Conflicts 4) End Users & Legacy Systems Challenges 5) Ethical & Legal Challenges	1) Real-World Data Awareness & Expertise during Pre and Post Deployment Phases 2) Stable Deployment Infrastructure 3) Awareness and Understanding on Legal and Social Implications		1
11	Communication or Interpretation Bias	S8	1) Lack of Context 2) Ambiguity in Presentation 3) Over-simplifications 4) Audience Understanding Level 5) Selective Reporting 6) Use of Inappropriate Metrics 7) Lack of Domain Expertise	1) Fair use of Metrics, Visuals & Communication Channels 2) Domain Expertise on all relevant Subjects 3) Tranparent Reporting w/o 'Pick and Choose' Criteria		1
						10

1.1.2 Step 2: Define the bias variables

The research has highlighted 11 biases and each bias is assigned with a weight:

$$B_i = \begin{cases} 1 & \text{if bias } i \text{ is present,} \\ 0 & \text{if bias } i \text{ is absent.} \end{cases}$$

The weights w_i for each bias are:

$$w_i = \begin{cases} 0.5 & \text{if bias } i \text{ is business-oriented or policy-driven,} \\ 1 & \text{otherwise.} \end{cases}$$

The total bias score S is calculated as:

$$S = \sum_{i=1}^{11} w_i \cdot B_i$$

where $S \in [0, 10]$.

1.1.3 Step 3: Expanded Bias Formula

The research has identified 11 types of biases B_i in air quality data analysis, where:

$$B_i = \begin{cases} 1 & \text{if bias } i \text{ is present,} \\ 0 & \text{if bias } i \text{ is absent.} \end{cases}$$

The weights w_i are defined as:

$$w_i = \begin{cases} 0.5 & \text{if bias } i \text{ is business-oriented or policy-driven,} \\ 1 & \text{otherwise.} \end{cases}$$

Thus, the total bias score S is calculated as:

$$S = \sum_{i=1}^{11} w_i \cdot B_i.$$

Expanding this for all 11 biases:

$$S = B_1 + B_2 + B_3 + B_4 + B_5 + B_6 + B_7 + B_8 + B_9 + 0.5B_{10} + 0.5B_{11}.$$

Since the maximum score is:

$$S_{\max} = 9 \cdot 1 + 2 \cdot 0.5 = 10,$$

the score $S \in [0, 10]$.

1.1.4 Step 4: An Example

For example, 6 out of the first 9 biases are present, the business-oriented bias (B_{10}) is present, and the policy-driven bias (B_{11}) is absent. In this case, formula calculation is:

$$S = 6 + 0.5 \cdot 1 + 0.5 \cdot 0 = 6.5.$$

Therefore, the total bias score is $S = 6.5$ out of 10.

References:

1. Conner, A., S. Hosking, J. Lloyd, A. Rao, G. Shaddick & M. Sharan, Tackling climate change with data science and AI . Mar. 2023.
2. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. and Nagar, S., 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), pp.4-1.
3. DEFRA, UK AIR - Air Information Resource. Department for Environment Food andRural Affairs.