

**A PROJECT REPORT ON**  
**CAN YOU PREDICT WHICH WATER PUMPS ARE FAULTY**

**A project report submitted in fulfillment for the Diploma Degree in AI & ML**

**Under**

**Applied Roots with University of Hyderabad**



**Project submitted by**

**PONAKA BHANU PRADEEP KUMAR**

**Enrolment No: 40AIML343-21/2**

**2021-2022**

**Under the Guidance of**

**Mentor: Harish**

**Approved by: Mentor: Harish**



**UNIVERSITY OF HYDERABAD**

**CENTRE FOR DISTANCE AND VIRTUAL LEARNING**

**GACHIBOWLI, HYDERABAD-500046**

## **Declaration of Authorship**

We hereby declare that this thesis titled” CAN YOU PREDICT WHICH WATER PUMPS ARE FAULTY” and the work presented by the undersigned candidate, as part of Diploma Degree in AI & ML.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name: PONAKA BHANU PRADEEP KUMAR

Thesis Title: CAN YOU PREDICT WHICH WATER PUMPS ARE FAULTY

## **CERTIFICATE OF RECOMMENDATION**

We hereby recommend that the thesis entitled “Can you predict which water pumps are faulty” prepared under my supervision and guidance by Ponaka Bhanu Pradeep kumar be accepted in fulfilment of the requirement for awarding the degree of Diploma in AI & ML Under applied roots with University of Hyderabad. The project, in our opinion, is worthy for its acceptance.

---

Mentor:Harish

Under Applied roots with



University of Hyderabad

## **ACKNOWLEDGEMENT**

Every project big or small is successful largely due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. I sincerely appreciate the inspiration; support and guidance of all those people who have been instrumental in making this project a success. I, P Bhanu Pradeep Kumar student of applied roots, is extremely grateful to mentors for the confidence bestowed in me and entrusting my project entitled “Can you predict which water pumps are faulty” with special reference.

At this juncture, I express my sincere thanks to Mentors Vishnu of applied roots for making the resources available at right time and providing valuable insights leading to the successful completion of our project that even assisted me in completing the project.

**Name: P Bhanu Pradeep Kumar**

## **Contents**

**1. Abstract**

**2. Problem Definition**

**3. Data set Description.**

**4. Key metric to optimize**

**5. Real world challenges and constraints**

**6. Similar problems solved in literature**

**Step 1: EDA and Feature Extraction**

**Step 2: Feature Encoding**

**Step 3: Base-line model and metrics**

**Step 4: Ensembling of Models**

**Step 5: Feature importance and Ranking**

**Step 6: Choosing a Final Model**

**Step 7: Model Deployment**

**7. Limitations**

**8. References**

## **Abstract**

World surface is covered with 75% of water, 97.5% of it is salt water and remaining 2.5% is fresh water. Only 0.3% of fresh water is on surface. Water is the major necessity for our body to deliver the vitamins and nutrients to our body cells. Due to global warming, water pollution and over population water crisis are occurring in many countries of the world. Among the countries Tanzania is also facing water crisis. To use the available water resources precisely and deliver the portable water to the people Using the machine learning/deep learning model optimal solution to be provided to Tanzania government to monitor the water resources and supply portable water to people.

### **1. Problem definition**

#### **1.1 Introduction**

Tanzania is an east African country with a population of 59 million and half of the population has not access to safe drinking water. It has water points of 59k approximately and some of them are functioning, non-functioning and need of some repair.

#### **1.2 Importance of problem**

Water pumps are located in different geographical locations and are funded and maintained by different organizations. Various features are impacting the functionality of water pumps. Manually correlating the functionality of water pumps with the features available is difficult and it won't give any insight for installing new pumps

further. Using the machine learning/Deep learning model on the data provided about water pumps and its features need to predict about water pump functioning, non-functioning and needs of repair. It will give an insight about functionality of water pumps.

### **1.3 Real world impact of this problem**

Using ML/DL model for Predicting about water pumps, resources will be allocated on priority based and it will provide the government a clear approach for installing new pumps. The features which are affecting the water pumps will be focused further. Above things will resolve many issues related water crisis and ensures government to supply portable water to people

## **2. Data set Description.**

### **2.1 Source of dataset**

Tanzania ministry of water, in collaboration with online source data platform taarifa has created and hosted data points about water pumps in Tanzania.

Data set link: <https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/>

### **2.2 Dataset properties and challenges:**

Features in the data set includes categorical and numerical as well. There are missing feature values for data points, filling the missing ones with appropriate values and removing the less correlated features for predicting the model are the main challenges.

#### **Dataset features:**

- **ID:** unique identification number for each water point
- **Amount\_tsh:** total static head available for water pump
- **Date\_recorded:** data recorded date
- **Funder:** financial supporter to water point

- **Gps\_height:** height of the water point from ground
- **Installer:** who installed the water point
- **Longitude:** GPS coordinate of water point
- **Latitude:** GPS coordinate of water point
- **Wpt\_name:** Name of the water point
- **Num\_private:**
- **Basin:** geographic region of water basin
- **Sub\_village:** village name of the water point
- **Region\_code:** unique identification number of region
- **District\_code:** unique identification number of district
- **Lga:** Geographic location of water point
- **Ward:** Geographic location of water point
- **Population:** no.of people using the water point
- **Public\_meeting:** public gathering to discuss available or not
- **Recorder\_by:** organization who recorded the data
- **Scheme\_management:** name of the management who operates water point
- **Scheme\_name:** name of the scheme under which water point operates
- **Permit:** For installing water point permission available or not
- **Construction\_year:** year in which water point constructed
- **Extraction\_type:** type of extraction used in water point
- **Extraction\_type\_group:** type of extraction used in water point
- **Extraction\_type\_class:** type of extraction used in water point
- **Management:** management who operates water point
- **Management\_group:** how the water point is managed
- **Payment:** charges for using the water
- **Payment\_type:** charges for using the water
- **Water\_quality:** quality of water available
- **Quality\_group:** quality of water available
- **Quantity:** amount of water available
- **Quantity\_group:** amount of water available
- **Source:** source of water



- **Source\_type:** source of water
- **Source\_class:** water source is surface water or ground water
- **Waterpoint\_type:** type of accessing water point
- **Waterpoint\_type\_group:** type of accessing water point

**Dataset description:** Dataset has a 59k data points and each data point has 40 features. Given dataset is a imbalanced data, as it has 54% of functional water points, 7% of them are functional and needs repair and 38% of them are non-functional. Due to this imbalanced dataset, model will be biased towards some classes.

### 3. Key metric to optimize

- a) Business metric chosen for this is micro averaged F1 score. Computing the True positive, true negative and false negatives of individual classes, then computing the micro averaged precision and recall. Finally computing the harmonic mean of precision and recall, i.e F1 score.

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

- b) As the data set is imbalanced one, Micro averaged F1 score will give the absolute performance of the model and is less affected by imbalanced dataset
- c) Precision, Recall, ROC/AUC, Log loss, accuracy can also be used as a performance metrics for multi class classification models. But, all of them are affected by imbalanced dataset and will not evaluate the different models in a correct way. They will give same result or biased result for imbalanced dataset
- d) Micro avg. F1 score is less impacted by the imbalanced data and it will measure the performance of the model accurately.
- e) F1 score is less interpretable metric. If the dataset is balanced, ROC/AUC is preferable as it is more interpretable for evaluating the performance of the model
- f) F1 score is preferably used for evaluating the binary and multi classification models.

### 4. Real world challenges and constraints

- a) Feature engineering and selection is one challenging task as it has 40 features and consists of both numerical and categorical features. Converting categorical

features to numerical features for performing the model is required. Domain expert inputs are required for selecting the suitable features as the performance of water point depends on various aspects. Moreover with the new technologies in water pumps and extraction type new data may perform better than the old ones. Then model performance may vary for new data.

- b) Selecting the appropriate features which are affecting the model is essential. Removing the outliers and preprocessing of data for improving the performance of model is needed. The model should be able to give the similar performance on both test and train data.

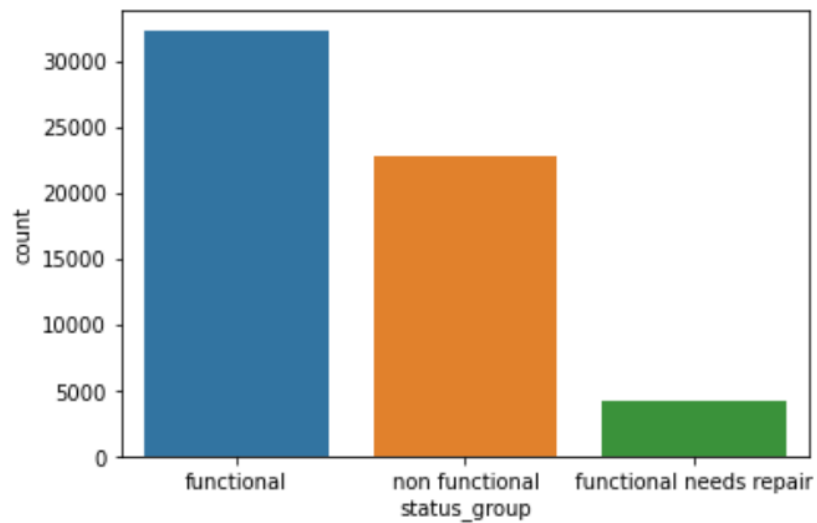
## **5. Similar problems solved in literature**

- a) The given problem of predicting about water point for given dataset as functioning, on-functioning and functioning needs repair is related to the multi class classification model in Machine learning/Deep learning. It is a three class classification problem. We can use One vs Rest or one vs one approach for this model.
- b) For multi class classification various algorithms can be used, such as Neural networks using multiclass perceptron, K-nearest neighbors, Naïve Bayes, Decision trees, support vector machines and random forest. After performing preprocessing of data, feature engineering and feature selection multi class classification algorithms to be applied on the data. Based on the performance evaluation metric for various algorithms for train and test data, appropriate algorithm will be selected for the model.

## STEP 1: EDA and Feature Extraction

### 1. Data-set level and output-variable analysis:

Given dataset have 59400 data points and 40 features. Among those features 10 are numerical features and 30 are categorical features. These data points have to classify into three categories as functional, non-functional and functional needs repair.



From the above histogram, we can conclude that given dataset is highly imbalanced as it has functional:32259, non-functional:22824, functional needs repair: 4317.

For numerical features, using correlation function formed the correlation matrix.

```
train_data.corr()
```

	id	amount_tsh	gps_height	longitude	latitude	num_private	region_code	district_code	population	construction_year
id	1.000000	-0.005321	-0.004692	-0.001348	0.001718	-0.002629	-0.003028	-0.003044	-0.002813	-0.002082
amount_tsh	-0.005321	1.000000	0.076650	0.022134	-0.052670	0.002944	-0.026813	-0.023599	0.016288	0.067915
gps_height	-0.004692	0.076650	1.000000	0.149155	-0.035751	0.007237	-0.183521	-0.171233	0.135003	0.658727
longitude	-0.001348	0.022134	0.149155	1.000000	-0.425802	0.023873	0.034197	0.151398	0.086590	0.396732
latitude	0.001718	-0.052670	-0.035751	-0.425802	1.000000	0.006837	-0.221018	-0.201020	-0.022152	-0.245278
num_private	-0.002629	0.002944	0.007237	0.023873	0.006837	1.000000	-0.020377	-0.004478	0.003818	0.026056
region_code	-0.003028	-0.026813	-0.183521	0.034197	-0.221018	-0.020377	1.000000	0.678602	0.094088	0.031724
district_code	-0.003044	-0.023599	-0.171233	0.151398	-0.201020	-0.004478	0.678602	1.000000	0.061831	0.048315
population	-0.002813	0.016288	0.135003	0.086590	-0.022152	0.003818	0.094088	0.061831	1.000000	0.260910
construction_year	-0.002082	0.067915	0.658727	0.396732	-0.245278	0.026056	0.031724	0.048315	0.260910	1.000000

For numerical features there is no strong correlation between the input variables from the above result. There is a strong correlation of 65.9% between gps\_height and construction\_year. But, actually there shouldn't be relation between those features and not considering those correlations for selecting the features. Remaining categorical features has to be converting into numerical features and observe the correlations among them.

### Missing values in data:

From the given dataset below are the no.of missing values are observed in Funder,Installer,SubVillage,public\_meeting,Scheme\_name,Scheme\_management,permit columns.

funder	3635
installer	3655
subvillage	371
public_meeting	3334
scheme_management	3877
scheme_name	28166
permit	3056

We can remove the rows of missing values, but, it will cause the loss of data. For categorical features replacing the missing values with frequent label and for numerical features, replacing the missing values with median will be appropriate.

### **Data pre-processing:**

- a. Using the mode operation, it is showing clearly ID is unique for all data points. Dropping the ID feature from dataset as it is not useful for classification.
- b. Removing the quantity group feature as it is same data as quantity.
- c. From the output of `train_data['recorded_by'].describe()`. It is same for all the datapoints.so,dropping this feature.
- d. In Num\_private feature 98.75% of data is same and its value is 0. Removing this feature from data.
- e. Date recorded feature has no correlation with the classification, as data may be recorded earlier or in delay.
- f. Comparing the feature labels in both the features, below features are similar, some of them are mentioned with spelling mistakes, upper and lower case letters. Removing the below features based on the above criteria.
  1. Extraction\_type\_group
  2. Installer
  3. Payment\_type
  4. Quality\_group
  5. Source\_type
  6. Waterpoint\_type\_group
  7. Extraction\_type
  8. region
  9. scheme\_management
  - 10.extraction\_type
- g. Region\_code and district\_code has a correlation of 67.8%, as both of them belong to geographical location, removing the region code feature as it has less no.of sub categories.

- h. Waterpoint\_type and extraction\_type\_class has correlation of 65% and same can be checked by comparing the labels in each feature.
- i. 88% of data is sub-categorized as user-group. It is biased towards user-group. So, removing management\_group feature from dataset
- j. 33% of data with amount of water head is 0 and it is functional. Practically without water availability functioning is not possible. So, removing amount\_tsh from data set.
- k. categorcial featrures wpt\_name, basin, lga, ward, sub\_village all will represent about the geographical location of the water point. Among all the features basin has the less no.of sub categories. So, dropping the remaining features from the dataset
- l.

### **Removing Outliers:**

Outliers can be detected by measuring the interquartile range and calculating the lower limit and upper limit for removing the outliers.

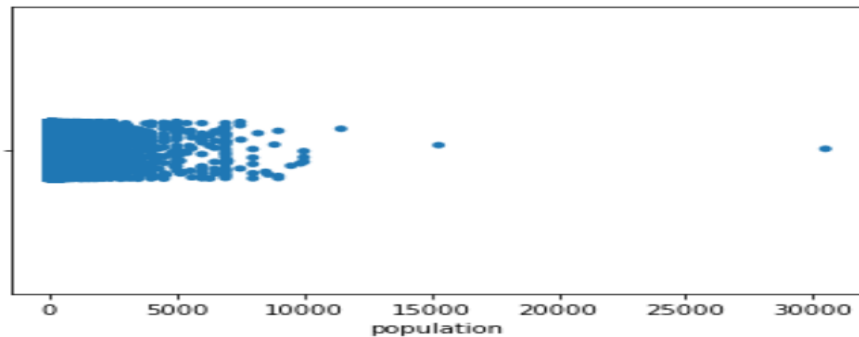
Data points lies in the range  $(Q1 - 1.5 * IQR, Q3 + 1.5 * IQR)$  will be considered for classification and remaining will be removed. From the numerical features population has only the outliers. But, in practical population will be very high in some region and less in other regions. Removing outliers will cause loss of data in other features as well. So, not removing the outliers

### **Univariate Analysis:**

In Univariate analysis each feature is analyzed one at a time. Below are the uni-variate analysis methods.

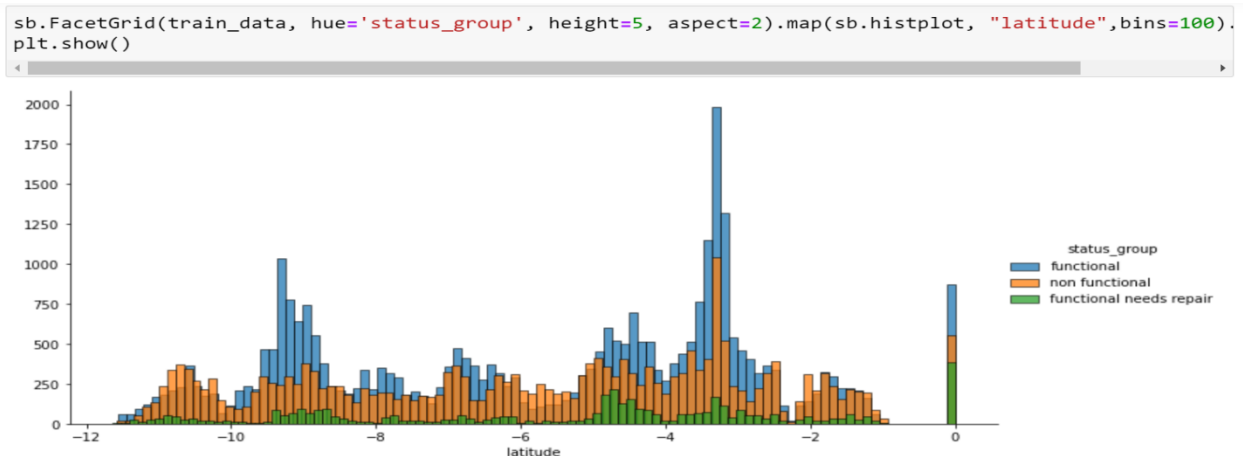
- a. 1D scatter plots
- b. Histogram
- c. .PDF
- d. CDF
- e. Boxplot
- f. Violin plot

a. 1D scatter plots:

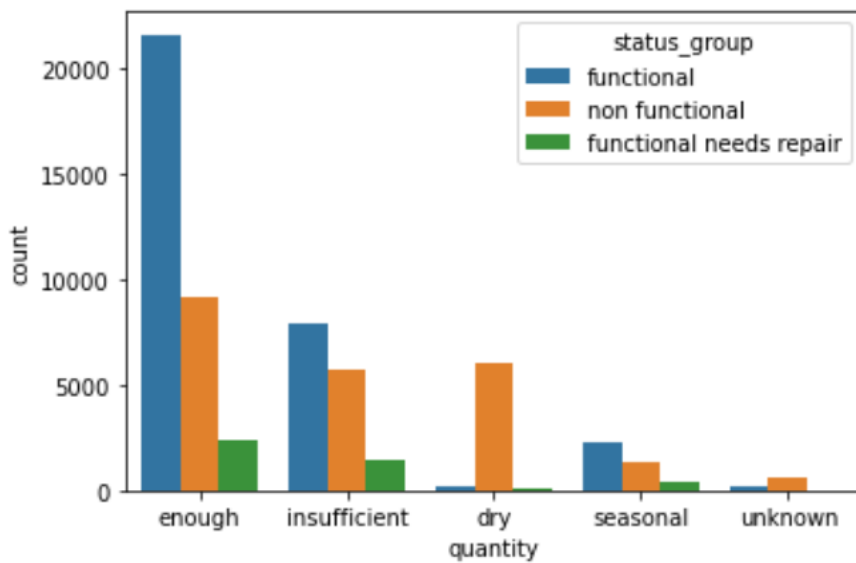


Scatter plots can able to represent the large quantity of data. In the above scatter plot, more data points are overlapped in the range of 0 to 7500 for population feature. We cannot draw any results using 1D scatter plot due to overlapping

b.Histogram

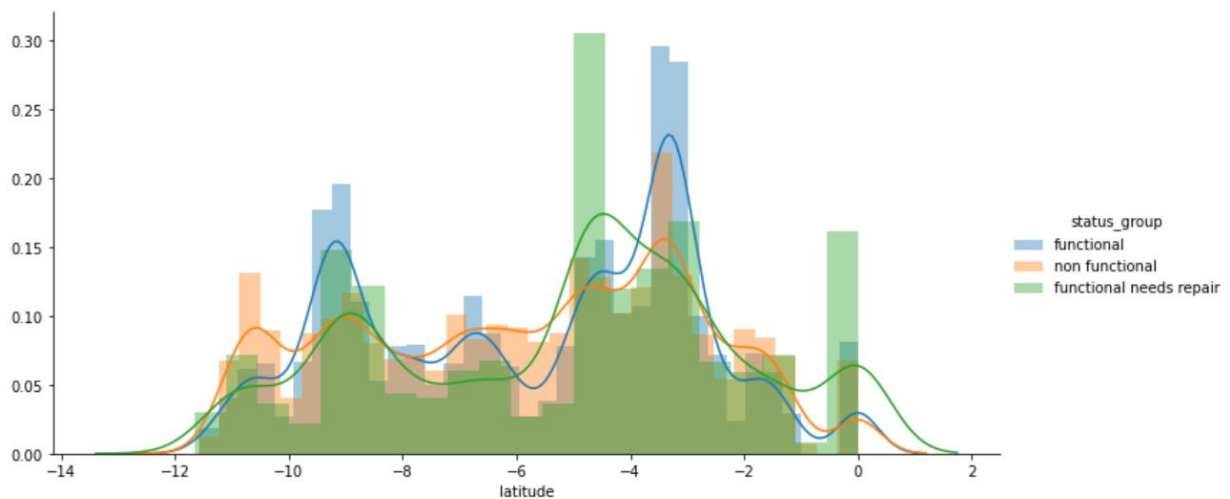


From the above histogram we can visualize that latitude values in between -4 and -2 has more points .Using the histogram, we can able to represent large amount of data in bins. Highest and lowest frequency intervals can be analyzed using histogram. Shape and spread of the continuous data will be visualized. But, we cannot able to visualize the density of the points in intervals and outliers' detection is not possible with histogram.



From the above plot we can observe that, water with enough quantity are functioning well and dry quantity water are non-functional.

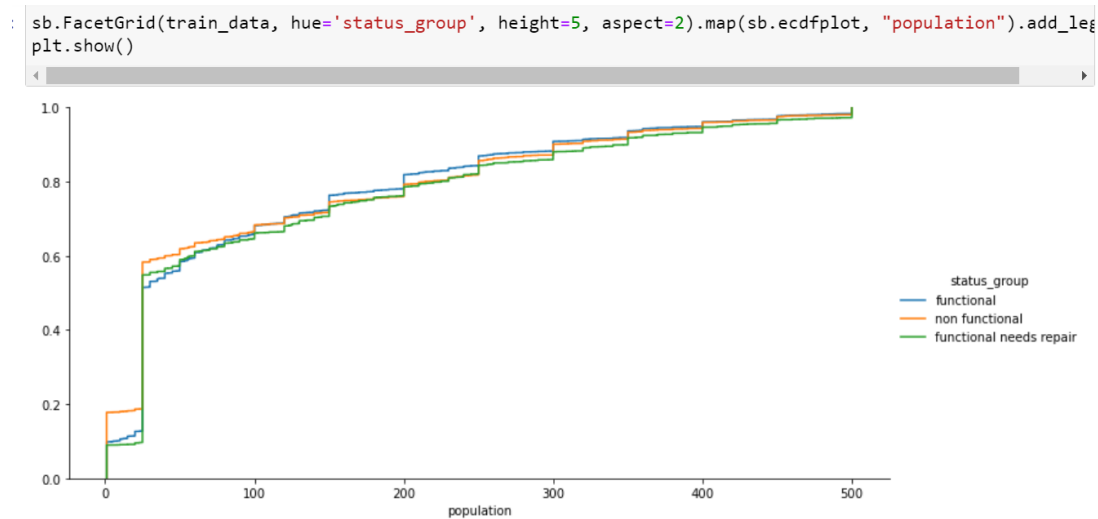
### c.PDF



PDF will give the probability of the points falling in intervals. It is better version of histogram. Using the PDF we cannot able to find out the outliers present in the data. From the above plot, more % of latitude points, i.e 32% points are lies in the interval of -5 to -4. <1% points are lies in the interval of -24 to -12

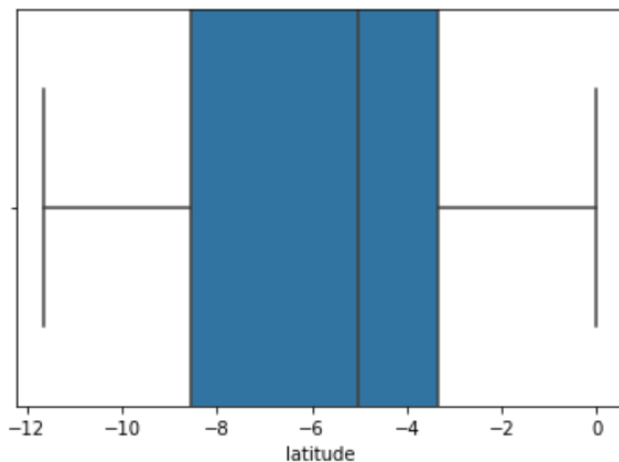


#### d.CDF



Using the CDF plot, data points can be visualized along with their range of points falling in particular interval. From the above plot, we can able to see the most % of points interval. But, we cannot able to point out the outliers. From the CDF plot, 60% of the water points are with a population of 40.

#### e. box plot



Irrespective of the distribution of the data, box plot will gives the visualization of data into 5 statistical points as minimum point, lower quartile, median, upper quintile and maximum point. Outliers can be detected using box plot by multiplying with 1.5 times of IQR and can be eliminated from the dataset. Box plot cannot able to represent the

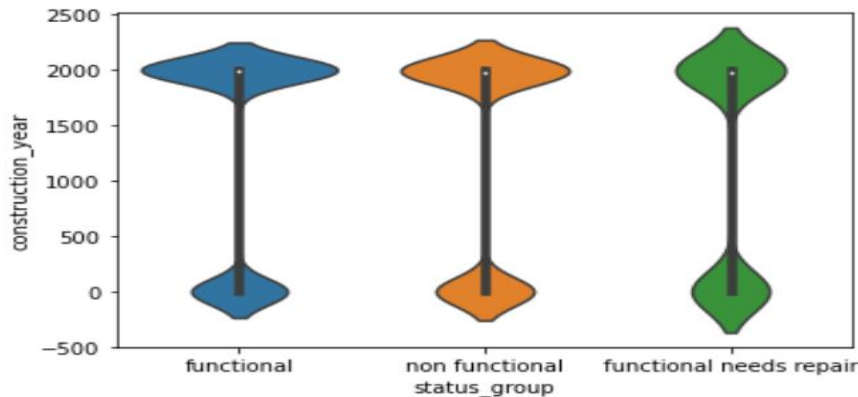
dataset distribution. From the box plot, min point =1960, max point=2013,IQR is 20 and Q1=1984,Q3=2004

Lower limit for outlier is = $1984 - 1.5 \times 20 = 1954$

Upper limit for outlier is = $2004 + 1.5 \times 20 = 2034$

From the above boxplot, outliers are not present in the dataset.

f. violin plot

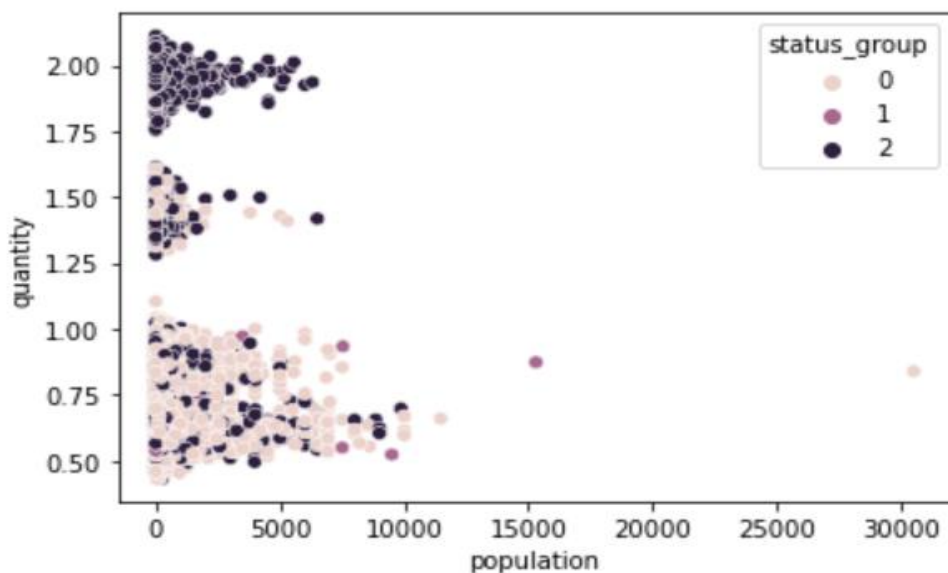


Along with the statistical points and outliers detection, violin plot also provides the distribution of data points in the dataset. From the violin plot, it is a non-Gaussian distribution.

Multivariate Feature analysis:

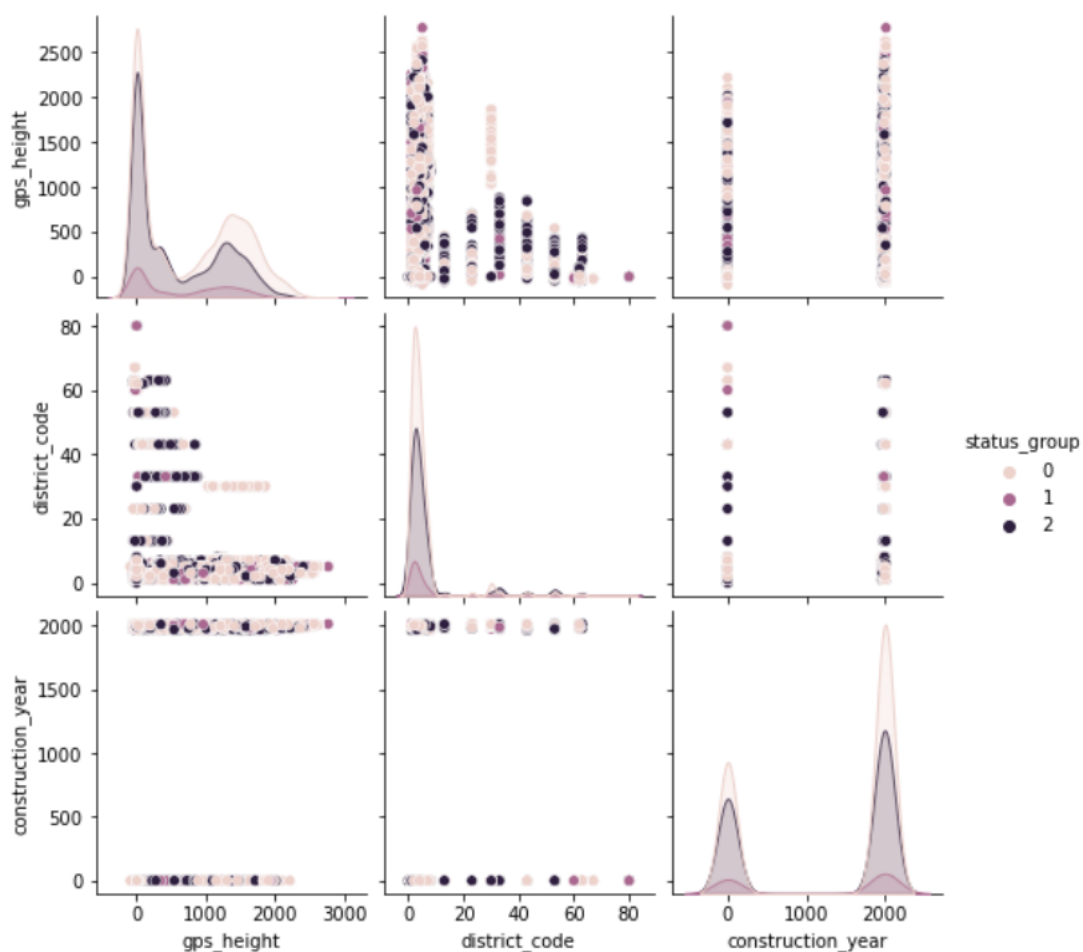
In multivariate feature analysis two or more features will be analyzed at a time.

1. Scatter plot



Data set is large, visualizing the features and feature selection with 2D scatter plot is difficult.

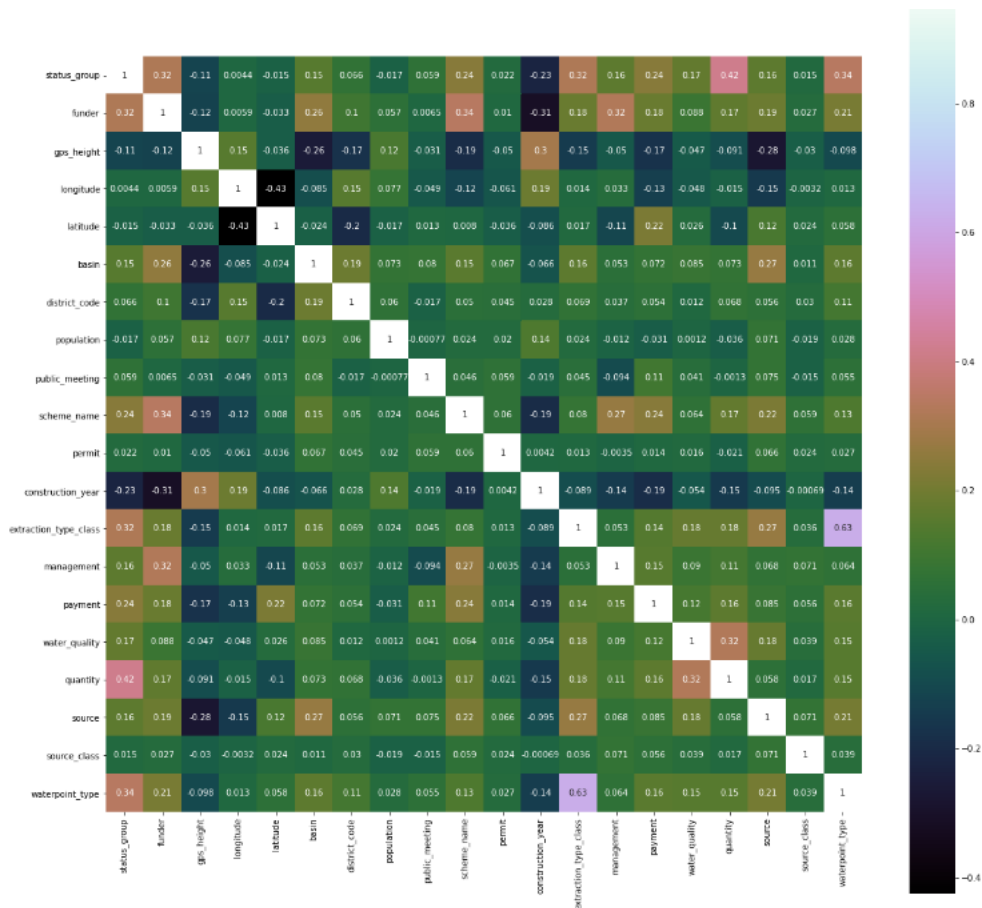
## 2. Pair plots



From the above pair plot with different numerical features, most of the pair of features is overlapping due to the more data points present in it. Increasing the scale of the plot and visualizing is difficult using pair plot. So, for feature selection pair plot is not useful if the data points are more.

### 3. Correlation matrix

Before performing the correlation matrix on the features, all the categorical features are converted into numerical features and correlation coefficient between all the pair of features is calculated and its range is from 0 to 1. Correlation coefficient near to 1 between the features has more correlation and suitable feature will be selected accordingly.



From the correlation heat map there is a correlation between gps\_height and construction\_year. But, practically there is no relation between them. so, not considering this relation

### Step 2: Feature Encoding

**Feature encoding:** Converting the categorical features to numerical features.

**Label Encoding:**

Assigning the values to label features based on their rank order. Firstly converting Target labels into numerical for modelling purpose.

1. Status\_group

### **Target encoding:**

Replacing a categorical value with the mean of the target variable. It avoids the higher dimensionality and sparsity in the features, which occurs due to one-hot encoding. But, due to target encoding data leakage and over fitting of model will occurs. To avoid this, Gaussian noise will be added to the encoded columns.

Features encoded with Target Encoding are:

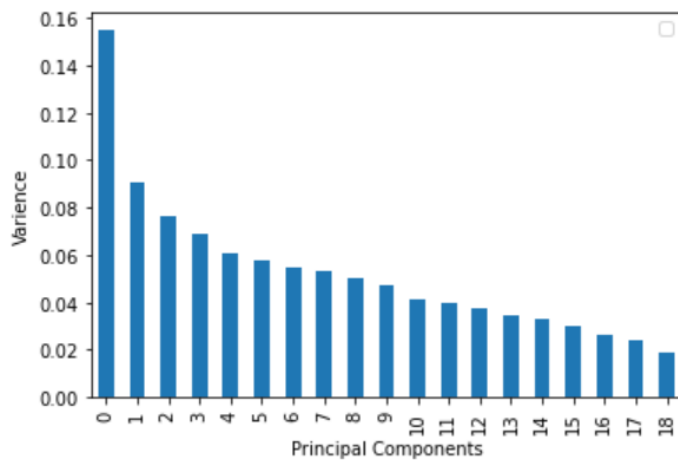
1. Permit
2. Public\_meeting
3. Water\_quality
4. quantity
5. Waterpoint\_type
6. Source
7. Source\_class
8. Management
9. Management\_group
10. Extraction\_type\_class
11. Basin
12. funder
13. scheme\_name

After removing the features and feature encoding, data set shape changed from (59400, 40) to (59400,19)

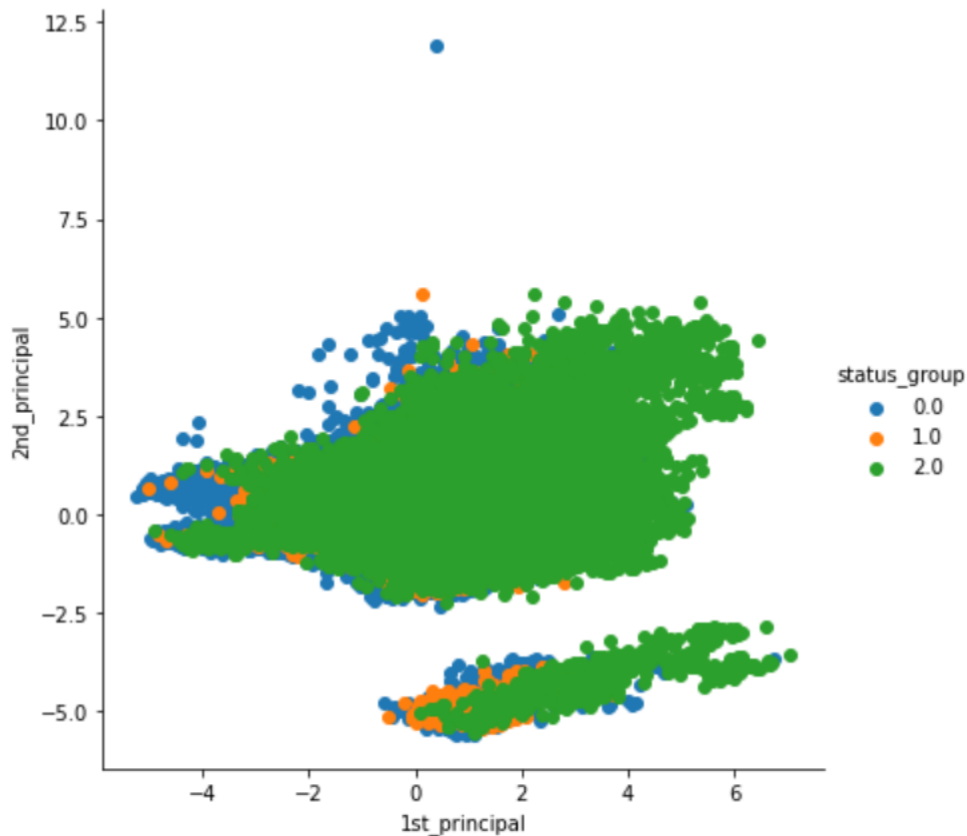
### **High dimensional data visualization:**

## 1. Principal component analysis:

Before performing the principal component analysis all the features in the dataset standardized to bring the all features with zero mean and unit standard deviation. Principal components and its variance to be plotted for selecting the no.of principal components holding the maximum variance.



From the above plot principal components vs variance, out of 19 principal components 16 are holding the 91% of variance of the data and some are having less variance of the data. Considering two principal components for visualizing the data



#### Advantages of PCA:

- a) it removes the correlated features as all the principal components are independent to each other
- b) Significantly improves model performance of high dimensional data by reducing the no.of features
- c) It reduces over fitting by reducing the no.of features
- d) It improves data visualization by converting from high dimensional to low dimensional using principal components.
- e) It preserves global structure

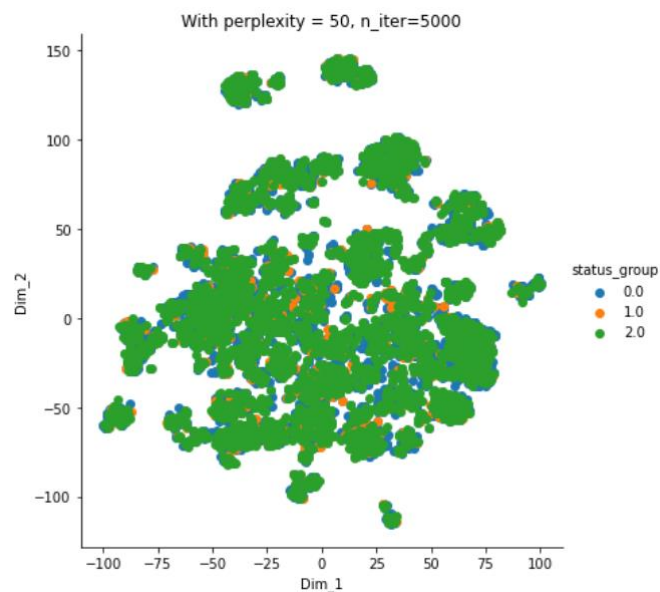
#### Disadvantages of PCA:

- a) principal components are less interpretable
- b) information loss due to selecting the principal components with max variance

- c) It doesn't preserve local structure

## 2. T-SNE

For performing high dimensional data visualization using T-SNE, standardization of data is required. It computes pairwise conditional probabilities for each data point.



Advantages of T-SNE:

- a) It handles non-linear data efficiently
- b) It preserves both local and global structures

Disadvantages of T-SNE:

- a) Computationally complex and times complexity is more
- b) It is Non-deterministic as different runs with same hyper parameters may produce different results
- c) It requires hyper parameter tuning and produces noisy patterns

## Step 4: Base-line model and metrics



### Base-Line Model:

A baseline model is essentially a simple model that acts as a reference in a machine learning project. Its main function is to contextualize the results of trained models.

Baseline models usually lack complexity and may have little predictive power.

Base line model chosen for the three class classification is K-Nearest Neighbours. KNN also called K- nearest neighbor is a supervised machine learning algorithm that can be used for classification and regression problems. K nearest neighbor is one of the simplest algorithms to learn. K nearest neighbor is non-parametric. It does not make any assumptions for underlying data assumptions.

### Evaluation Metric:

From the dataset given, it is evident that it is an imbalanced dataset. So, micro averaged F1score is chosen as the performance metric. Micro averaging computes a global average F1 score by counting the sums of the True Positives (TP), False Negatives (FN), and False Positives (FP). Sum the respective TP, FP, and FN values across all classes and then put them into the F1 equation to get micro F1 score.

## 1. Classification Models used for Modeling are:

### a. K-Nearest Neighbours

The k-nearest neighbors algorithms a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

#### Advantages:

- I. KNN modeling does not include training period as the data itself is a model which will be the reference for future prediction
- II. It is very time efficient in term of improvising for a random modeling on the available data.

- III. KNN is very easy to implement as the only thing to be calculated is the distance between different points on the basis of data of different features and this distance can easily be calculated using distance formula such as- Euclidian or Manhattan

Disadvantages:

- i. Does not work well with large dataset as calculating distances between each data instance would be very costly.
- ii. Does not work well with high dimensionality as this will complicate the distance calculating process to calculate distance for each dimension.
- iii. Sensitive to noisy and missing data
- iv. Data in the entire dimension should be scaled (normalized and standardized) properly.

Applications:

Text mining, Agriculture, Finance, Medical, Facial recognition, Recommendation systems (Amazon, Hulu, Netflix, etc)

## **b. Naive bayes**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

Advantages:

- i. It is simple and easy to implement and It doesn't require as much training data
- ii. It handles both continuous and discrete data
- iii. It is highly scalable with the number of predictors and data points
- iv. It is fast and can be used to make real-time predictions

Disadvantages:

- I. If test data set has a categorical variable of a category that wasn't present in the training data set, the Naive Bayes model will assign it zero probability.
- II. This algorithm is also notorious as a lousy estimator
- III. It assumes that all the features are independent. Which is not possible in real life

Applications:

Real time Prediction, Multi class Prediction, Text classification/ Spam Filtering/ Sentiment Analysis, Recommendation System

### **c. XGboost**

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

Advantages:

- I. It is Highly Flexible and uses the power of parallel processing.
- II. It is faster than Gradient Boosting and supports regularization.
- III. It is designed to handle missing data with its in-build features.
- IV. The user can run a cross-validation after each iteration.

Disadvantages:

- I. Difficult interpretation , visualization tough
- II. Over fitting possible if parameters not tuned properly.
- III. Harder to tune as there are too many hyper parameters.

Applications:

Anomaly detection, quality prediction

### **d. Random Forest:**

It builds decision trees on different samples and takes their majority vote for classification

Advantages:

- I. It reduces over fitting problem in decision trees and also reduces the variance and therefore improves the accuracy.
- II. Random Forest can be used to solve both classifications as well as regression problems.
- III. Random Forest can automatically handle missing values and no feature scaling required
- IV. Handles non-linear parameters efficiently and automatically handle missing values.
- V. Random Forest is usually robust to outliers and can handle them automatically.
- VI. Random Forest algorithm is very stable and less impacted by noise.

Disadvantages:

- I. Complexity
- II. Longer Training period

Applications:

Credit card default, fraud customer/not, easy to identify patient's disease or not, recommendation system for ecommerce sites.

## e. Decision Tree

Decision Trees are a non-parametric supervised learning method used for classification

Advantages:

- I. Normalization or scaling of data not needed and handling missing values
- II. No considerable impact of missing values.
- III. Easy visualization Automatic Feature selection
- IV. Irrelevant features won't affect decision trees.

Disadvantages:

- i. Prone to over fitting.
- ii. Sensitive to data. If data changes slightly, the outcomes can change to a very large extent.
- iii. Higher time required to train decision trees.

Applications:

Identifying buyers for products, prediction of likelihood of default,

### 2. Sampling of dataset:

Dataset is divided randomly into Train data and train data with a ratio of 80:20. After dividing the dataset into two parts. Same can be used for all the classification models. So that, all models will work on the same splitted data and performance of the models can be compared accordingly.

### 3. K-fold cross validation

In K-fold cross validation method, total dataset will be used in both training the model and testing the model as well. Based on choosing the K value, data set is divided into k times and each time some portion will be used for testing and remaining will be used for training. Optimum value k=5 is chosen for cross validation

	Model	Train data_F1 score	Test data_F1score	Crossvalid_mean	Crossvalid_std	Grid_bestscore
0	K Nearest Neighbours	0.75	0.68	0.67	0.004	0.67
1	Naive Bayes	0.70	0.70	0.70	0.004	0.70
2	XGboost	0.81	0.77	0.77	0.002	0.78
3	Randomforest	0.87	0.74	0.74	0.005	0.75
4	Decision Tree	0.70	0.70	0.70	0.004	0.74

From the above table cross validation mean and standard deviation shows that, classification models chosen are performing the same level after changing the train and test data values 5 different combinations also. XGboost is having the highest mean value and less standard deviation. Random forest model has more standard deviation compared to remaining models.

#### 4. Hyper-parameters tuning methods:

Grid search:

For modeling on the given dataset for classification, hyper parameters are initialized. But, for getting the optimum hyper parameters for improving the model performance grid search is used.

	Model	Train data_F1 score	Test data_F1score	Crossvalid_mean	Crossvalid_std	Grid_bestscore
0	K Nearest Neighbours	0.75	0.68	0.67	0.004	0.67
1	Naive Bayes	0.70	0.70	0.70	0.004	0.70
2	XGboost	0.81	0.77	0.77	0.002	0.78
3	Randomforest	0.87	0.74	0.74	0.005	0.75
4	Decision Tree	0.70	0.70	0.70	0.004	0.74

After using the grid search for best hyper parameters except for Decision Tree model, remaining models F1\_score is same and it shows, best hyper parameters are already initialized in the model. Whereas, after getting the best hyper parameters for Decision Tree and using the same, model F1\_score increased from 70 % to 74%.

```
#Finding the best hyperparameters for improving the performance of the model using gridsearch
grid_p = {"criterion":['gini','entropy','log_loss'], "random_state":[50,100,200],
          "max_depth":range(1,10), "min_samples_leaf":range(1,5)}

grid_search = GridSearchCV(DT, grid_p, n_jobs=-1, cv=5, scoring='f1_micro')
grid_search.fit(X_train, y_train)
best=grid_search.best_score_
grid_search.best_params_
```

```
{'criterion': 'gini',
 'max_depth': 9,
 'min_samples_leaf': 1,
 'random_state': 200}
```

Best hyperparameters for Decision Tree using gridsearch are:{'criterion': 'gini', 'max\_depth': 9,'min\_samples\_leaf': 1,'random\_state': 200}

#### 5. Error analysis on models

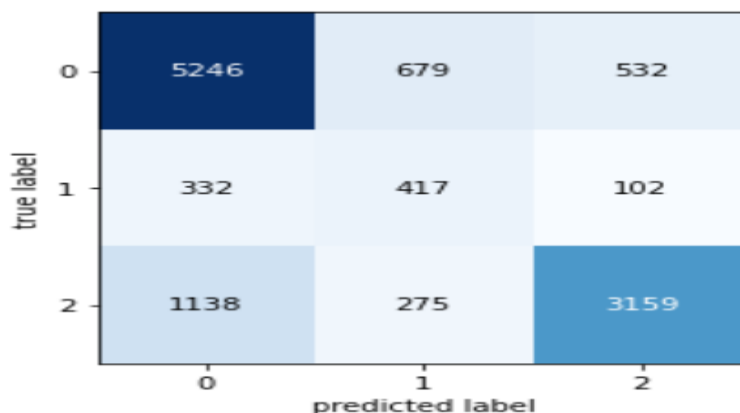
	Model	Train data_F1 score	Test data_F1score	Crossvalid_mean	Crossvalid_std	Grid_bestscore
0	K Nearest Neighbours	0.75	0.68	0.67	0.004	0.67
1	Naive Bayes	0.70	0.70	0.70	0.004	0.70
2	XGboost	0.81	0.77	0.77	0.002	0.78
3	Randomforest	0.87	0.74	0.74	0.005	0.75
4	Decision Tree	0.70	0.70	0.70	0.004	0.74

Increasing order of classification models based on evaluation metric F1\_score

Naïve Bayes>Decision Tree>K nearest Neighbours>XGBoost>Random Forest

Compared to baseline model K nearest neighbors (F1 score= 0.75) XGBoost and Random forest model performance is more. Random Forest has the highest micro F1 score 0.87 on train data and XGBoost has highest F1\_score 0.77 on the test data.

Confusion Matrix of Random Forest



Dataset has a shape of 59400 data points with 40 features. After EDA and preprocessing no.of features are reduced to 90. In sampling the dataset for modeling into train and test data, it is splitted into 80:20 ratio. Now, test data has 11880 points. Adding all the values in confusion matrix will be equivalent to test data size.

Observations from the confusion matrices of all the classification models:

- Models with more diagonal elements sum predicts the better compared to remaining and it classification is better. XGBoost has highest diagonal elements sum
- Decision Tree model unable to classify the test data into class 1 classification and it is biased towards class 0 and class 2
- Random Forest model correctly classified more no.of class 2 data points compared to remaining models.

- Random Forest model correctly classified more no.of class 1 data points compared to remaining models.

#### **Step 4: Ensembling of Models**

Ensemble is a general approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.

Ensemble methods used for classification are:

##### **a. Max. Voting**

It is mainly used for classification problems. The method consists of building multiple models independently and getting their individual output called 'vote'.

The class with maximum votes is returned as output.

##### **b. Extreme Gradient Boosting(XGB)**

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

Advantages:

- I. It is Highly Flexible and uses the power of parallel processing.
- II. It is faster than Gradient Boosting and supports regularization.
- III. It is designed to handle missing data with its in-build features.
- IV. The user can run a cross-validation after each iteration.

Disadvantages:

- I. Difficult interpretation, visualization tough
- II. Over fitting possible if parameters not tuned properly.
- III. Harder to tune as there are too many hyper parameters.

#### **4. Gradient Boosting**

Advantages:

- I. train faster especially on larger datasets,
- II. most of them provide support handling categorical features,
- III. Some of them handle missing values natively.

Disadvantages:

- i. Prone to over fitting.
- ii. models can be computationally expensive
- iii. Hard to interpret the final models.

## 5. Light GBM

Advantages:

- I. Faster training speed and higher efficiency.
- II. Lower memory usage
- III. Better accuracy than any other boosting algorithm and compatibility with Large Datasets

Disadvantages of Light GBM

- i. Over fitting: Light GBM split the tree leaf-wise which can lead to over fitting as it produces much complex trees.
- ii. Compatibility with Datasets: Light GBM is sensitive to over fitting and thus can easily overfit small data.

## **Stacking:**

Stacking is an ensemble learning technique to combine multiple classification models using meta-classifier. The individual classification models are trained based on the complete training set. The meta-classifier is fitted based on the outputs of the individual classification models in the ensemble. The meta-classifier can either be trained on the predicted class labels or probabilities from the ensemble.

Classification models chosen are KNN, Decision Tree, Random Forest and LGBM. Four different combinations will be performed on stacking by considering each time each model as a Meta classifier.



Hyper parameters for stacking classifier are choosing the no.of base classifiers and choosing the Meta classifier. We have chosen 3 base classifiers each time for performing stacking.

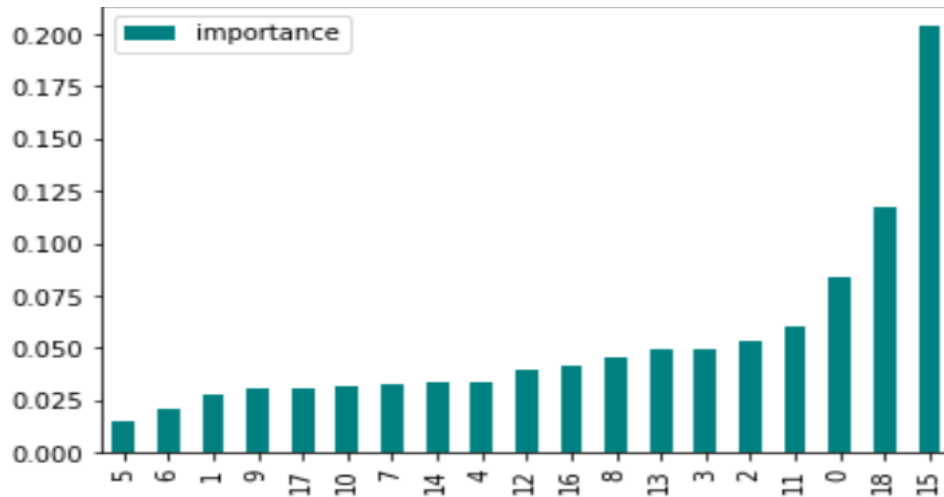
	Classifiers	meta classifier	Train data_F1 score	Test data_F1score
0	KNN DT RF	LGBM	0.97	0.77
1	KNN DT LGBM	RF	0.94	0.75
2	KNN LGBM RF	DT	0.94	0.75
3	LGBM DT RF	KNN	0.77	0.68

From the above table we can observe that considering Decision Tree or Random forest as a meta classifier has highest F1\_score.

### Step 5 : Feature importance and Ranking

For improving the performance of the model, most important features are selected based on their variance. Random forest regressor and Principal component analysis is used for selecting the model.

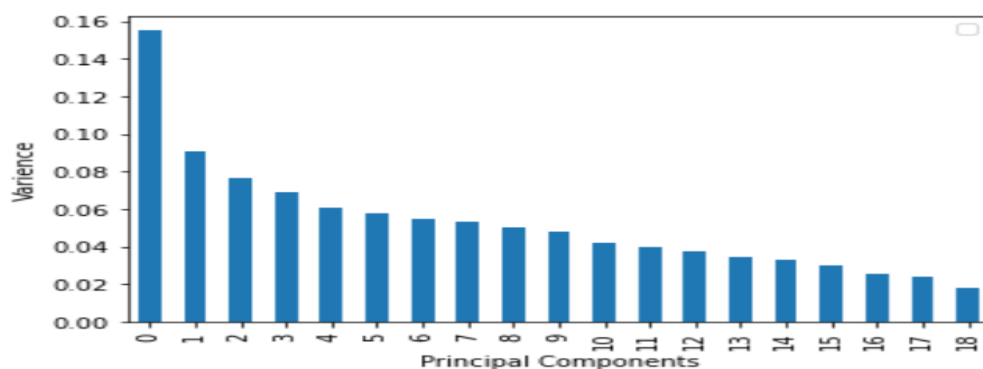
Random Forest Regressor:



Features are plotted in increasing order in the above bar graph with column numbers. Only high importance features are selected for improving the performance. Features with highest variance is selected with feature column numbers and stored in another data frame for modeling.

Using the above method for feature selection, modeling with top 12 important features on base line model K-Nearest neighbors F1\_score increased on train data increased from 0.75 to 0.83 and on test data F1\_score increased from 0.68 to 0.83

Principal Component Analysis:



Using the above method for feature selection, modeling with top 12 principal components on base line model K-Nearest neighbors F1\_score increased on

train data increased from 0.75 to 0.79 and on test data F1\_score increased from 0.68 to 0.74

From the both feature selection methods used on same base line model K nearest neighbours model performance significantly improved with the features selected from random forest regressor.

### Step 6: Choosing a Final Model

	Model	Train data_F1 score	Test data_F1score
0	K Nearest Neighbours	0.75	0.68
1	Randomforest	0.87	0.74
2	Decision Tree	0.76	0.74
3	MAx Voting ensemble	0.82	0.77
4	XGboost	0.91	0.79
5	Gradient boost	0.77	0.76
6	Light GBM	0.79	0.77

For Max voting ensemble method, K nearest neighbors, Decision Tree and Random Forest are used. The F1\_score of max voting ensemble model is better than decision tree and k Nearest Neighbor. The test data F1\_score is significantly improved to 77%.

XGboost ensemble has highest Train data F1\_score 0.91 and test data F1\_score 0.79 compared to all the models.

### Step 7: Model Deployment:

Model deployment is the process of placing a finished machine learning model into a live environment where it can be used for its intended purpose. Models can be deployed in a wide range of environments, and they are often integrated with apps through an API so they can be accessed by end users.

Streamlit is used for model deployment. Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.

For Deploying model in streamlit, Github account is to be created. A repository is created to upload the model.py file and requirements.txt.

A text file with name requirements.txt to be created for accessing the necessary libraries by streamlit after deploying in it. Libraries with versions to be mentioned in requirements.txt file as below.

Libraries used for deployment are:

category-encoders==2.5.0

mlxtend==0.21.0

numpy==1.22.4

plotly==5.3.1

ply==3.11

scikit-image==0.18.1

scikit-learn==1.1.2

scipy==1.8.1

seaborn==0.11.0

statsmodels==0.12.2

xgboost==1.6.2

streamlit==1.13.0

Using streamlit, model is created in such a way that, train data values and train data labels will be accessed directly with the url. Input to the model is given by

Python code for model is downloaded as .py file and uploaded in github repository.

Github repository link: <https://github.com/Bhanupradeep543/musical-doodle>

Streamlit webapp link: <https://bhanupradeep543-musical-doodle-pump-model-hityrs.streamlitapp.com/>

Python code for model is downloaded as .py file and uploaded in github repository.

Any changes in github repository model will replicate the same in streamlit and no need to deploy the model from beginning in streamlit.

Steps for Deploying model in streamlit.

1. Creating account in streamlit
2. Click on the New app.
3. Enter the github repository name, Branch and model.py file name.
4. After clicking on deploy, model will be deployed in streamlit.

[← Back](#)

## Deploy an app

Repository

[Paste GitHub URL](#)

bhanupradeep543/repo

Branch

master

Main file path


streamlit\_app.py

[Advanced settings...](#)

Deploy!

Using streamlit, model is created in such a way that, train data values and train data labels will be accessed directly with the url. Input to the model is given by uploading the CSV file.

Upload CSV

 Drag and drop file here  
Limit 200MB per file • CSV

Browse files

After uploading the file Name Error will be rectified.

Process

After clicking the Process model, Model will work on the input data. Input data parameters will be displayed on the web app.

### Waterpoint dataset parameters

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude
0	50785	0.0000	2013-02-04	Dmdd	1996	DMDD	35.2908
1	51630	0.0000	2013-02-04	Government Of Tanzania	1569	DWE	36.6567
2	17168	0.0000	2013-02-01	<NA>	1567	<NA>	34.7679
3	45559	0.0000	2013-01-22	Finn Water	267	FINN WATER	38.0580
4	49871	500.0000	2013-03-27	Bruder	1260	BRUDER	35.0061
5	52449	0.0000	2013-03-04	Government Of Tanzania	1685	DWE	36.6853
6	24806	0.0000	2011-03-02	Government Of Tanzania	550	Gover	36.3980
7	28965	0.0000	2013-01-25	Finw	234	FinW	39.6074
8	36301	30.0000	2013-01-23	Unicef	584	LGA	39.2630
9	54122	0.0000	2013-03-18	Lawatefuka Water Supply	1083	Lawatefuka w	37.0961

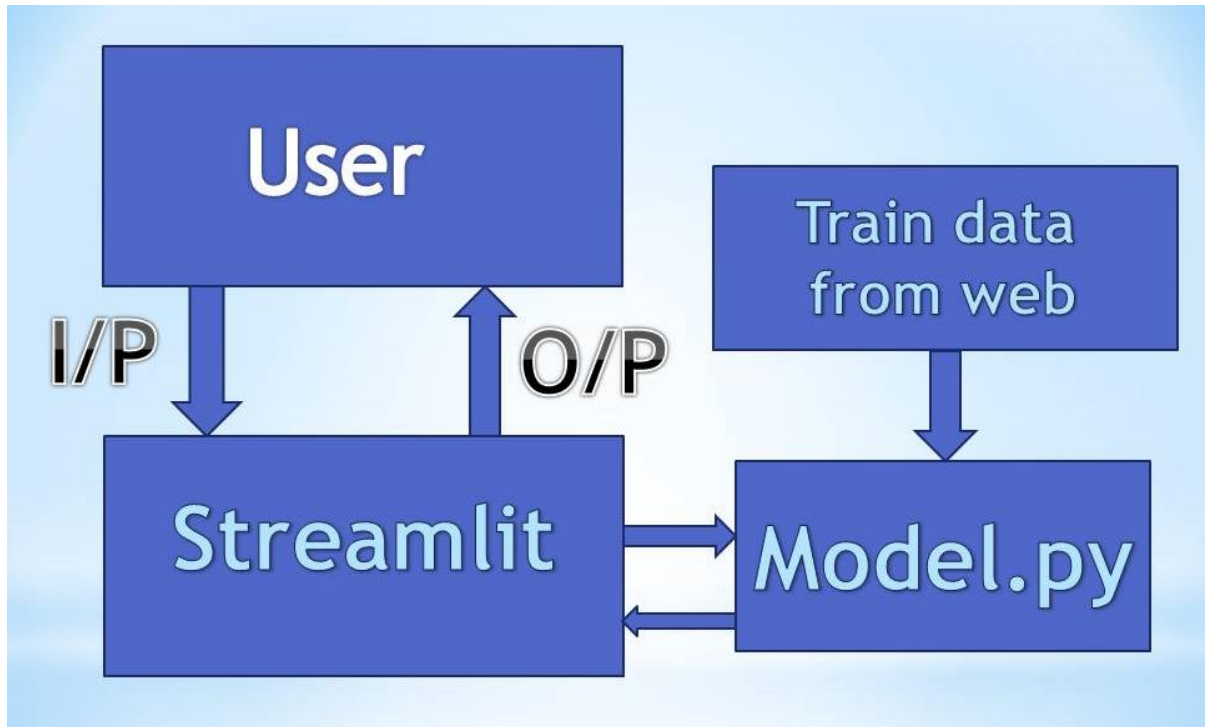
Output will be displayed in the web application and Downloading the output in a single file option also available.

### Prediction

	0
0	functional
1	functional
2	functional
3	functional needs repair
4	functional
5	functional
6	functional
7	functional
8	functional needs repair
9	functional

Download data as CSV

### Architecture diagram



Model.py will be trained by accessing the train data from web URL provided in the code. User will enter the dataset of water points by uploading it in streamlit. Streamlit will process the data through pump\_model.py and after classification the test labels will be displayed and also user can able to download the data from streamlit web app.

Throughput and Latency of the system:

Input data to the model is given up to nearly a data of 200MB file size in CSV format. Due to the amount of input data, time taken for the process will be increased. System is tested with a file size of 4.78MB with 14850 data points. Time taken for processing the above data is nearly 15 min.



## 7. Limitations of the system:

- Input to the system has to be given in CSV file format only
- Input data should have 40 features and each feature has to be given in a desired data type only, otherwise it won't predict the model
- User cannot able to select the features dynamically from the interface. They have to provide the input data to the model in CSV file format only.
- To reduce the latency of the system input data can be given with only 19 features. Which are selected through feature selection method.
- For Prediction selecting various models options are not available.
- To compare the predicted labels from various models option is not available.

## 8. References

<https://towardsdatascience.com/machine-learning-to-help-water-crisis-24f40b628531>

<https://davidcastineira.medium.com/ai-for-social-good-predicting-failure-for-water-pumps-in-tanzania-using-automated-machine-8d75b28fe9c8>

<https://itnext.io/predicting-functional-water-pumps-in-tanzania-using-random-forests-and-logistic-regression-in-ffa04b0617f2>

[https://nycdatascience.com/blog/student-works/linlin\\_cheng\\_proj\\_5/](https://nycdatascience.com/blog/student-works/linlin_cheng_proj_5/)

<https://machinelearningmastery.com/types-of-classification-in-machine-learning/>

<https://www.kaggle.com/code/nkitgupta/evaluation-metrics-for-multi-class-classification/notebook>

[https://en.wikipedia.org/wiki/Multiclass\\_classification](https://en.wikipedia.org/wiki/Multiclass_classification)

<https://vitalflux.com/classification-problems-real-world-examples/>

<https://online.stat.psu.edu/stat508/lesson/1a/1a.5>