Coursework Task Sheet 3

Data Analytics Pipeline: Application and Analysis

CMP5366: Data Management and Machine Learning Operations

Birmingham City University

Kacper Popis – 23161791 – [kacper.popis@mail.bcu.ac.uk](mailto:kacper.popis@mail.bcu.ac.uk)

[Step 1 2](#_Toc792272860)

[Step 2 5](#_Toc208306754)

[Step 3 6](#_Toc1847735709)

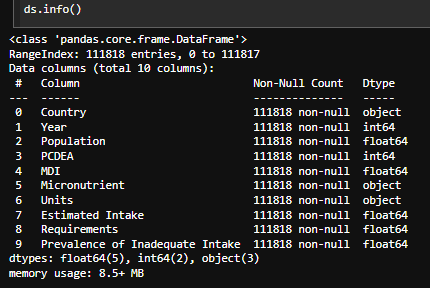
[References 7](#_Toc1855890633)

# Step 1

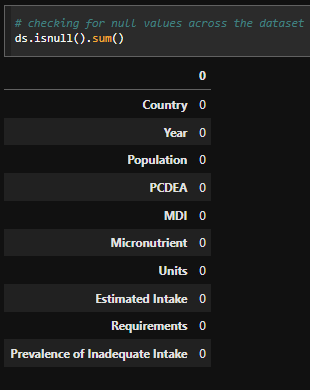
Majority of the data inside of the dataset was already numerical (as mentioned in Task Sheet 2), with some non-numerical data like countries, which could be classified as categorical data, but because the techniques that have been used to create the AI models are regression techniques, the data inside of the dataset will have to be strictly numerical.

Also, as mentioned in Task Sheet 2, there is a disparity between the amount of data for each country, as some countries hold 728 entries of data, while others contain only 128 entries of data. This leads to an imbalance of data across all countries which will affect how accurate the model will be on countries with less data entries, as not only does it limit the time frame the data was collected, but also limits the quantity of micronutrients measured during a specific time period (some countries might have collected all micronutrient data in a single year, when some countries may only have a fraction of it).

However, the dataset itself does not contain any null values, as it is shown below in *Figure 1*, when running the **.info()** function on the dataset, the non-null count value is the same across all the columns. This can be further checked using the **.isnull().sum()** function (shown in *Figure 2*) to count the amount of null values per column in the dataset.

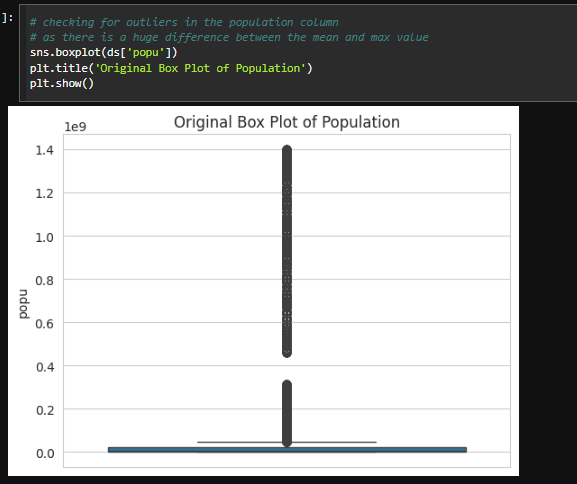


*Figure 1: Using the* ***.info()*** *function to check the dataset*



*Figure 2: Using the* ***.isnull().sum()*** *function to count all null values*

Outliers were also checked for in the dataset, and this was done on the column that showed some potential outliers – the population column. Using the box plot to display the potential outliers in the column , there is a large amount of outliers present, however this is most likely due to (as the mentioned before) disparity of data between the countries, as such these outliers were kept in the dataset otherwise some countries would end up with no data.



*Figure 3:**Box plot diagram of the population column*

As mentioned in Task Sheet 1, the original dataset downloaded from FigShare contains a collection of 5 CSV files, and some of the files are connected to one another, providing potential links within the dataset. However, to streamline the process (due to the massive amounts of data – as shown with one of the planned regression techniques failing to function) only the main CSV file (S4.csv) has been used to train and develop the regression models and stored inside of the database, but the files linked to the main csv file could have potentially been stored to help aid the training progress – but this can be looked at as an upgrade to the test data in the future.

# Step 2

Using the SKlearn Metric library, all of the models developed are constantly having their accuracy reassessed whenever they are built. On the airflow build of the pipeline, the models are not created with a predefined **random\_state** (as mentione in Task Sheet 2), this prevents the models from utilising the exact same data every single time they are built – this is done to allow for each model to learn on somewhat different data to compare the accuracy between them.

Some of the processes throughout the pipeline process could potentially be simplified, or at least use less processing power, as one of the planned regression models stopped functioning (I believe the reason was the amount of data for that specific regression type) - as such methods could be implemented to allow for that specific regression model to function while not being disadvantaged compared to the other regression models. Another potential change to the pipeline could be looking into algorithm complexity throughout the whole process to speed up processing, and prevent redundant code being executed to only elongate the process.

I believe the data utilised is not fully fair, as previously stated in the previous Task Sheets, the data per country is not even, some countries have more, while some have less. This creates a bias towards countries with more entries as the ones with less have a lesser chance to be chosen for training data. Additionally, when it comes to predicting using the lesser countries, the accuracy for that specific country would be lesser than the acuraccy of a country with 700+ entries. If the dataset would contain a better balance of data this model would be more beneficial.

# Step 3

Model drift is a term that describes the process of an AI model slowly degrading in accuracy / performance due to a number of reasons, some of which are changes in relationships of data, quantity / quality of data used to train the model and the disparity of input and output data **[1]**. This can greatly impact the deployed machine learning system as the systems function on the data that is being fed to them, and if there are changes in the data, like the data being fed being invalid or fake, then this can affect how well the AI model is able to process incoming requests from users.

Model drift goes hand in hand with concept drift **[2]**, as the change of data will affect the performance of the model, but it will also affect the relationships between different features of the dataset used for the model. As some features may have a greater impact that other features, feeding untrue data to the model will depreciate the impact of those features or potentially weaking the relationships between specific features – resulting in a different outcome compared to the desired one.

The best way to combat model drift would be to reassess the model regularly and utilise metrics **[3]** that evaluate performance of the model (as mentioned previously – SME, SAE, R2) and choosing the appropriate ones based on the model type that is being used. With these metrics, you are able to notice dips in accuracy compared to the usual model’s performance (by keeping a track of metrics).

# References

[1] - IBM (2024). Model drift. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/model-drift. [Accessed. 20th May 2025]

[2] - evidentlyai. (n.d.). What is concept drift in ML, and how to detect and address it. [online] Available at: https://www.evidentlyai.com/ml-in-production/concept-drift. [Accessed. 20th May 2025]

[3] - Polat, G. (2024). Model Drift: Best Practices to Improve ML Model Performance. [online] Encord.com. Available at: https://encord.com/blog/model-drift-best-practices/. [Accessed. 20th May 2025]

‌