Data Science Project

Detection of Automatic Flush System

Wang Shihua
Ramesh Rebba
Himpens Dorian
Felderhoff Noé

CONTENT OF THIS PRESENTATION

- Data Set Description

- Predictions Models

- Discussion

Data Set Description

Our Data Set

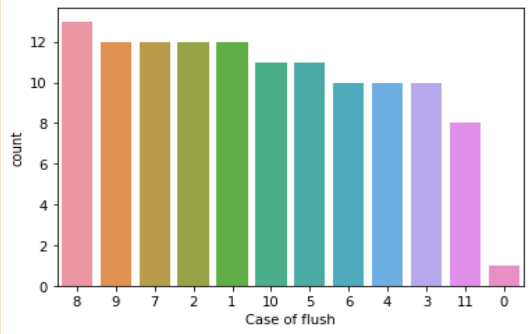
Dependant Variables

- Case of Flush (Y)
- Flush Volume

Independant Variables

- Leds Photoides Values
- Waste Volume

Counts

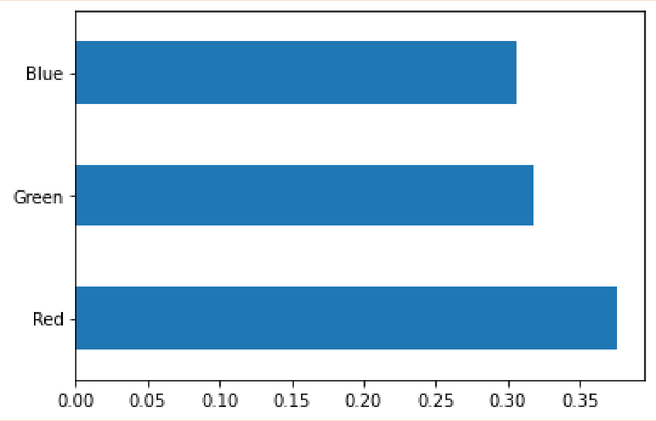


Assumptions

- Colours have the same evolution
- Higher the intensity of light, less volume of water

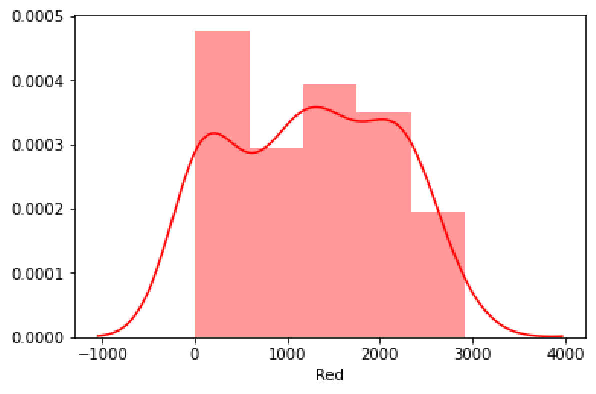
Means

Case of flush	Red	Blue	Green	Sum
0	2240	1795	928	4963
1	2275.6	770.8	754.2	3800.6
2	2230.2	632.5	655.5	3518.2
3	2198	650.5	654.9	3503.4
4	1823	456.8	533.7	2813.6
5	1298.4	456.27	447.6	2202.3
6	1173.2	404	396.4	1973.6
7	970.5	344.75	349.3	1664.6
8	645.9	172.69	210	1028.6
9	514.7	203	193.6	911.3
10	288.4	136.72	122.5	547.6
11	28.2	16.5	11.1	55.9

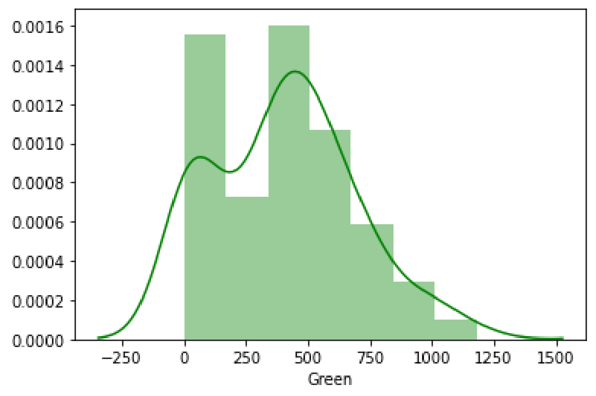


Importance of Features

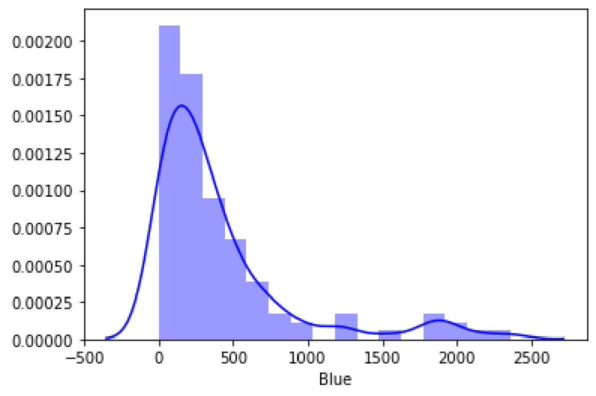
P-value Shapiro (Red):
1.83e-22



P-value Shapiro (Green):
8.23e-25



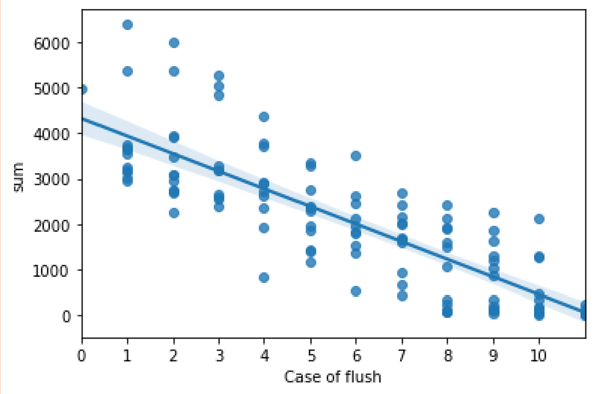
P-value Shapiro (Blue):
3.66e-32



Aggregated Distributions

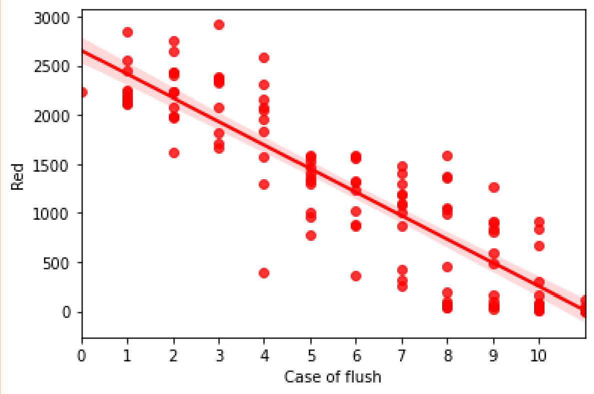
Regplot of all the diodes

Pearson’s correlation coefficient: -0.82



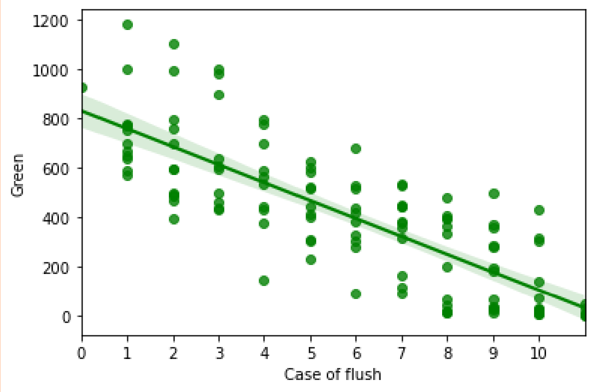
Regplot Red diodes

Pearson’s correlation coefficient: -0.88



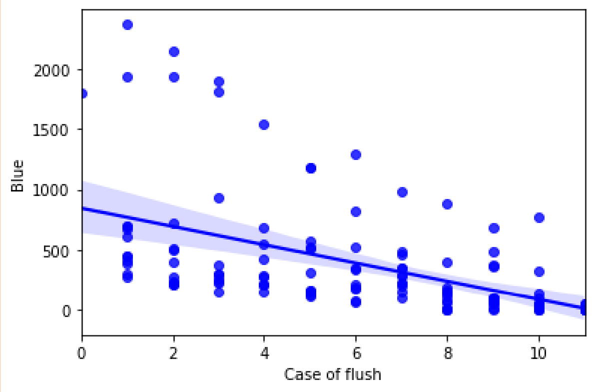
Regplot Green diodes

Pearson’s correlation coefficient: -0.81

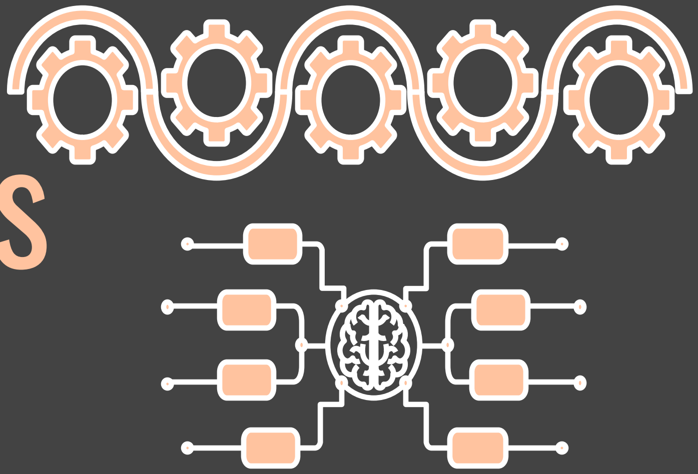


Regplot Blue diodes

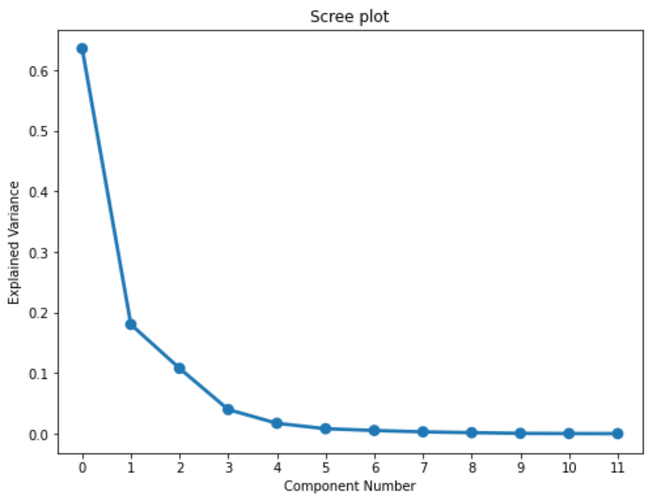
Pearson’s correlation coefficient: -0.49



Predictions Models



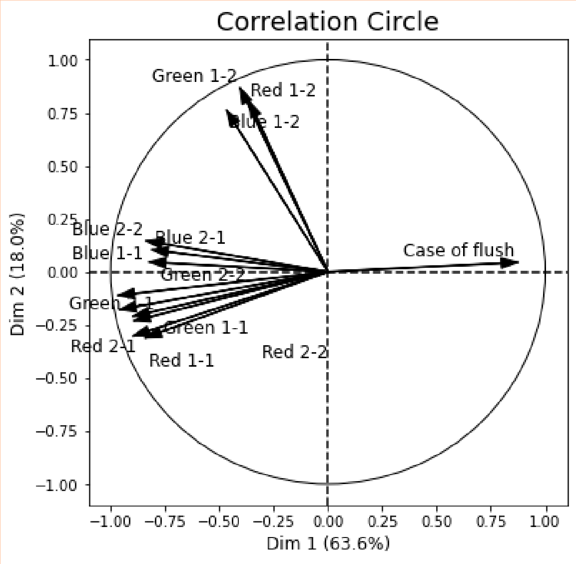
PCA



- 4 components explain 96% of the dataset
- 2 components explain 82% of the dataset
- 82% is enough to draw an interpretable correlation circle

Explained Variance per Component Number (rounded)											
0	1	2	3	4	5	6	7	8	9	10	11
0.64	0.18	0.10	0.04	0.017	0.008	0.005	0.003	0.002	0.0005	0.0002	0.00004
Cumulative Sum of Explained Variance (rounded)											
0.64	**0.82**	**0.92**	**0.96**	0.977	0.985	0.99	0.993	0.995	0.9955	0.9957	100 rounded

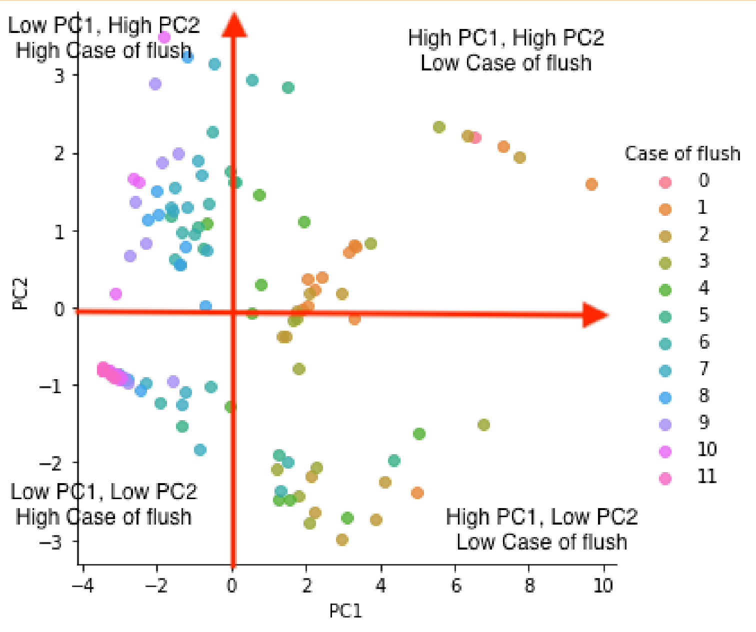
Correlation Circle



DIM 2	<u>Top Left:</u> Test with low intensities of 1-1, 2-1, 2-2 Led-Photodiodes but high intensities of 1-2 Led-Photodiodes and high case of flush	<u>Top Right:</u> Test with high intensities of 1-1, 2-1, 2-2 Led-Photodiodes and high intensities of 1-2 Led-Photodiodes and low case of flush
	<u>Bottom Left:</u> Test with low intensities of 1-1, 2-1, 2-2 Led-Photodiodes and low intensities of 1-2 Led-Photodiodes and high case of flush	<u>Bottom Right:</u> Test with high intensities of 1-1, 2-1, 2-2 Led-Photodiodes but low intensities of 1-2 Led-Photodiodes and low case of flush
	DIM 1	

Does this interpretation fit the observations of our dataset?

Data Recasting Along the PCA

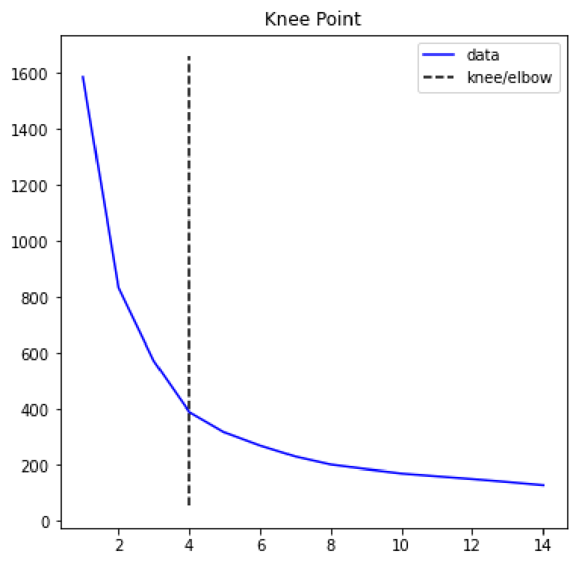


- Low Case of flush: right side of the scatter plot - when PC 1 is positive
- High Case of flush: left side of the scatter plot - when PC 1 is negative
- PC 1: biggest predicator of the Case of flush
- This recasting follows the interpretation of the matrix

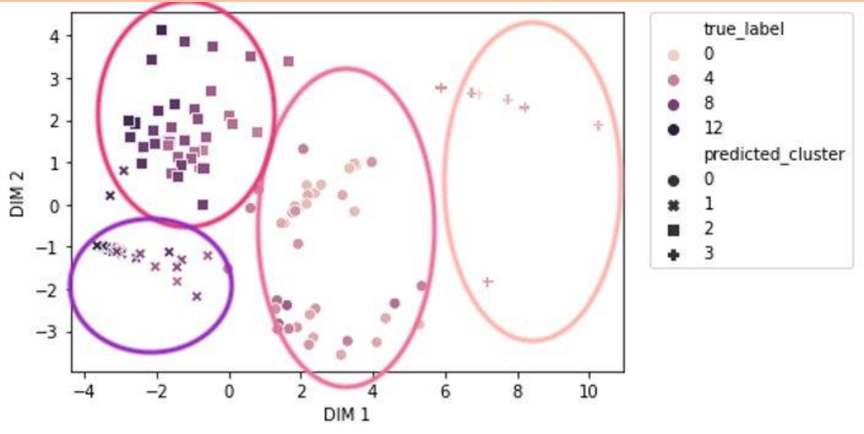
K-Means

Elbow Method

Minimization of the SSE (Sum Squared Error):
4 Clusters (SSE: 390)

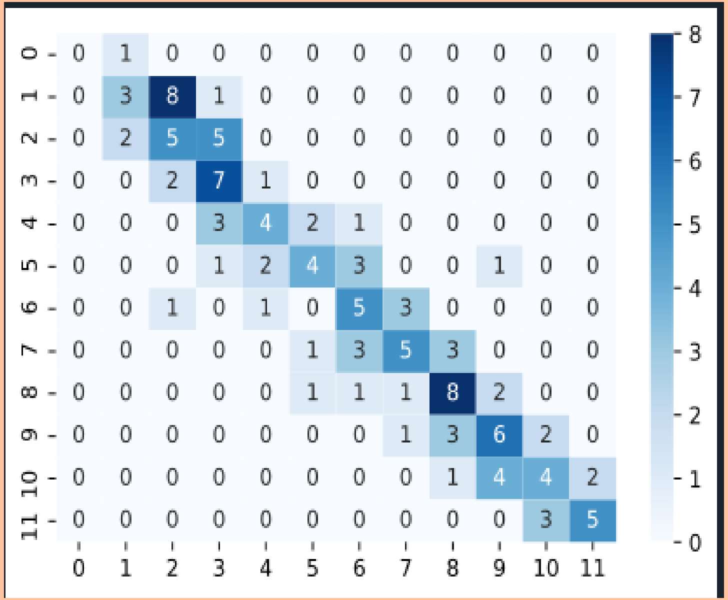


Scatter Plot of the 4 Clusters



- 4 clusters is the sweet point regarding the SSE
- Clusters linked to the intensity of PC 1 and PC 2
- Clusters match well the different observations of the dataset regarding their Case of flush values

OLS



The testing process:

- Treat Y as qualitative data (although it is discrete variables)
- Use Chi-square test to test the fitting optimization
- Null hypothesis: the predicted value and true value is independent
- Find the lowest P-value of Chi-square test
- Draw the heatmap to analyse the fitting result like this

OLS

Initial Assumption:

$Y=aX1^{(½)} +bX2^{(½)}+cX3^{(½)}\ldots +lX12^{(½)}$

$Y=aX1+bX2+cX3+dX4+eX5.... +lX12$

$Y=aX1^2+bX2^2+cX3^2... +lX12^2$

$Y= a*\log(X1+1)+b*\log(X2+1)+c*\log(X3+1)+d*\log(X4+1)+e*\log(X5+1)....+l*\log(X12+1)$

Result of four assumption:

Assumption	R-square	P-value
^(½)	0.895	4.203382217526121e-25
^1	0.871	4.625267489647 67e-22
^2	0.813	2.549091996368603e-13
log	0.870	5.629189891557796e-21

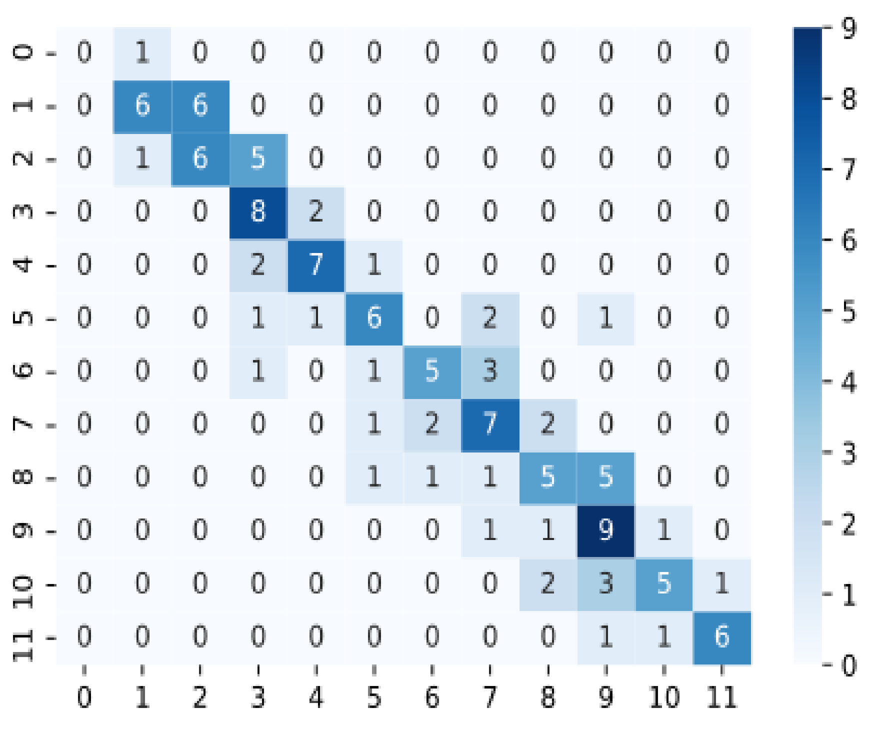
OLS

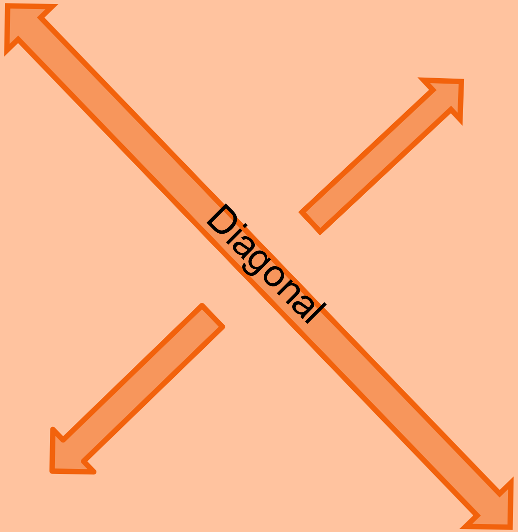
R-square: 0.904
P-value: 3.326956159301724e-42

Parameter	Independent variable	Power
0.0137	Blue LED 1 Photodiode 1	0.6049757
0.0044	Blue LED 1 Photodiode 2	0.80858354
-0.2363	Blue LED 2 Photodiode 1	0.257002
-0072	Blue LED 2 Photodiode 2	0.63890724
0.0061	Green LED 1 Photodiode 1	0.5909667
-0.0909	Green LED 1 Photodiode 2	0.59050965
-0.0542	Green LED 2 Photodiode 1	0.61059362
-0.1216	Green LED 2 Photodiode 2	0.10776922
0.0840	Red LED 1 Photodiode 1	0.33897334
0.2966	Red LED 1 Photodiode 2	0.31753335
0.0058	Red LED 2 Photodiode 1	0.67646697
-0.1574	Red LED 2 Photodiode 2	0.49213317

Y= 0.0137*X1^0.6049757+0.0044*X2*0.80858354-0.2363*X3^0.257002 -0072*X4^0.63890724+0.0061*X5^0.5909667-0.0909*X6^0.59050965 -0.0542*X7*0.61059362 -0.1216*X8^0.10776922 + 0.0840*X9^0.33897334 +0.2966*X10^0.31753335 +0.0058*X11^0.67646697 -0.1574*X12^0.49213317

OLS





Heatmap of predicted flush volume and true flush volume

GLM

We try Gaussian, Poisson, Gamma and Negative Binomial model in GLM family

Parameter	Independent variable	Power
0.2307	F (Blue LED 1 Photodiode 1)	0.17182156
-0.0552	F (Blue LED 1 Photodiode 2)	0.42594121
0.0037	F (Blue LED 2 Photodiode 1)	0.8522537
0.1546	F (Blue LED 2 Photodiode 2)	0.513989
0.2494	F (Green LED 1 Photodiode 1)	0.53666289
0.9801	F (Green LED 1 Photodiode 2)	0.17705916
-0.4866	F (Green LED 2 Photodiode 1)	0.655312
-0.8017	F (Green LED 2 Photodiode 2)	0.12910251
-0.3429	F (Red LED 1 Photodiode 1)	0.42398372
-0.2425	F (Red LED 1 Photodiode 2)	0.34662343
0.0280	F (Red LED 2 Photodiode 1)	0.94068168
-0.0553	F (Red LED 2 Photodiode 2)	0.81186728

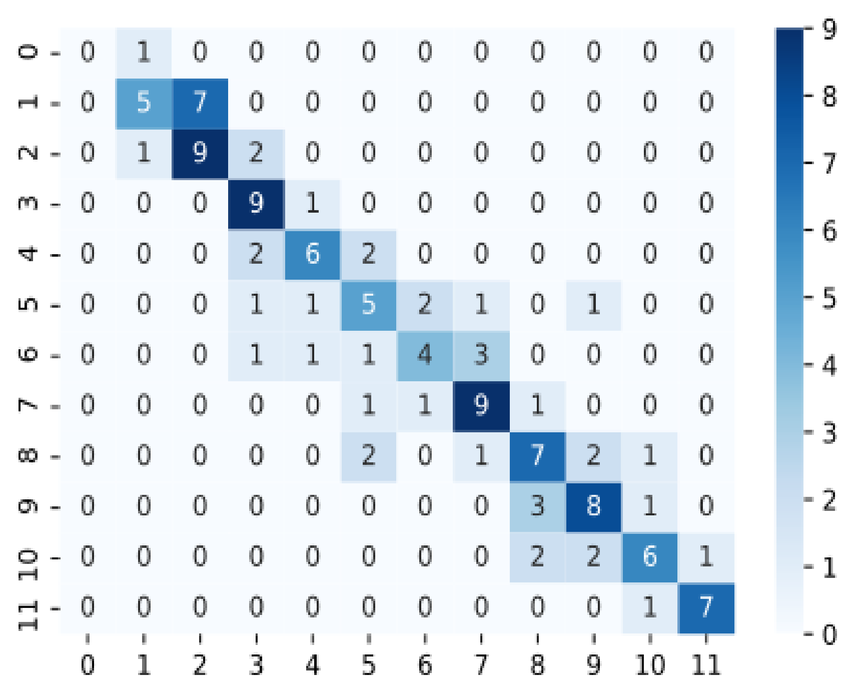
Gaussian model

Accuracy rate : 61.475%

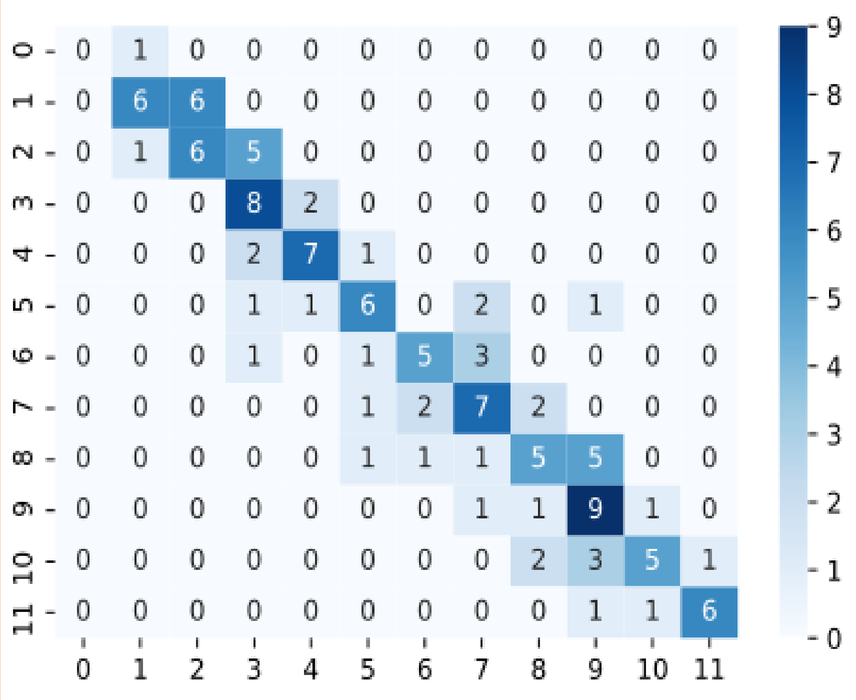
P value: 1.0278368867117633e-48

Y=0.2307*f(x1)^0.17182156-0.0552*f(x2)^0.42594121+
0.0037*f(x3)^0.8522537+ 0.1546*f(x4)^0.513989+
0.2494*f(x5)^0.53666289+ 0.9801*f(x6)^0.17705916 -0.4866*f(x7)^0.655312-
0.8017*f(x8)^0.12910251-0.3429*f(x9)^0.42398372-
0.2425*f(x10)^0.34662343+ 0.0280*f(x11)^0.94068168- 0.0553*f(x12)^0.81186728

GLM



Heatmap of predicted flush volume and true flush volume **in GLM**

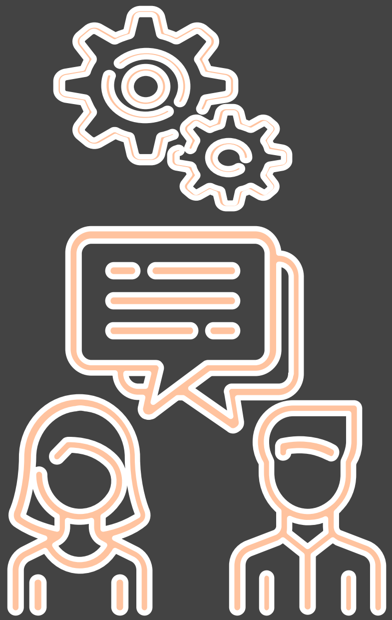


Heatmap of predicted flush volume and true flush volume **in OLS**

GLM

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.71	0.42	0.53	12
2	0.56	0.75	0.64	12
3	0.60	0.90	0.72	10
4	0.67	0.60	0.63	10
5	0.45	0.45	0.45	11
6	0.57	0.40	0.47	10
7	0.64	0.75	0.69	12
8	0.54	0.54	0.54	13
9	0.62	0.67	0.64	12
10	0.67	0.55	0.60	11
11	0.88	0.88	0.88	8
accuracy			0.61	122

Discussion about the Models



Size of the dataset

- Trade-off in terms of train/test split: since our dataset contains 122 observations, if we put 20% as our test size 20%, only 25 predictions will be made. It creates a trade-off between the accuracy of our predictions versus the quantity of predictions made

- Only one test where the outcome of Case of flush is 0: very difficult for our model to predict this outcome (at least 10 tests of the same outcome is needed to make accurate predictions)

- From the following observations, it is concluded that due to the small size of the data set it causes high bias and low variance which results under fitting

- Therefore, it will not perform well under training data set and testing data set

Solution : Test our model on a dataset with more iterations of observations with a Case of flush of 0

Predictions

- When we don't make the correct prediction, we will always prefer to predict the closest value of the true one

- Principal need of the case is to clean the toilet efficiently: predictions putting way too much or not enough water must be avoided
  - For Case of flush 0: we have an accuracy of 0, but we predicted 1, it's the closest value to 0

- Need to be cautious when the predicted value is distant of more than 1 unity of Case of flush
  - For the Case of flush 8, two predictions are 5: it's too distant of the true value, an improved model will need to work on this case

Most of the variables in the given data set have not followed the normal distribution. So non-parametric models like random forests, decision trees can be implemented to get more accuracy if needed.

Solution : Focus more on predictions which are too distant of the true value so as to decrease/erase them.

THANKS!

Does anyone have any questions?