Sudanese Primary Schools Dataset Analysis and Classification based on Facility Availability

Data Preparation/Feature Engineering

1. Overview

The main aim of this project is to:

- analyze Sudan school's dataset and visualize it
- classification of schools according to facilities

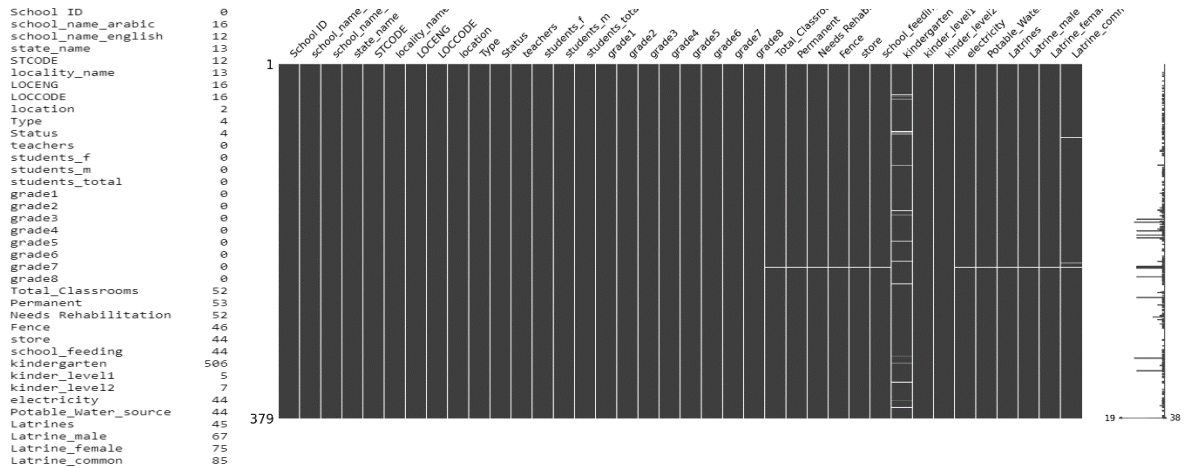
1. Dataset
   - Dataset preparation
   - Exploratory data analysis (EDA)
   - Prepare dataset for machine learning model (Feature engineering)
2. Classification
   - Machine learning models (random forest, knn, svm, naive bayes, nn)
   - Comparison

2. Data Collection

The project aims to analyze Sudanese schools' dataset and classify it according to the facilities to enhance the understanding of the current state of Sudanese schools. As known, Sudan one of the least development countries which still need many steps of development in many aspects but the most important is to provide well based education environment and suffers from proper dataset that can help in analyzing the situation and making good decisions. An impressive collaboration between the Ministry of Education, UNICEF, and OCHA Sudan have been done in 2021 in collecting the school's data in order to ease the decision-making process and to know the actual statics of students, teachers and facilities availability as well.

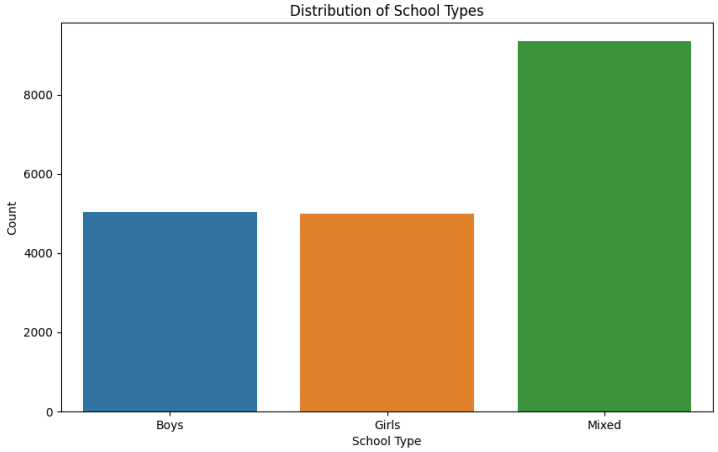
3. Data Cleaning

As the dataset is quite big, dropping all missing data rows seems a suitable solution. Dataset size changed from 19379 ==> 18716 only 663 rows have been dropped. Below is visualization of missing data distribution across dataset.

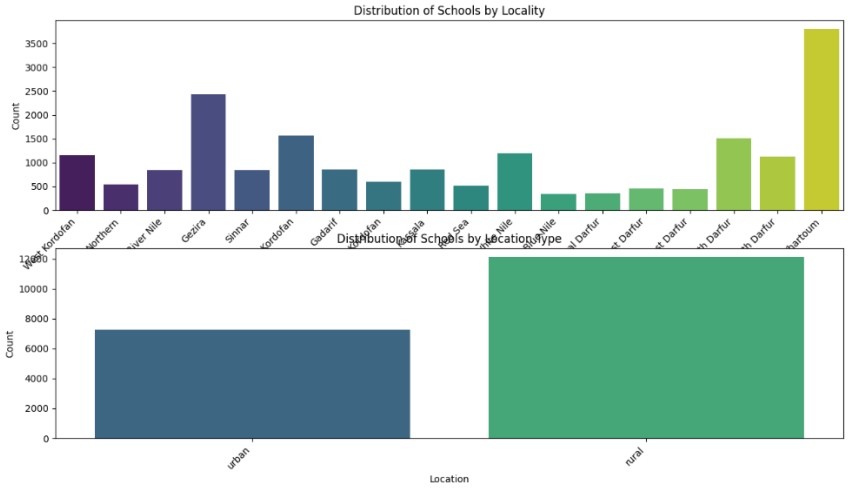


4. Exploratory Data Analysis (EDA)

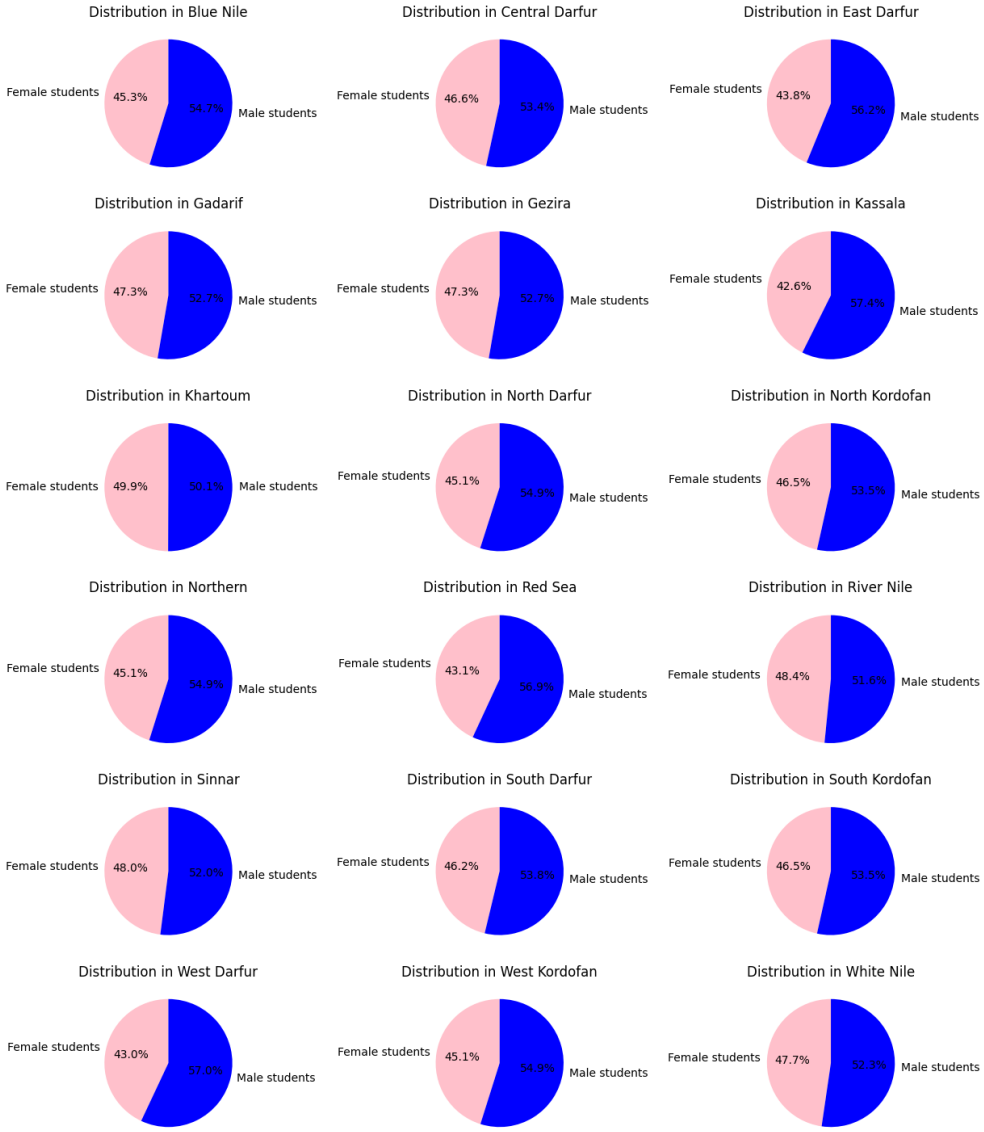
4.1. Distribution of Sudanese school types



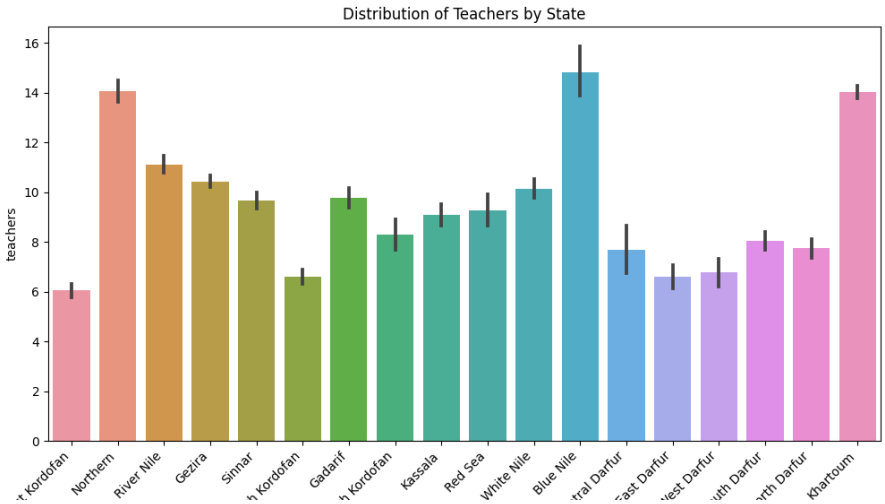
4.2. Distribution of Sudanese by locality and areal type (urban, rural)



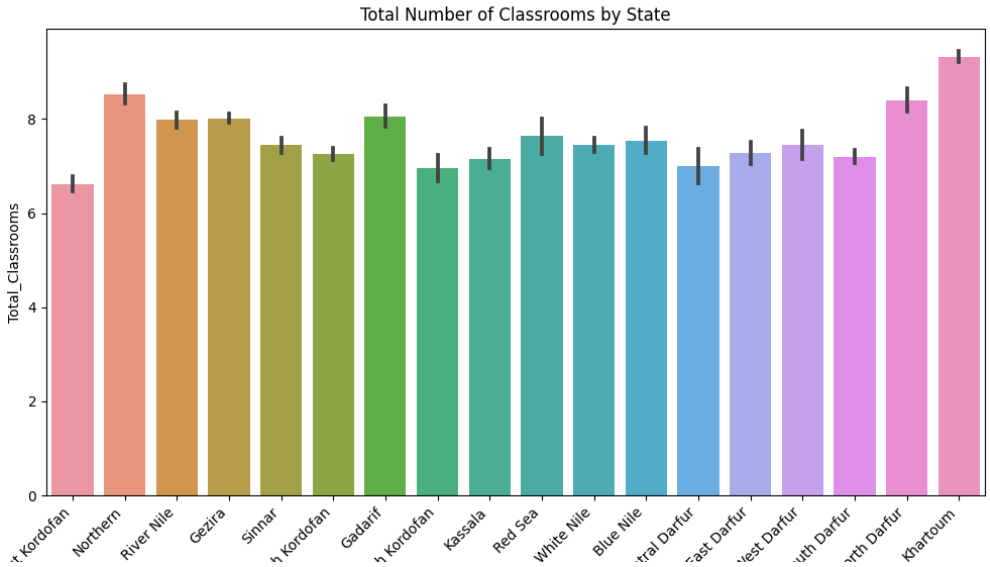
4.3. Distribution of Sudanese school male-female across states



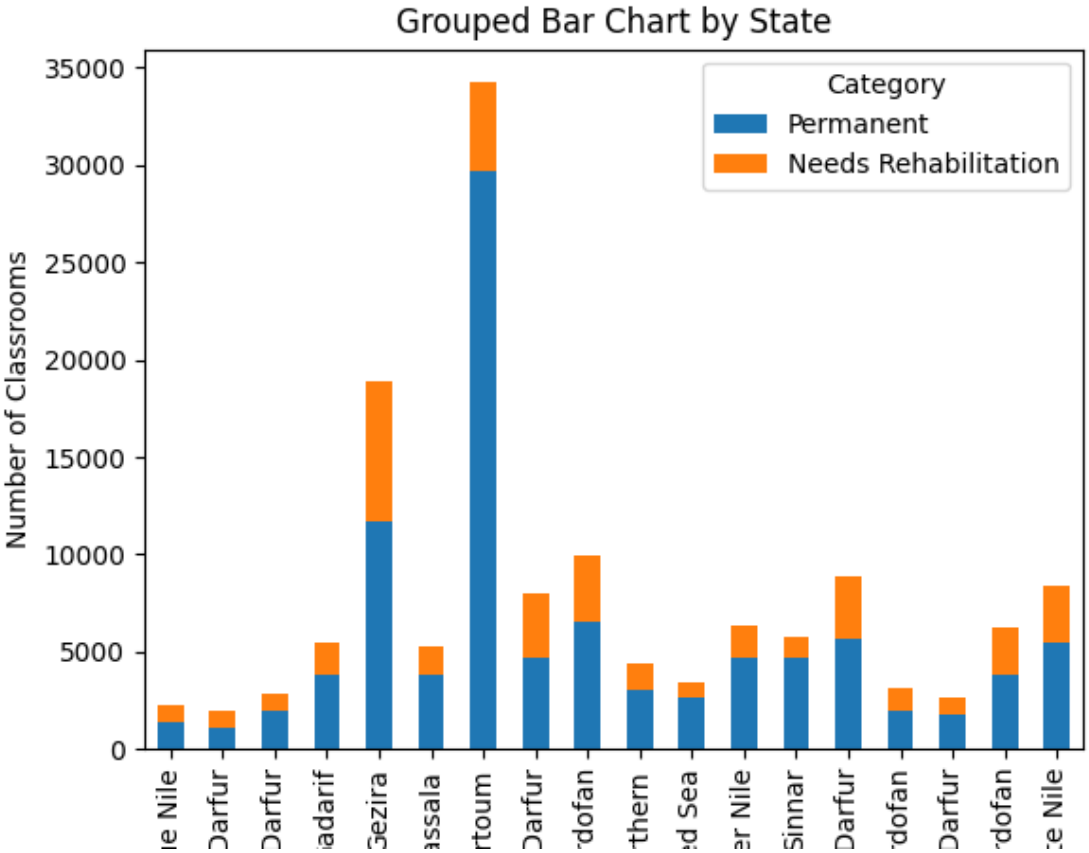
4.4. Distribution of teachers across different states



4.5. Total number of classrooms in each state



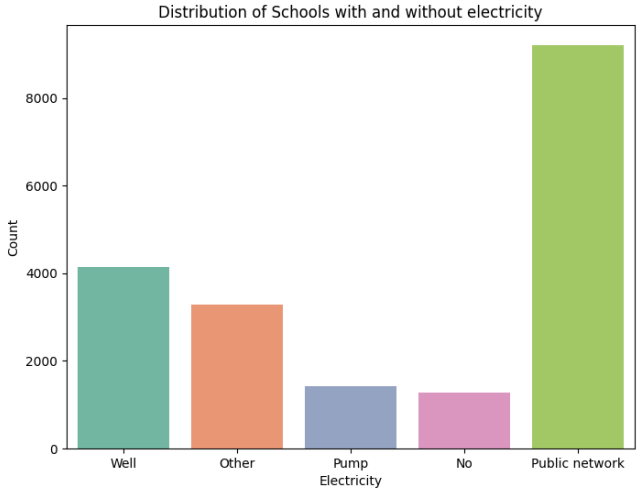
4.6. Classes status based on permanent or need rehabilitation.

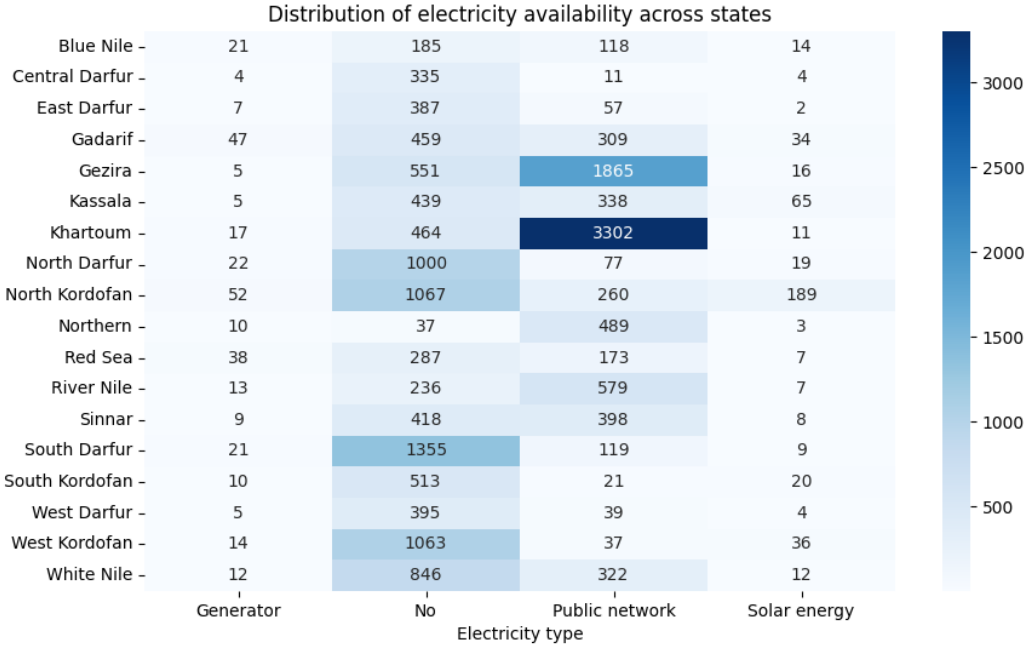


4.7. Distribution of schools based on their status.

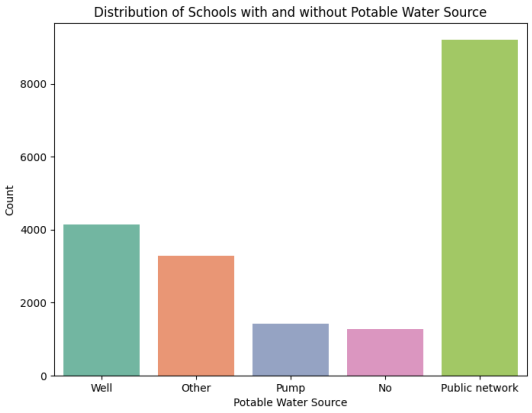
Status	Count	Percentage
normal	14745	76.10%
nongovernmental	3141	16.21%
nomadic	1130	5.83%
special needs	233	1.20%
quranic	72	0.37%
displaced	33	0.17%
complementary	21	0.11%

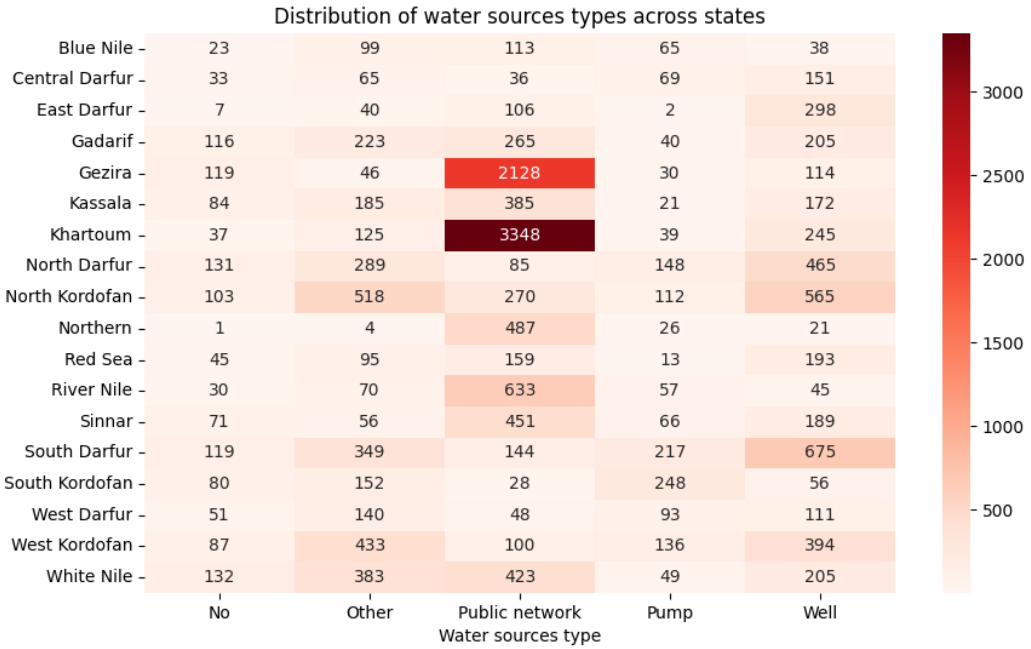
4.8. Distribution of Schools with and without electricity



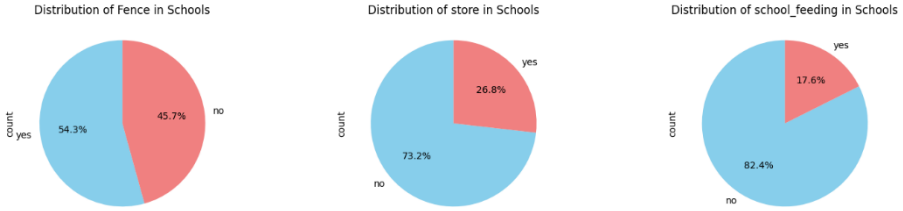


4.9. Distribution of Schools with and without water resources





4.10. Distribution of fence, store, school feeding availability



3. Feature Engineering and Data Transformation

3.1. Drop unnecessary features

- 1st and 5th columns written in Arabic and as both have similar information column written in English
- School name, state name, location (LOCENG)
- School grade1 to grade8 columns

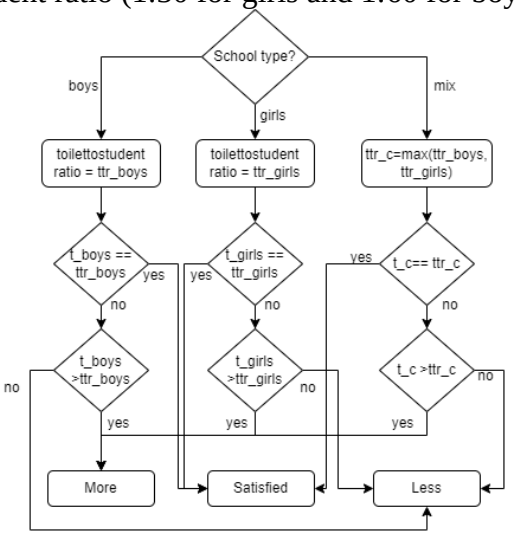
3.2. Convert categorical to numerical

- edit STCODE and LOCCODE columns to be only the numerical part without "SD" (SD01 – SD18)
- Convert yes/no to 1/0 in columns; Fence, Store, School_feeding, Latrines
- check unique classes in 'location', 'Type', 'Status', 'electricity', 'Potable_Water_source' columns convert them to numerical.

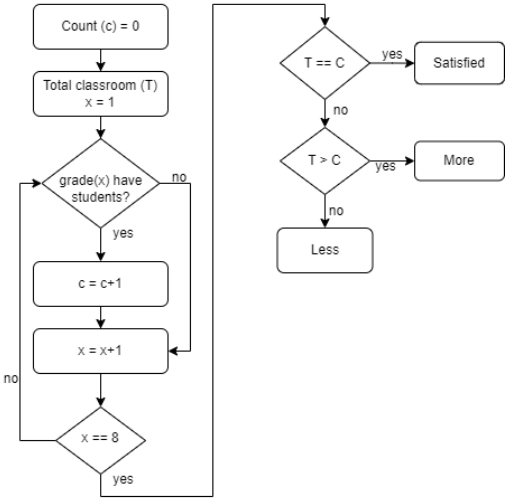
```
Unique classes in location: ['urban' 'rural']
Unique classes in Type: ['Boys' 'Girls' 'Mixed']
Unique classes in Status: ['normal' 'nomadic' 'nongovernmental' 'special needs' 'quranic'
 'complementary' 'displaced']
Unique classes in electricity: ['No' 'Solar energy' 'Generator' 'Public network']
Unique classes in Potable_Water_source: ['Well' 'Other' 'Pump' 'No' 'Public network']
```

3.3. Create new features

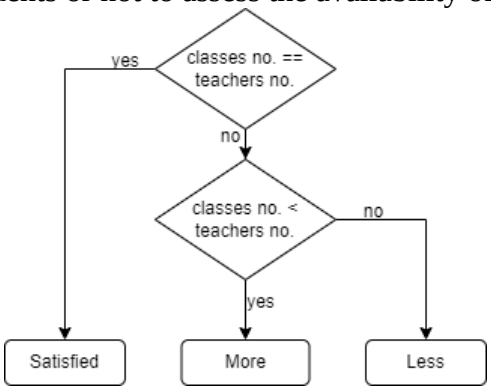
- Toilet to student feature, according to UNICEF report (UNICEF Sudan - Investment Case - Back to School, Back to Learning) the average recommended ratio of toilet to student ratio (1:30 for girls and 1:60 for boys) [1].



- Class to student, this feature made to check if the number of classrooms equals the number of grades that's have student. It will check every grade if it have student or not then compare it with the total classrooms to assess the availability of classrooms.

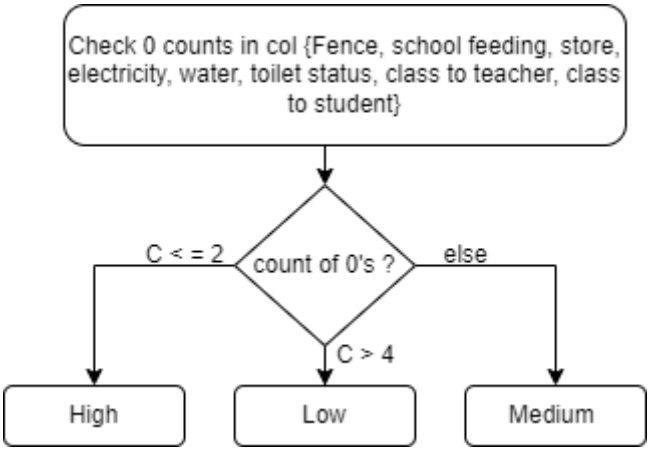


- Class to teacher, this feature to check if number of teachers equals to number of grades that have students or not to assess the availability of the teachers.

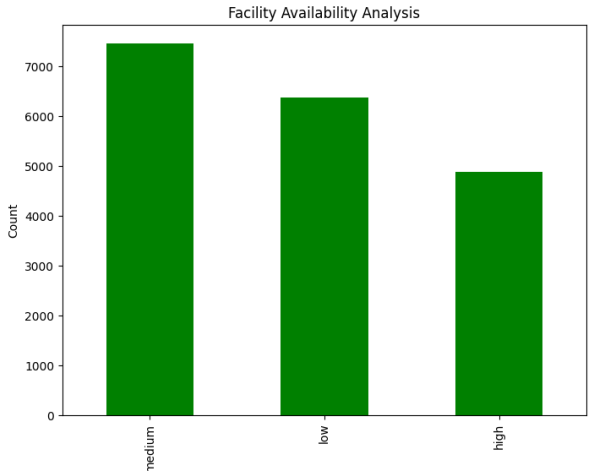


3.4. Labeling

As the data is not labeled, facility features are taken into consideration to make labeling. The algorithm works to check if the facility is available or not. It counts the number of non-available facilities. If more than 4 it considers the school as low facility if less than or equal 2 it consider it as high facility otherwise it consider as medium.



The following images show the distribution of school labels across the Sudanese primary schools.



3.5. Prepare data for machine learning

- Determine x, y values.
- Standardization using StandardScaler.
- Feature importance (29 features, 8 most important have been selected).
- Splitting (train/test)

Classification Model Exploration

1. Models, training, evaluation and code implementation

Training set = (14972, 6) (14972,)
Test set = (3744, 6) (3744,)

Model	Metrices
Naïve bayes	Fitting 10 folds for each of 20 candidates, totalling 200 fits Best Parameters: {'var_smoothing': 0.11288378916846892} Classification report precision recall f1-score support 0 0.96 1.00 0.98 1292 1 0.95 0.97 0.96 1468 2 1.00 0.92 0.96 984 accuracy 0.97 3744 macro avg 0.97 0.96 0.97 3744 weighted avg 0.97 0.97 0.97 3744 Class Confusion Matrix [[1292 0 0] [48 1419 1] [0 78 906]]

Random forest	Classification report for Random Forest precision recall f1-score support 0 1.00 1.00 1.00 1292 1 1.00 1.00 1.00 1468 2 1.00 1.00 1.00 984 accuracy 1.00 3744 macro avg 1.00 1.00 1.00 3744 weighted avg 1.00 1.00 1.00 3744 Confusion Matrix for Random Forest: [[1291 1 0] [1 1466 1] [0 0 984]]
KNN	Classification report with k=1 precision recall f1-score support 0 1.00 1.00 1.00 1292 1 1.00 1.00 1.00 1468 2 1.00 1.00 1.00 984 accuracy 1.00 3744 macro avg 1.00 1.00 1.00 3744 weighted avg 1.00 1.00 1.00 3744 Class Confusion Matrix [[1292 0 0] [0 1467 1] [0 1 983]]
SVM	Time taken for linear kernel: 0.1210 seconds Accuracy with linear kernel: 100.00% Time taken for poly kernel: 0.4380 seconds Accuracy with poly kernel: 99.97% Time taken for rbf kernel: 0.6661 seconds Accuracy with rbf kernel: 100.00% Time taken for sigmoid kernel: 4.0587 seconds Accuracy with sigmoid kernel: 74.20% Classification report with linear kernel precision recall f1-score support 0 1.00 1.00 1.00 1292 1 1.00 1.00 1.00 1468 2 1.00 1.00 1.00 984 accuracy 1.00 3744 macro avg 1.00 1.00 1.00 3744 weighted avg 1.00 1.00 1.00 3744 Class Confusion Matrix [[1292 0 0] [0 1468 0] [0 0 984]]

Neural network
Training set = (11977, 27)
(11977, 5)
Validation set = (2995, 27)

(2995, 5)
Test set = (3744, 27)
(3744, 5)

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_2 (Conv1D)            (None, 27, 32)            128

batch_normalization_3 (Batc  (None, 27, 32)            128
hNormalization)

max_pooling1d_2 (MaxPooling  (None, 13, 32)            0
1D)

dropout_3 (Dropout)          (None, 13, 32)            0

conv1d_3 (Conv1D)            (None, 11, 64)            6208

batch_normalization_4 (Batc  (None, 11, 64)            256
hNormalization)

max_pooling1d_3 (MaxPooling  (None, 5, 64)             0
1D)

dropout_4 (Dropout)          (None, 5, 64)             0

flatten_1 (Flatten)          (None, 320)               0

dense_2 (Dense)              (None, 128)               41088

batch_normalization_5 (Batc  (None, 128)               512
hNormalization)

dropout_5 (Dropout)          (None, 128)               0

dense_3 (Dense)              (None, 3)                 387


117/117 [==============================] - 0s 2ms/step
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.91      0.88      1292
           1       0.89      0.78      0.83      1468
           2       0.88      0.97      0.92       984

    accuracy                           0.87      3744
   macro avg       0.87      0.89      0.88      3744
weighted avg       0.88      0.87      0.87      3744

Confusion Matrix:
[[1175  117    0]
 [ 192 1142  134]
 [   0   27  957]]
```

References:

[1] UNICEF Sudan - Investment Case - Back to School, Back to Learning (EU), https://www.unicef.org/sudan/media/9711/file/UNICEF%20Sudan%20-%20Investment%20Case%20-%20Back%20to%20School,%20Back%20to%20Learning%20(EU).pdf