Panu Khumchoo-Joonpijarn

23135551

AI and Machine Learning Project

Predictive modelling of Global Region based on Micronutrient Deficiency using Supervised Machine Learning Algorithms

Table of Contents

[1 Report Introduction 1](#_Toc184846945)

[1.1 Introduction 1](#_Toc184846946)

[1.2 Dataset Identification 2](#_Toc184846947)

[1.3 Supervised Learning Task Identification 3](#_Toc184846948)

[2 Exploratory Data Analysis Process and Results 3](#_Toc184846949)

[2.1 Question(s) Identification 3](#_Toc184846950)

[2.2 Exploratory Data Analysis Process and Results 5](#_Toc184846951)

[2.3 Exploratory Data Analysis Conclusions 9](#_Toc184846952)

[3 Experimental Design 10](#_Toc184846953)

[3.1 Identification of Chosen Supervised Learning Algorithm(s) 10](#_Toc184846954)

[3.2 Identification of Appropriate Evaluation Techniques 11](#_Toc184846955)

[3.3 Data Cleaning and Pre-Processing Transformations 11](#_Toc184846956)

[3.4 Limitations and Options 23](#_Toc184846957)

[4 Predictive Modelling / Model Development 24](#_Toc184846958)

[4.1 Splitting the Dataset 24](#_Toc184846959)

[4.2 The Predictive Modelling Process 26](#_Toc184846960)

[4.3 Evaluation Results on “Seen” Data 27](#_Toc184846961)

[4.3.1 Grid Search 27](#_Toc184846962)

[4.3.2 Predictions Using Over Sampling (SMOTE) 29](#_Toc184846963)

[4.3.3 Predictions Using Under Sampling (Near Miss) 32](#_Toc184846964)

[4.3.4 Deep Learning 35](#_Toc184846965)

[5 Evaluation and Further Modelling Improvements 43](#_Toc184846966)

[6 Conclusion 44](#_Toc184846967)

[6.1 Summary of Results 44](#_Toc184846968)

[6.2 Reflection on Individual Learning 44](#_Toc184846969)

[7 References 44](#_Toc184846970)

# 1 Report Introduction

## Introduction

As stated in journal (Beal et al., 2017), we characterise the global trends in dietary quality by estimating micronutrient density of the food supply. Undernourishment remains very high in about 11% of the global population, such hidden hunger is associated with pregnancy complications and child growth failure, increase susceptibility to disease, and impair cognitive development. On the other hand, excessive consumption is associated with increased incidence of obesity and risk of diabetes and heart disease.

The dataset being used in this study is from the journal detailing the "Global trends in dietary micronutrient supplies and estimated prevalence of inadequate intakes" between 1961 and 2011. The dataset uses a variety of figures to visualise the prevalence of inadequate intake of 14 micronutrients globally. The journal suggests "at the global level, micronutrients with the lowest levels of adequate estimated intake are calcium, iron, vitamin A, and zinc, but there are strong differences between countries and regions". The aim of this report is to explore this statement, and whether it is viable to apply a variety of Supervised Machine Learning Algorithms, to correctly predict and classify a record's global region based on the micronutrient features found in this dataset.

## Dataset Identification

The “S4\_Dataset” from the journal, can also be found on fig share. The dataset contains 111817 rows of records, ordered by country and year for each micronutrient.

|  |  |
| --- | --- |
| Attribute | Description |
| Zone | Global Region the record is from in string form. |
| Country | The country the record is sourced from in string form. |
| ISO3 | The 3-letter code identifier for a country in string form. |
| Year | The year in which the data record is collated from between 1961-2011 in int form. |
| Population | The total population of the country record is sourced from in float form. |
| Fortification | The binary classifier which identifies if a country supplements micronutrients to a nation’s food supplies (via food and water) in int form. |
| PCDEA | National per capita daily energy availability in int form measured in Kilocalories (Kcal). |
| MDI | Micronutrient Density Index for each country-year in float form. |
| Tagname | Abbreviated identifier for Micronutrient in string form. |
| Micronutrient | Identifies the specific Micronutrient data the record is referring to. One of the 14 Micronutrients in string form. |
| Units | The measurement unit specified for all 14 micronutrients in string form. |
| Estimated Intake | Explained in journal as an estimate of micronutrient intake "per capita derived from FBS data by calculating coefficient of variation of intake based on within-subject variation from published dietary intake studies” in float form. |
| Requirements | Refers to "Estimated average requirements (EARs) obtained by the Institute of Medicine for all nutrients except zinc, iron, and calcium" in float form. |
| Prevalence of Inadequate Intake | The result from the "EAR cut-point method to estimate based on approximated micronutrient intakes. It "relies on three assumptions: the distribution of intakes varies more than the distribution of requirements; the distribution of requirements is symmetrical; and intakes and requirements are not correlated" in float form. |

Table 1

## Supervised Learning Task Identification

As mentioned in the introduction, from the S4\_Dataset, the aim is to develop and apply a machine learning model which will predict the global region (Zone) attribute, basing its prediction on the other features and attributes of the dataset shown in Fig. 1.

This model will be using classification algorithms to correctly identify each record’s corresponding Zone. In these models (other than Deep Learning), the Zone feature is label encoded, meaning each Zone is represented by a corresponding numerical value (0-7). As such, traditional binary classification methodology is limited in this application and requires visualisations to be adjusted.

The applied models which will be explored are Logistic Regression, Support Vector Machine (SVC), Random Forests, K-Nearest Neighbour, Multinomial Naïve Bayes, Decision Tree Classifier, and Deep Learning. Their performance will be compared and tested under different pre-processing and sampling methods.

# Exploratory Data Analysis Process and Results

## 2.1 Question(s) Identification

|  |  |  |
| --- | --- | --- |
| Question No. | Question | Assumption |
| 1 | What is the spread of Zones in the dataset? Is it balanced data? | The dataset in the Zones column is spread among 8 categories, with a count of 29120 values at highest and 3640 at lowest. This means the Zone category is extremely unbalanced. |
| 2 | Does MDI effect Micronutrient intake? | Beal states the MDI decline from “1979 to 1993 in sub-Saharan Africa was due to the increased availability of lower micro-nutrient density grains and vegetable oils”, whilst a decreased proportionality of micronutrient rich products. |
| 3 | Does the dataset have any null or missing values to be handled? | There are no null values or missing data that needs to be imputed. |
| 4 | Does computational speed of each model provide any relevance? | As the aim is correct classification of data, accuracy of prediction algorithm should be prioritised, but computational speed will be considered in this report to provide comprehensive performance review for each algorithm on this set of data. |
| 5 | What is the importance of the global nutrient database? | As mentioned in the discussion of the (Schmidhuber et al., 2018) article, “the global nutrient database provides the opportunity to answer important questions about macronutrients and micronutrients across nations. For example, the database can be used to identify countries with insufficient supplies of specific nutrients that are therefore at risk of nutrient deficiencies. Additionally, the database can be used to determine the dependency on specific foods for each nutrient in a country and over time. This information is necessary to make both agriculture and trade more nutrition-sensitive and to inform food-based interventions to prevent micronutrient deficiencies in high-risk countries.” |
| 6 | What is the supply trend of micronutrients globally? | Schmidhuber also mentions the “supply of micronutrients has generally increased between 1980 and 2013 across levels of development”. |
| 7 | Does Fortification influence a population’s micronutrient consumption? | It is within a government’s best interest to ensure their population is well-nourished, thus national policy for fortification of food products is common. As mentioned in (Mannar and Sankar, 2004), studies by World Bank shows that “countries whose populations suffer from micronutrient deficiencies encounter economic losses as high as 5% of gross domestic product (GDP)” and later goes on to suggest solutions as implementing education programmes, targeted distribution programmes, or “fortifying commonly eaten foods with the missing micronutrients” ensuring the “stability and bioavailability of the nutrient”. |
| 8 | Where globally is the highest PCDEA, MDI, Estimated Intake? | Schmidhuber mentions the energy availability widely varied across levels of development ranging from 2170 kcal (2090–2250) per person per day in low-SDI countries to 3270 (3220–3310) kcal per person per day in high-SDI (Socio-demographic Index) countries. Among the most populous countries, the highest level of energy availability was in the USA (3500 kcal [3450–3560] per person per day) and the lowest was in Ethiopia (1880 kcal [1810–1940] per person per day).  Beal finds the 5 countries with highest MDI in 2011 were “Antigua and Barbuda, Kazakhstan, Armenia, Albania, and Serbia”.  Additionally, Beal finds estimated intake increased globally through fortification of certain micronutrients. |

Table 2

## 2.2 Exploratory Data Analysis Process and Results

Using the head function, the first few data records in the data-frame can be displayed as shown in Fig. 1, allowing for the visualisation of the data and give a better understanding of the data held in the dataset and how they interact.

A table with numbers and letters

Description automatically generated

Fig. 1

Using the Pandas plot function, Fig. 2 shows the percentage of Zone category data. The pie chart shows large imbalance in data records to corresponding categories. This visualisation helps better illustrate the scale of imbalance.

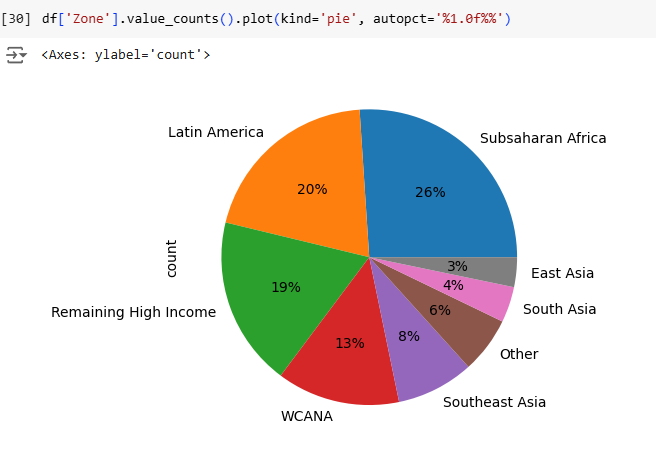


Fig. 2

Histograms are another useful visualisation to show the distribution of data for each category, showing the skewness of a specific feature. Commonly the aim is to have data in a normal distribution often referred to as a ‘bell curve’ shape. As the data is from the real world, it is accepted that skewness and imbalance is present in the visualisations. However, to such a large extent with large outliers can cause bias and skew predictions towards these outliers. As such, it is important to recognise what data must be kept, maintaining the characteristics of a feature, and what information is likely to induce bias.

A graph of different sizes and shapes

Description automatically generated with medium confidence

Fig. 3 (Using logarithmic transformation on population and requirements)

An example of a case where the graph must be interpreted is fortification in Fig. 3 where data is 0 or 1. Due to this, the 1 value are not outliers, but a fundamental characteristic to contextualise the following data of prevalence of inadequate intake.

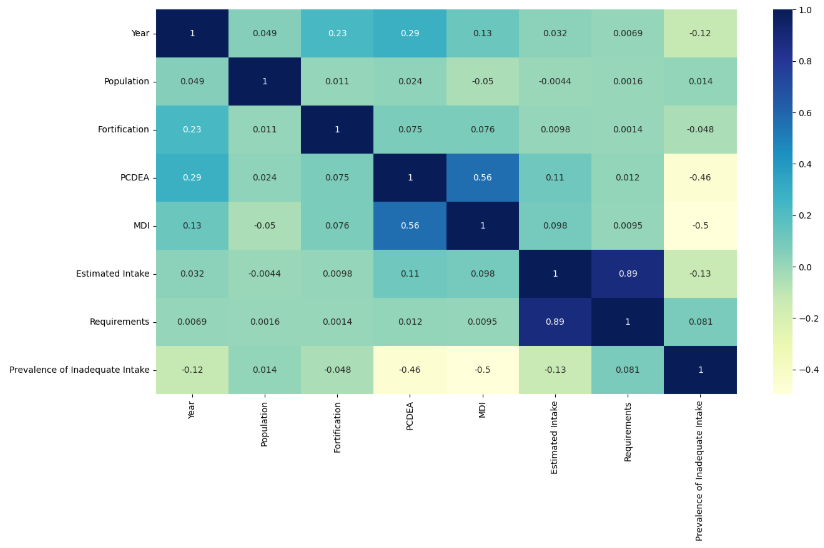


Fig. 4 (Heatmap)

The heatmap in Fig. 4 is produced using the ‘.corr’ function on the data-frame shows the diagonal which correctly shows each feature correlates with itself.

The Heatmap on the raw data additionally shows strong positive correlation between Requirements and Estimated intake - a score of 0.89. This states that as a population's requirements for micronutrients increase, the estimated intake of a population also increase. This is interpreted as a government being able to meet the growing micronutrient on average.

The Heat map also shows a positive correlation between PCDEA and MDI - a score of 0.56. This states as National per capita daily energy availability increases, the micronutrient density index increases each year. This is interpreted as a government being better able to facilitate its growing population's micronutrient needs on average, thus increasing MDI.

The Heatmap shows a negative correlation between Prevalence of inadequate intake and the features: PCDEA and MDI - a score of -0.46 and -0.5 respectively. This is interpreted as the increase in Prevalence of inadequate intake correlates to decreased decreasing national capita daily energy capacity and decreasing Micronutrient density index.

A screen shot of a computer screen

Description automatically generated

Fig. 5 (Scatterplot of population, estimated intake, prevalence of inadequate intake)

Visualisations such as the scatter-graph in Fig. 5 give a broader view of the features of a dataset and how they interact.

The scatterplot shows that the prevalence of inadequate intake disregards population factors and is more heavily concentrated where estimated intake is at 0.5. The highest estimated intake records have the lowest Prevalence of Inadequate Intake value which aligns with the view mentioned in the 2.1 Questions section, that high intake can lead to higher levels of obesity and other health issues. A further point for other reports to explore would be to see if these values are associated with global regions with high GDP or low GDP per capita and other forms of socio-economic analysis.

## 2.3 Exploratory Data Analysis Conclusions

1. Target column data (Zone) is imbalanced with regions having counts ranging from 29,120 to 3,640 values.

2. There is some correlation between some features, but most have no correlation to one another (Fig. 4).

3. The distribution of data is skewed but relevant to the context of the problem (Fig. 3).

4. There is a mix of categorical and numerical features in the dataset which need to be encoded.

5. The target feature has 8 different outputs.

# Experimental Design

## Identification of Chosen Supervised Learning Algorithm(s)

This report will focus on K Nearest Neighbour, Random Forests, Decision Tree Classifier, and Deep Learning. The other three algorithms: SVM, Logistic Regression and Multinomial Bayes will be referenced but used as comparative algorithms.

|  |  |
| --- | --- |
| Algorithm | Description |
| Deep Learning | Deep learning utilises layered nodes with input and activation nodes which output based on their function type. Deep Learning utilises a loss function which determines the weight change in each epoch until the loss function is successfully minimised or epoch number is reached. |
| Decision Tree Classifier | (www.javatpoint.com, 2021) states Decision Tree is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.  In essence:   1. Split data 2. Recursively iterate through tree 3. Decision making with Entropy   Entropy in a decision tree measures the impurity of a node – i.e. the uncertainty. |
| K-Nearest Neighbour | (IBM, 2021) explains that K-nearest neighbours algorithm uses proximity to make classifications or predictions about the grouping of individual data points. It works off the assumption similar points can be found near one another. It is like Random Forests, utilising majority voting concept (requires > 50%). |
| Random Forests Classifier | Random Forests Classifier Algorithm merges the outputs of multiple decision trees to produce a single outcome. Random Forests uses ensemble modelling, where a classification goes through majority voting.  In essence:   1. subset of data points and features are selected for constructing each decision tree 2. Individual decision trees are constructed for each sample 3. Each decision tree will generate output 4. Final output is considered based on majority voting |

## Identification of Appropriate Evaluation Techniques

Using accuracy, precision, recall and f1-score helps to explain effectiveness for datasets with imbalance .

Accuracy, precision and recall will give the surface level insight into the predictions the model makes and will be easy to compare to the other models.

However, most comparisons will be made utilising the F1-score, which gives a more contextualised view of the confusion matrix, considering false positive and negative predictions. In essence, the F1-score is the harmonic mean of precision and recall.

## Data Cleaning and Pre-Processing Transformations

As shown in Fig. 6, the dataset has 111,818 entries with each feature holding 111818 non-null values, meaning no data needs to be imputed. It should be noted that the data types differ for each feature.

A screenshot of a computer

Description automatically generated

Fig. 6 (Before Pre-Processing)

A screenshot of a computer

Description automatically generated

Fig. 7 (After Pre-Processing)

In Fig. 7 a mix of both integer and float values are present. It is important to keep nuance of each feature which is a float, as rounding the values would give an inaccurate representation of data, and with the specified quantity or micronutrients being in smaller units, it is important that the precision remains unchanged.

A screenshot of a computer

Description automatically generated

Fig. 8

A close-up of a computer screen

Description automatically generated

Fig. 9

In Fig. 8, the dataset can be determined to be already clean with no null values and there are no duplicate values as shown in Fig. 9.

A screenshot of a computer

Description automatically generated

Fig. 10

A screenshot of a computer

Description automatically generated

Fig. 11

Synthetic Minority Oversampling Technique (SMOTE) is applied in Fig. 10 and Near-Miss Under-sampling Technique Fig. 11 is applied as a comparison to the execution of the unbalanced dataset data. As seen in both figures, X (the features to train model) and y (the target feature). X and y variables are then applied to fit their respective resampling technique before they are split into train and test variables, which are then scaled by a scaler algorithm.

A screenshot of a computer

Description automatically generated

Fig. 12 (Hyperparameter tuning table from (Google Cloud, 2024))

The algorithms in focus for this report and their adjustable hyperparameters are shown in Fig. 12 which the article (Google Cloud, 2024) states “in machine learning, hyperparameter tuning identifies a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins”.

A screenshot of a computer

Description automatically generated

Fig. 13

A screenshot of a computer

Description automatically generated

Fig. 14

As shown in Fig .13, values for each feature are scaled. (Punyakeerthi BL, 2024) states feature scaling is a fundamental preprocessing step in machine learning aimed at ensuring that numerical features have a similar scale.

A graph of a train accuracy

Description automatically generated with medium confidence

Fig. 15 (Hyperparameter testing for Random Forest Classifier)

A screenshot of a computer

Description automatically generatedFig. 16 (Grid Search for Hyperparameter tuning for Random Forest Classifier)

A screenshot of a computer screen

Description automatically generated

Fig. 17 (Comparing Standard and Min Max scaler output)

Principal components Fig. 17 in compute the percentage of variance (information) accounted for by each component.

The full data is a 14-dimensional point cloud, and these points are the projection of each data point along the directions with the largest variance (highest information axes). Essentially, we have found the optimal stretch and rotation in 14-dimensional space that allows us to see the layout of the data in two dimensions, and we have done this in an unsupervised manner—that is, without reference to the labels.

i.e. There are clear splits and groupings of data represented by both scalers with similar performance as shown in Fig. 13 when compared to results in Fig. 31 and Fig. 32.

A blue line graph with numbers

Description automatically generated

Fig. 18 (PCA result Random Forest)

PCA analysis in Fig. 18 shows 80% information from dataset can be maintained with 5 components. This would reduce the computational requirement without losing performance of Random Forest model.

A graph with blue lines

Description automatically generated

Fig. 19 (Loss score each epoch in Deep Learning)

A graph with blue lines

Description automatically generated

Fig. 20 (Accuracy of training and validation over epochs)

In Fig. 19 it is seen that as number of epochs increase from around 10, validation and training loss decreases significantly before beginning to plateau with little to no improvement as epochs increase. Similarly in Fig. 20  after 10 epochs, there is a linear positive correlation in training accuracy. Validation accuracy increases to extreme peaks and troughs throughout the increasing epochs but generally trends in line with Training accuracy in a positive correlation.

## Limitations and Options

A screenshot of a computer

Description automatically generated

Fig. 21

* + Limitations is the near miss and smote do not like y being encoded with multiple categories, causes a classifier issue (as seen in (fig. 21).
  + Deep learning confusion matrix does not like multiple categories, will compare zone by zone.
  + Deep Learning multiple labels could be selected by prediction, not limited to one category prediction.
  + Overfitting, model too specific and ungeneralised due to some redundant features or unintentional data leakage.

# Predictive Modelling / Model Development

## 4.1 Splitting the Dataset

A computer screen shot of a computer code

Description automatically generated

Fig. 22

Except for Deep learning, the split of X and Y split keeps the target data ‘zone’ unseen to the model predicting the data. This will split will be utilised for each model. (Tim Mucci, 2024) states Data Leakage occurs when a model uses information during training that would not be available at time of prediction and is a key concept to be avoided. He states there are two types of leakage: Target and Train-Test contamination. The commented X in figure 1, shows a previous iteration which reduced the overfitting/data leakage issue by hiding more data from the model. However, this made the classification more akin to a black box, with features and patterns which were not visible to the human in the minimal data given, would give a prediction. The aim of developing the model is to have it generalised when test with unseen data.

A screenshot of a computer program

Description automatically generated

Fig. 23

Except for Deep Learning, the train test split is 60/40, which provides an opportunity to utilise validation. However, in this case it is used to prevent the overfitting of each model on the training data.

A screenshot of a computer program

Description automatically generated

Fig. 24

A screenshot of a computer

Description automatically generated

Fig. 25 (Encoding for Deep Learning Example of y values)

For deep learning, the use of one-hot encoding in pre-processing as shown in Fig.25 separated each categorical input of Zone, utilising the ‘pd.get\_dummies’ to encode into a data-frame.

## 4.2 The Predictive Modelling Process

A screenshot of a computer code

Description automatically generated

Fig. 26

A screenshot of a computer

Description automatically generated

Fig. 27

A screenshot of a computer program

Description automatically generated

Fig. 28

To fit values before applying them to algorithms such as in Fig. 26, Fig. 27, Fig. 28 with a scaler, SMOTE, Near-Miss adjusted variables, change the predict parameter to the adjusted variable.

A screen shot of a computer code

Description automatically generated

Fig. 29

As stated in (Furnieles, 2022), SoftMax is an activation function like sigmoid. Soft max applies to multiclass problems which is suited to the problem as 'zone' is separated over 8 potential classifications. Additionally soft max is already vectorised "meaning that takes a vector with the same number of entries as classes we have and outputs another vector where each component represents the probability of belonging to that class".

## 4.3 Evaluation Results on “Seen” Data

### 4.3.1 Grid Search

As shown in Fig.16 the use of Grid Search helps set the hyperparameters utilised to optimise prediction performance, and computational processing speed e.g. max\_features = 8 and n\_estimators = 30 for Grid search on Random Forests.

A screenshot of a graph

Description automatically generated

Fig. 30 (Grid Search Random Forest Result)

Random forests using the grid search optimised parameters leads to an extremely high prediction accuracy. This is extenuated by the now scaled yet unbalanced data. Additionally, the data leakage of the most important features is another cause of the extremely high accuracy. Evidently performing better than the previous Random Forests application.

### 4.3.2 Predictions Using Over Sampling (SMOTE)

Formulating data for features with a lower count to balance dataset output values.

A screenshot of a graph

Description automatically generated

Fig. 31

A screenshot of a graph

Description automatically generated

Fig. 32

A screenshot of a graph

Description automatically generated

Fig. 33

Despite synthetic data, K nearest neighbour (KNN) is still able to predict with high f1-score and accuracy and should be tested on test data without synthesised data. The Random Forest Classifier with and without grid-search perform similarly well for differing computational load. Decision Tree classifier performs expectedly worse than Random Forests yet still outperforms KNN.

### 4.3.3 Predictions Using Under Sampling (Near Miss)

It works by selecting samples closest to the minority class and deletes records so all categories have the same number of samples.

A screenshot of a graph

Description automatically generated

Fig. 34

A screenshot of a graph

Description automatically generated

Fig. 35

A screenshot of a graph

Description automatically generated

Fig. 36

KNN performs less accurately than when oversampled, Random Forest with Grid Search params performs the best of all algorithms with minimal spread on inaccuracy.

### 4.3.4 Deep Learning

A blue square with numbers and a blue square

Description automatically generated

Fig. 37

A screen shot of a graph

Description automatically generated

Fig. 38

A blue and white graph

Description automatically generated

Fig. 39

A blue squares with numbers and a number of labels

Description automatically generated with medium confidence

Fig. 40

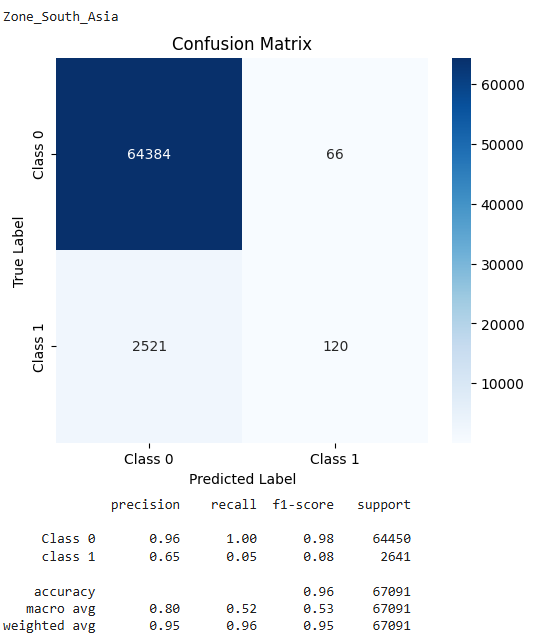
#

Fig. 41

A blue and white graph

Description automatically generated

Fig.42

A blue squares with numbers and a blue rectangle

Description automatically generated

Fig. 43

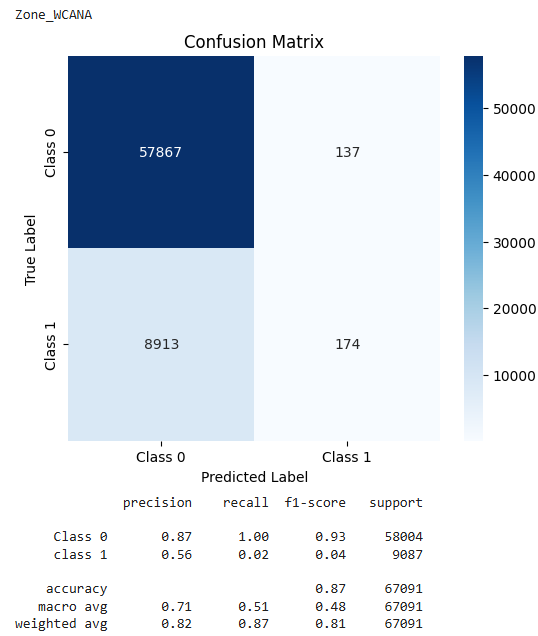


Fig. 44

Initially y was converted from “nd.array” format back to a data-frame type (refer to Fig. 25) and compared the prediction for each category individually and independent of each other. Overall, the result for each category is predicted to a satisfactory f1-score and accuracy. Many False Negative predictions can be seen in Latin America and WCANA predictions. Many false positives in Southeast Asia and Remaining High income.

The Class1 true and predicted id data correctly classifying the item not being in the column "0" for this category (already classified in another category).

Train test split for Over sampling and near miss struggles to deal with one hot encoded train test split.

# Evaluation and Further Modelling Improvements

|  |  |
| --- | --- |
| Discussion Topic Questions | Answer |
| Which models performed the best? Were there big differences? | Models involving decision tree, majority voting or layers performed better for this dataset than other algorithms that do not use any one of the listed features did. As seen by some models predicting with <50% accuracy. |
| What were uniquely prevalent issues with the dataset? | Due to being a classification model with multiple categories for outputs, there were issues with classification reports for deep learning. It should also be noted huge imbalance in output data meant predictions is susceptible to skew and bias, meaning the model is affected by data leakage. |
| Do you think it's more important to be sensitive to negativity or positivity? Do we want more positive things incorrectly marked as negative, or more negative things marked as positive? | Whether an error is positively or negatively predicted is irrelevant for this dataset, as it is not a binary classification, the error could be one of 7 other categories. In the case of deep learning, the focus is on the accuracy, precision, recall and f1-score of the prediction for each correctly predicted item. |
| They all had very different training times. Which ones offer the best combination of performance and not making you wait around for an hour? | Random Forest Classifier had the best combination of prediction performance and training time. Performing well in all scenarios and under different sampling methods. |
| Is 75% accuracy good? | In this case, accuracy is a viable metric of consideration as a multiple category output. The aim is for each item to be precisely and uniquely identifiable. So, 75% across all categories is "good". However, in the last visualisation of deep learning, when comparing one category at a time, accuracy is a less prevalent metric of consideration and does not properly highlight the nuance of the confusion matrix that f1-score does. |

* + Further modelling improvements could be applying hyperparameter tuning to all machine algorithms in focus for the report.
  + Finding a method to apply SMOTE and Near Miss without errors to Deep Learning to see how the algorithm dals with synthetic data.
  + Applying the one-hot encoding to all algorithms and comparing the results to the label transformed Zone feature which means it is easier to identify if the algorithm is predicting multiple regions for one record.
  + Utilising PCA more to help fix the overfitting and overly accurate predictions on training data and test data.
  + Making the models more generalised by introducing data from 2012-2024.

# 6 Conclusion

## 6.1 Summary of Results

In summary, with such high accuracy values in the initial data (baseline testing seen in colab) demonstrates the model is likely not generalised and suffers from data leakage, likely sourced from one of the higher importance features in the importance table in Fig. 14. However, the performance in cases where the values are scaled, under or over sampled, perform statistically worse, but are better generalised to the task and are less likely to be overfitting as severely. Deep Learning with the one-hot encoded Zone categories gives great one to one data analysis, but an overall confusion matrix cannot be produced and there is a possibility the algorithm is predicting the record fitting into multiple categories unlike the other algorithms predicting 0-7 in Zone. Due to data leakage in the baseline, it is difficult to comment on the reliability and effectiveness of each model and their f1-scores. With further testing and generalisation, the application of classification prediction of global region based on micronutrient values is possible.

## 6.2 Reflection on Individual Learning

This module has sparked an excitement in my development as a computer scientist and will likely be a prominent feature driving my progression and passion for the field. It has been an eye-opening course brilliantly delivered, going through the trial by fire, which is the development of the code, encountering errors in all sorts of parts of the project has been fantastic, empowering me to no end. I feel I understand the machine learning rabbit-hole better, but I understand that this is merely the surface. From watching YouTube videos about reinforcement learning on a visual model, to having learnt its fundamentals in lectures has been surreal. I am glad the choice was made to apply this project on real-world and a dataset I sourced myself. I will most definitely look forward to undertaking more machine learning projects in future, big data, autonomation, etc.

# 7 References

[1] Beal, T., Massiot, E., Arsenault, J.E., Smith, M.R. and Hijmans, R.J. (2017). Global trends in dietary micronutrient supplies and estimated prevalence of inadequate intakes. *PLoS ONE*, [online] 12(4), pp.e0175554–e0175554. doi:https://doi.org/10.1371/journal.pone.0175554.‌

[2] Schmidhuber, J., Sur, P., Fay, K., Huntley, B., Salama, J., Lee, A., Cornaby, L., Horino, M., Murray, C. and Ashkan Afshin (2018). The Global Nutrient Database: availability of macronutrients and micronutrients in 195 countries from 1980 to 2013. *The Lancet Planetary Health*, [online] 2(8), pp.e353–e368. doi:https://doi.org/10.1016/s2542-5196(18)30170-0.

[3] Mannar, M.G.V. and Sankar, R. (2004). Micronutrient fortification of foods — rationale, application and impact. *The Indian Journal of Pediatrics*, [online] 71(11), pp.997–1002. doi:https://doi.org/10.1007/bf02828115.

[4] Tim Mucci (2024). *Data leakage machine learning*. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/data-leakage-machine-learning [Accessed 11 Dec. 2024].

[5] Google Cloud. (2024). *Hyperparameter tuning overview*. [online] Available at: https://cloud.google.com/bigquery/docs/hp-tuning-overview#:~:text=In%20machine%20learning%2C%20hyperparameter%20tuning,a%20linear%20model%20are%20learned. [Accessed 11 Dec. 2024].

[6] Furnieles, G. (2022). *Sigmoid and SoftMax Functions in 5 minutes - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/sigmoid-and-softmax-functions-in-5-minutes-f516c80ea1f9 [Accessed 11 Dec. 2024].

[7] www.javatpoint.com. (2021). *Decision Tree Algorithm in Machine Learning - Javatpoint*. [online] Available at: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm [Accessed 11 Dec. 2024].

[8] IBM (2021). *KNN*. [online] Ibm.com. Available at: https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20non,used%20in%20machine%20learning%20today. [Accessed 11 Dec. 2024].

[9] Google Cloud. (2024). *Hyperparameter tuning overview*. [online] Available at: https://cloud.google.com/bigquery/docs/hp-tuning-overview#:~:text=In%20machine%20learning%2C%20hyperparameter%20tuning,a%20linear%20model%20are%20learned. [Accessed 12 Dec. 2024].

[10] Punyakeerthi BL (2024). *Understanding Feature Scaling in Machine Learning - Punyakeerthi BL - Medium*. [online] Medium. Available at: https://medium.com/@punya8147\_26846/understanding-feature-scaling-in-machine-learning-fe2ea8933b66 [Accessed 12 Dec. 2024].