## 

## Presentation

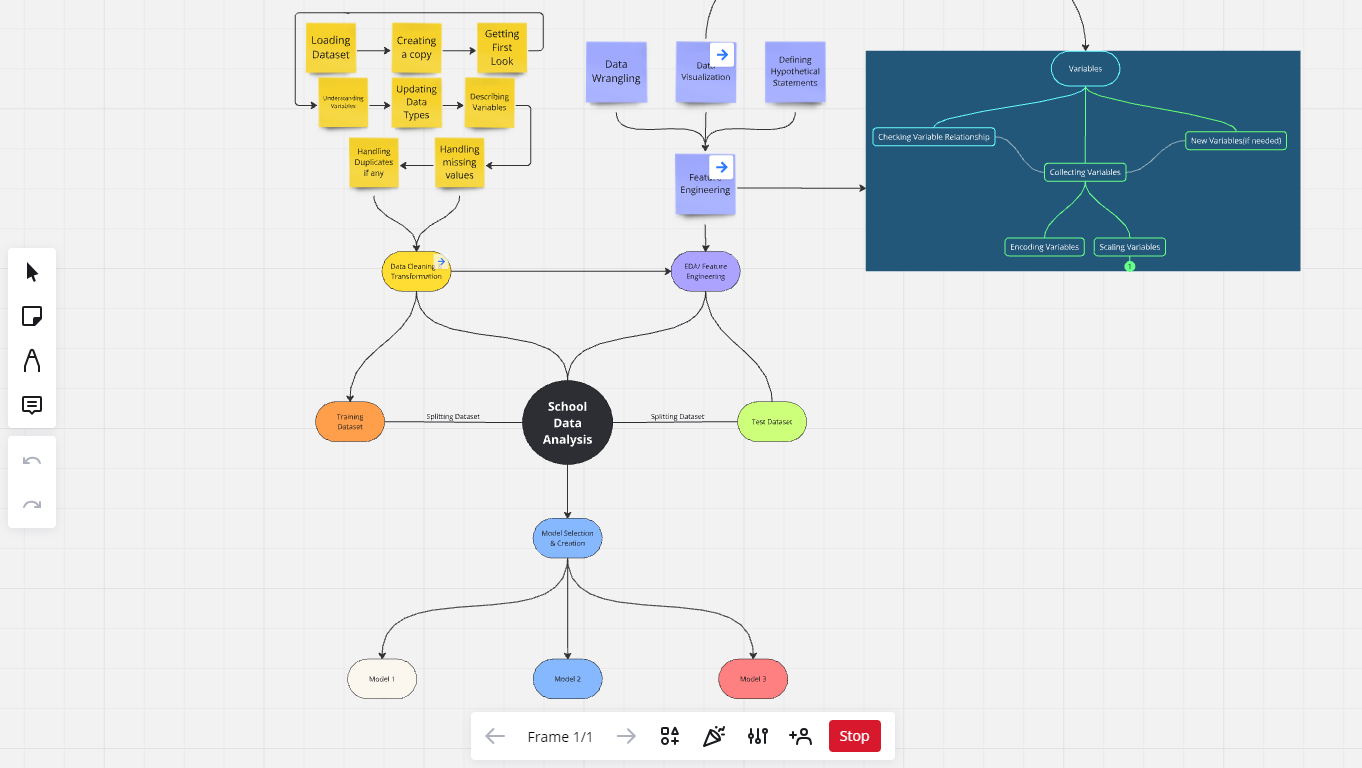
## Presentation this task

* In this presentation I will be explaining the approach towards the project and the logics/reasoning used for the outcome of the project.
* The purpose of this presentation is to make the client clear about what they can expect from the outcome of this project.
* The ultimate outcome of the project is to create an ML model in order to predict the need for schools to be built in the regions that are required to have a school.

## Steps Involved: -

* Data Cleaning and Transformation
  + Making a copy of the original file
  + Getting First Look
  + Understanding Variables
  + Updating Data Types
  + Describing Variables
  + Handling Missing Values
  + Handling Duplicates (if any)
* Data Wrangling
  + Listing Out Dependent and Independent Variables
  + Summarizing and Aggregating data as per requirement
  + Deriving insights to better understand the effect of different variables on data
  + Dividing Dataset into 70-3 or 80-20 for training and testing
  + Checking different relationships between different variables
  + Collecting independent variables that affect the dependent variable
* Data Visualization
  + Creating charts and visuals to visualize the relationship between variables
  + Collecting independent variables that best works for our datasets
  + Checking the relationships for each of them by visualizing the relationship
* Defining Hypothetical Statements
  + Hypothesis Testing
  + Measuring Accuracy
* Feature Engineering
  + Handling Outliers
  + Categorical Encoding
  + Feature Manipulation and Selection
  + Creating new features to use in our model
  + Data Transformation (If needed)
  + Data Scaling
  + Data Splitting
  + Handling imbalances
* ML Model Implementation
  + Measuring Results
  + Model Explanation
* Conclusion & Dashboarding
  + Defining Collusion of the project and creating Report.
  + Creating linked Dashboard of the EDA performed.

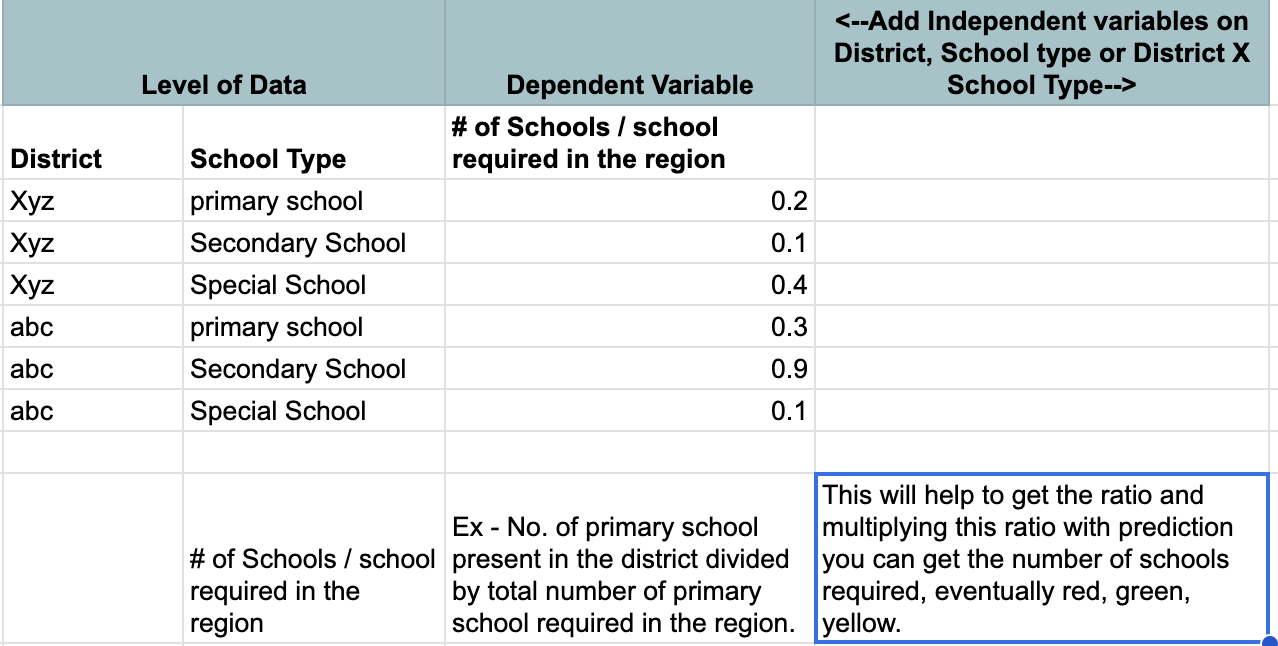
## Approach towards the solution

**[](https://miro.com/app/board/uXjVKxZbMjU=/?share_link_id=577504383926)**

## Approach towards finding dependent and independent Variables

There can be 2 ways we can define the dependent variable:

* Direct Method
  + We can simply check the difference of the total of adjusted capacity per region and the total number of students.
* School Type wise Method
  + In this method we can check the ratio of the number of schools required for each type of school. We can follow the logic below.

We will choose as per the results, whichever makes more sense.

Now for the independent variables we will be performing EDA on our dataset and will also be doing feature engineering for the same, we will check as per the visuals, that which features are having the most impact on our dependent variable. Once the variables are collected, we will then start making the model and check for the best results.

For the data split we will be splitting the data into 70-30 split and will train and validate our model respectively, the validation will be having few accuracy measures along with the confusion matrix.

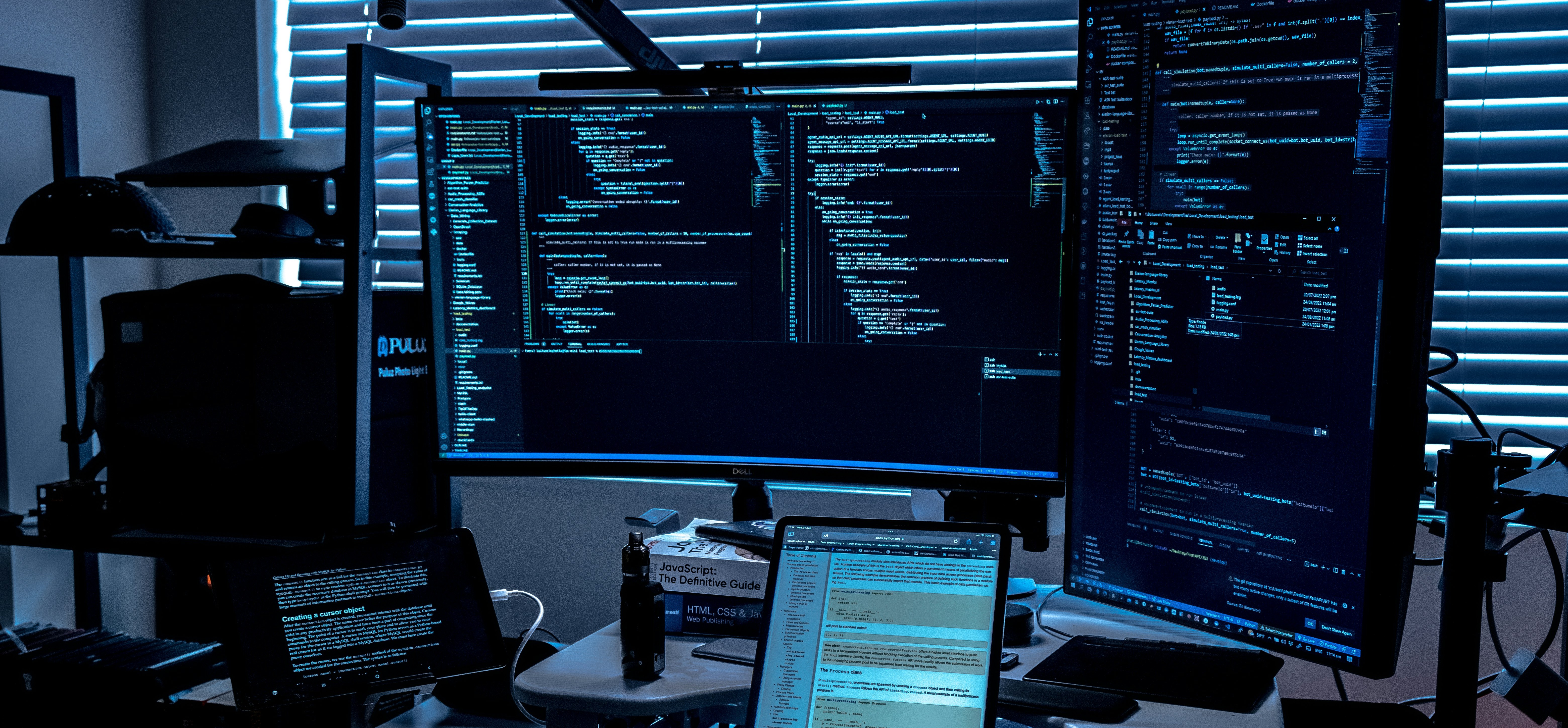
## Result Deliverables:

I will be performing the above mention tasks and will provide all the services mentioned above.

The result will be given as such:

* The .ipynb file along with the transformed dataset + a tableau dashboard + project report.

Project Files and Dashboard



|  |  |
| --- | --- |
| Status | Done |
| Priority | High |
| Assignee | Kartikey Rastogi |
| Due | @August 1, 2024 |
| AI  summary | The document discusses insights from various visualizations related to school capacity and student population, highlighting the need for new schools in underutilized and overburdened districts. It presents multiple charts that analyze relationships between total capacity, future students, and school conditions,  concluding that models like Logistic Regression, Decision Tree, and Random Forest show high accuracy in predicting school  needs. Recommendations for improvement include obtaining  more specific data on school needs per student population to enhance model accuracy. |

# Project Summary

In this project we have been given 2 data sets: **School Data** and **Population Data**. We are asked to create 4 ML model that can predict the urgency of the need of new schools in each district.

The school data contains the information of different schools present in each district, and also the estimated population of students present in that district. Using this information, we are required to wrangle the data in such a way that it can help us create an ML model to be used to predict the need for new schools.

The “Need\_Categoryˮ is divided into 3 categories: -

 **RED:** Where the urgency is most.

 **Yellow:** Where the urgency is moderate.

 **Green:** Where the urgency is least.

## Work Flow of the project

We started the project with the School Dataset as it was our main dataset in which we were required to work mostly. There are different steps that were taken in this project which will included in this report subsequently.

Variables Description - School Dataset (Numerical Columns)

 **The Year the school was built**

 We can see the dates are all 01-01 hence we can take only the years from this column.

 The time is also not relevant hence we can simply focus on the year.  We can clearly see that:

 the min year is 1908

 50% of data as per years goes till 1987  25% of data goes till 1975

We can see the year gap between 0%-25% is huge however the gap between 25%-50% is not much, which means that the frequency to build schools was less till 1975 and it started increasing after that.

 **Capacity**

Here we can see the 50% data is evenly distributed approx.



 However, the next 25% (i.e., 75%) is a bit less, but considerable.

 The max as we see is of course an outlier and we can say it has to be treated as an exception.

 If we check the standard deviation of this column, we can clearly see that this column is spread over a wide range and is not saturated,

however, it can also be due to few outliers that may be impacting this, we need to check that.

**Shifts**

 Looking at the shifts we can see it is a pretty direct column and is evenly distributed.

 We have 1, 2, and 3 shifts.

**Total number of Students**

 It is a sum of the column’s boys and girls.

 The min value is pretty concerning and it says 3 which clearly means it is an outlier.

 Here the data distribution as we can see is around 250-300.

 However here also the difference between the 75% and Max value is huge which means here we can see the upper limit outliers.

 Here also we can see there is a huge deviation in our data range as per the total number.

**Boys and Girls**

 As we can see these columns are very closely divided and they are almost even.

 However, there is a huge difference at the max values, other than that they are almost evenly divided as we can see.

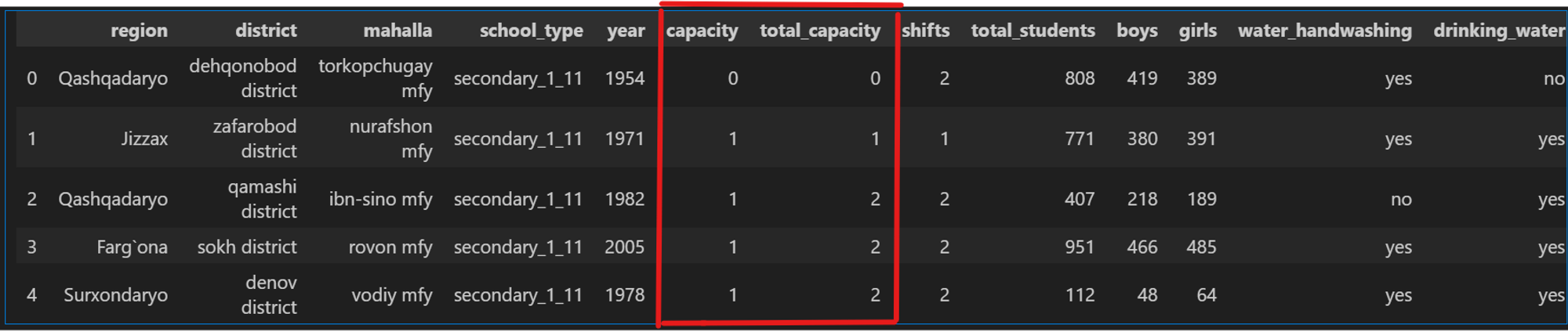
**Transport Stops**

 This is the column where we were having the most missing values  We will have to be careful at the time of consideration.

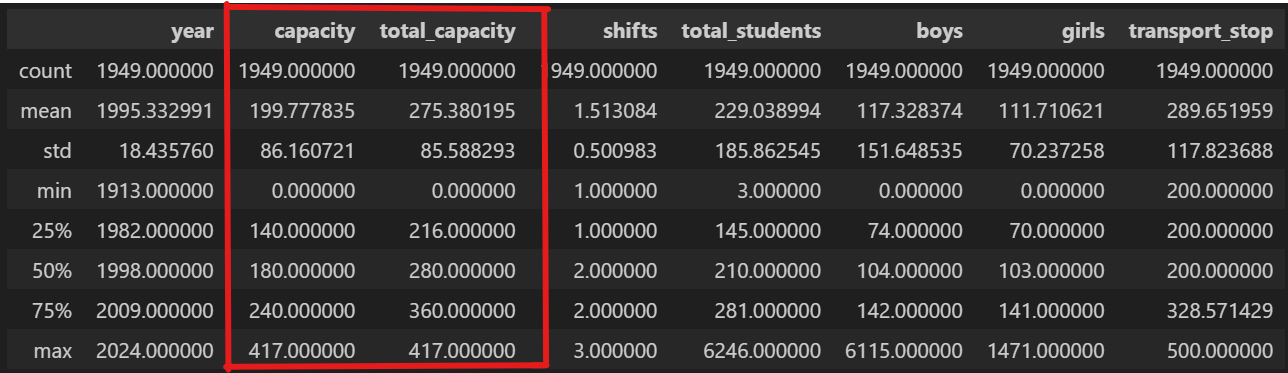
 This column is having most data around 200 and the max is 500 hence we can say that there is a good connectivity as the transport stops are

nearly within 500 meters and mostly within 200 meters.

When starting with the data wrangling, we came across few Outliers that were there in the total capacity column that we created. Hence, we checked for the outliers and got to know that they are within the starting 25% of the dataset.



As we can see here that the Total Capacity column is not making sense, and this data is not good for our machine learning model, hence we removed these outliers to keep only that data which is good for our ML model to be trained.



Now here we can see that the 25% of this division, has started to make sense, as the data before that is obviously less than 216 which is likely to be outliers hence, we will exclude them from our data, so that we can keep the data that really makes sense for our model to train.

Now in the population dataset, we created a new column with the name of Growth Rate. Where we measured the population growth rate of each district over the years.



This is the sample look of the Data frame. Using this growth rate in the population we calculated the estimated population of the students in each district after 5 years. We compounded this growth rate to get the estimated population of the students after 5 years.

# Merging the dataset on district

work\_df = work\_df.merge(pop\_work\_df[['district', 'Growth Rate

# Estimating the student population after 5 years

work\_df['future\_students'] = round(work\_df['total\_students']

The approach behind this is that, we are looking forward to create new schools with the future perspective hence we need to see the requirements as per 5 years, that how much will be the requirement, after 5 years as per the estimated population.

## Calculating the need of population using Adjusted Deficiency.

Here are all the manipulations that were done in order to do so. The steps involved in the manipulation are as follows: -

 Created a total capacity column ⟶ (capacity)\*(shifts)  Identified outliers in the total\_capacity column.

 Checked for the range in which outlies were present \* 25% of the total dataset

 Checked for the range of outliers \* 0-216

 Projected future population of students after 5 years.

 Used the population dataset to get the population growth rate of each district.

 Took pct\_change over the years and checked it's average

 Used the population growth rate in each district to calculate student population after 5 years as per the growth rate.

 Calculated adjusted capacity as per other factors.

 Checked for the current utilization of the capacity

 Checked for the deficiency of capacity as per estimated student population

 Increased the deficiency by different precents according to the weightage of the factor.

 +50% deficiency \* Unsatisfied; +70% deficiency \* Extremely Dissatisfied

 +20% deficiency \* No infrastructure

 +10% deficiency \* No water for drinking or handwashing  +10% deficiency \* Transport stops  800 meters.

 Created adjusted deficiency considering these factors ⟶ adjusted\_deficit

= deficit \* adjustment\_factor

 Created Red, Yellow, Green categories as per the deficiency.  adjusted\_deficit  10% \* Red

 adjusted\_deficit  5% \* Yellow  adjusted\_deficit  5% \* Green

Note: In these manipulations the updates will be made and the updates will be mentioned below.

UPDATES

 As we checked the 'state\_of\_your\_column' as per the distribution of

Need\_Category we came to know that even if the state was unsatisfied many values were coming into the Green portion, hence we applied the penalty directly on the capacity column.

 50% penalty on unsatisfied and 80% penalty on extremely dissatisfied.

# Visualizations in our Analysis for Checking relationship

These visualizations can be better seen in the python file as they are dynamic in nature.

## Visualization chart - 1

**Let's study the relationship of other columns with the Need Category**



## Why did we pick the specific chart?

Scatter plot is the best choice to check the relationship between any two numerical variables, and as total\_students and total\_capacity as the columns that contribute the most in our dependent variable we had to see the

relationship. Plotly Express also helped us to identify the points for each category.

## What is/are the insight(s) found from the chart?

In the chart we can clearly see that the relationship is pretty linear and we can also see that the categories are also making sense as the total\_capacity and total\_students are having clear difference in the values, hence giving us the urgency indicator for the values. We can see that there are few outliers as well, however they may also be some real facts depending on the situation.

## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, as we can see the linear pattern that is being formed as per the difference

between the total capacity and the total students and also the accuracy of the indicator is also good enough we can say it help our model to understand the impact of these columns on the result.

## Visualization chart - 2

**Let's visualize the future students vs the total\_capacity**



## Why did you pick the specific chart?

Using this chart, I can clearly see that what would be the estimated population of the students in the future and as per that, how much capacity are we having now. Here the division of the data seems pretty clear, as the higher the height will be the higher will be the need for new schools to be built.

## What is/are the insight(s) found from the chart?

In this chart I can see mostly where the capacity is pretty low and also, there are few places where there is enough capacity but still not enough for the

estimated population, it makes us understand that there are many places where there is huge gap between capacity and population.

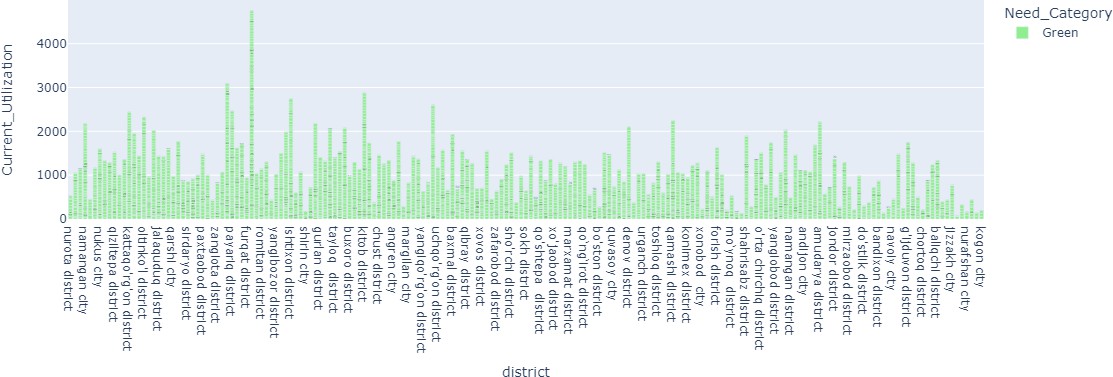
## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, as we can see the relationship is pretty direct that we need to look at what our population is going to be and we need to predict as per future perspective only depending upon how much capacity do, we have now.

## Visualization chart - 3

**Let's check which districts are having mostly underutilized schools**



## Why did you pick the specific chart?

In this bar chart we can clearly see that how much these districts are underutilized, as the higher the bar, the more underutilization is there in the district.

## What is/are the insight(s) found from the chart?

Here we can clearly see that the most underutilized district is the - ishtixon district also there are many other that are only 3-4% utilized which is a great concern, as this directly indicates the wastage of the resources.

## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

This is not directly related to our prediction model; however, it is a good

parameter or insight so as to manage the wastage that is going on as there are many places where the utilization is very high and there is an utmost need for new schools there.

## Visualization chart - 4

**Over utilization among districts**



## Why did you pick the specific chart?

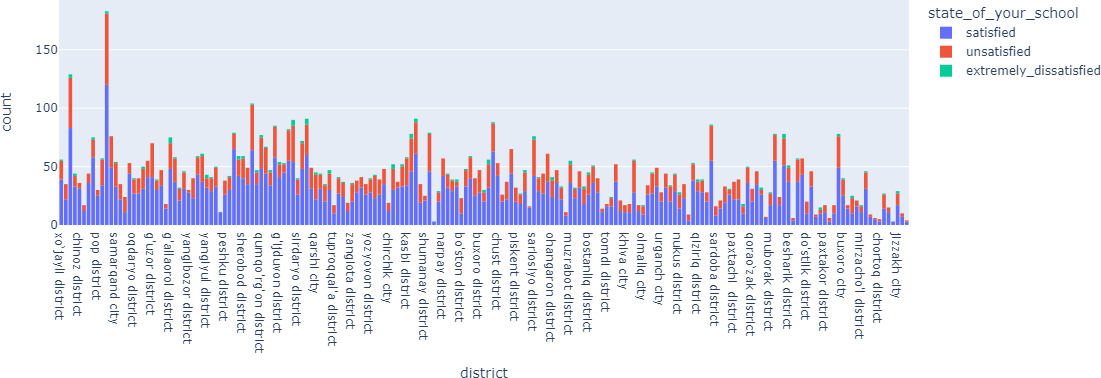
Here we can see the districts where there is over utilization of the capacity it can be clearly seen using a bar chart as the previous one.

## What is/are the insight(s) found from the chart?

We can see there are many districts that are highly over burdened, and the Urgut district is having the highest number of over burdened schools hence making it clear the need for creating new schools.

## Visualization chart - 5

**Districts vs State of School**



## Why did you pick the specific chart?

In this histogram we can easily see the distribution of the state of our schools in each district, on how much schools are not meeting the expectations in each district.

## What is/are the insight(s) found from the chart?

We can see there are 3 categories: -

 Extremely Dissatisfied: Which means they should not have much weightage in our data and the capacity in these schools should be highly adjusted.

 Unsatisfied: These values are distributed highly as they are having around:  MAX freq. RANGE: 0-20 schools each district

 MID freq. RANGE: 20-40 schools each district  MIN freq. RANGE: 40-60 schools each district

 Satisfied: The maximum freq. range is till 50 and then till 50-100

## Will the gained insights help in our prediction model?

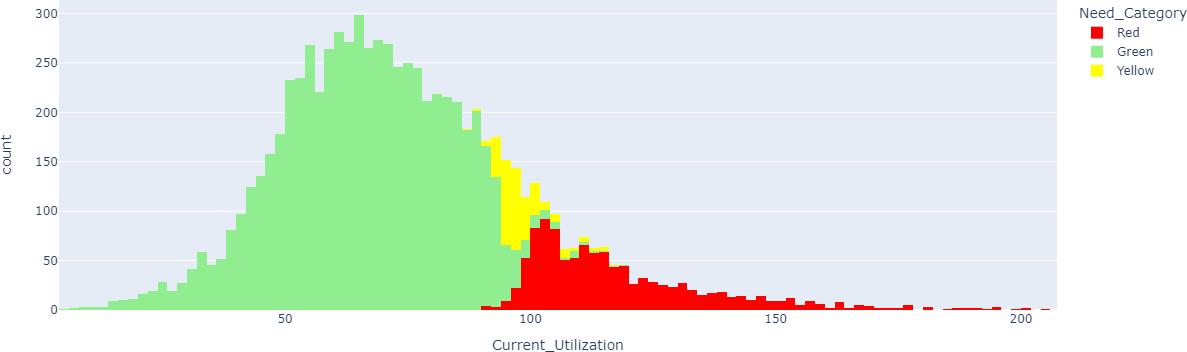
Are there any insights that lead to negative growth? Justify with specific reason.

Looking at this data we can come to this conclusion that in each district there is a good number of schools that are not satisfactory, hence the capacity in these schools shouldn't be considered completely as due to lack of schools and good environment the students are required to study here only hence, we need to penalize their capacity values as the capacity is playing a major role in us

target column.

## Visualization chart - 6

**Current Utilization**



## Why did you pick the specific chart?

In this histogram I am easily able to see the distribution of the % of the current utilization of the total\_capacity, I am also able to see the impact on what happens to our data as the utilization increases.

## What is/are the insight(s) found from the chart?

As seen in the chart, most of the utilization is around 64-65% which is good and it is clear that there is no need now to build any school in those areas,

however, as we move ahead, the utilization is still intact, and starting from

around 90-100 (which means that the capacity is completely filled), the bars start to turn yellow, hence there is a moderate need to build new schools.

Moving ahead we can see that in some schools the utilization of the capacity is more that 100% which means there is an utmost need for new schools to be built.

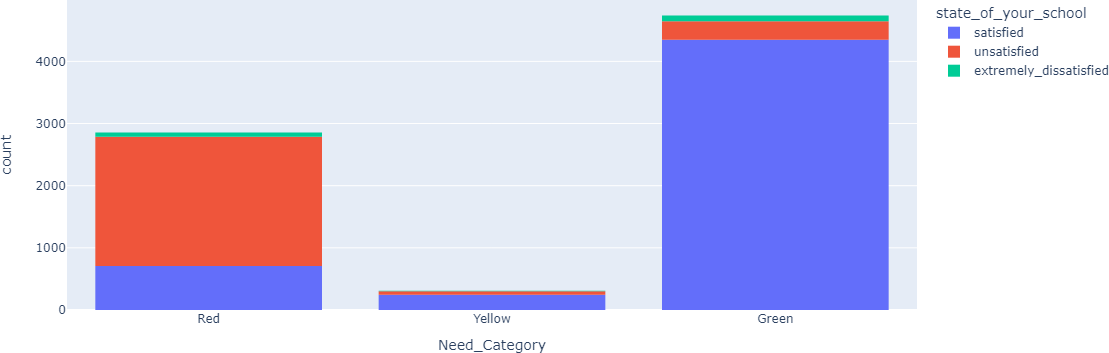
## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, definitely this will have a good impact on our model as, when the capacity will start to increase it is likely to happen that after 1 point the need for schools will start to grow, it has a direct impact.

## Visualization chart - 7

**State of School vs Need Category**



## Why did you pick the specific chart?

In this histogram we can clearly see that even if there are many unsatisfied and extremely dissatisfied values, the majority is still coming in green hence we

need to adjust the values accordingly in order to fix the distribution.

UPDATES

Now we have adjusted the values as per the results and now we can see the change which makes better sense.

## What is/are the insight(s) found from the chart?

Now after making all the adjustments as per requirement, we can clearly mention that if the state of school is unsatisfactory then it is more likely to come into the red category, because already if the capacity is way more than it is still coming into the Green or yellow one, however when there is a close call the values are most likely to fall in the red category.

## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, as the state of school is a good factor to consider when we see if there is a need for school or not, as it couldn't be left without taking into consideration.

## Visualization chart - 8

**Infrastructure vs Need Category**



## Why did you pick the specific chart?

Now in this case histogram helped us to clearly identify the difference in the distribution of the infrastructure, we can see that there is not much difference in the yes, no values, but we can clearly see the difference in the category values of our dependent variable.

## What is/are the insight(s) found from the chart?

This data as we see makes much more sense as the majority of the "RED" category can be seen where there is no infrastructure, however we can see that there are few reds in the yes column as well which can be due to high

deficiency in those areas, we can check it separately what would be the reason of these reds. Checking this in the Chart-9.

## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, as we can see that infrastructure is playing a vital role in the need category and is also allowing us to identify those locations where there is good infrastructure but lack of capacity. Which can help us to conclude accordingly.

## Visualization chart - 9

**Total Capacity vs Future Students**



## Why did you pick the specific chart?

In this plot we can clearly see the difference line between the red and green categories, scatter plot is the best to plot the point differences.

## What is/are the insight(s) found from the chart?

As we suspected the red ones that are there in the schools where the infrastructure is yes is due to the high deficiency of total capacity. Which makes it clear that our data is making correct sense.

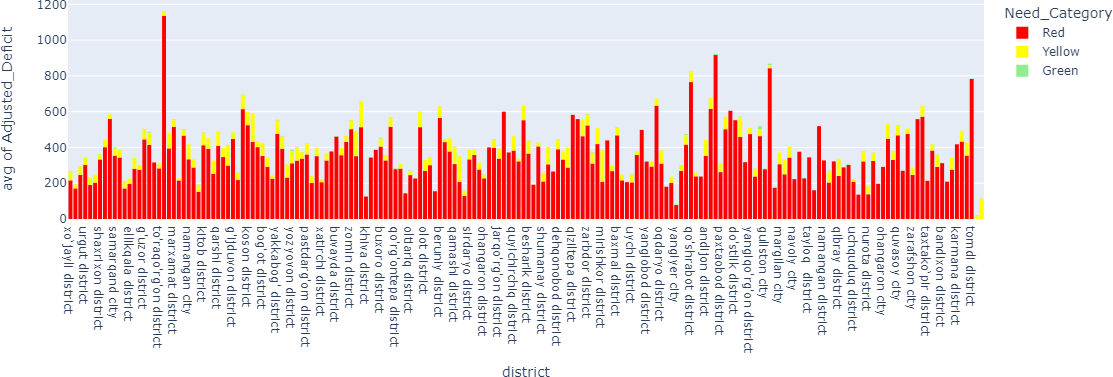
## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, the reason has been already mentioned above in the Chart-8 this one is to prove the correction.

## Visualization chart - 10

**District vs Avg Adjusted Deficit**



## Why did you pick the specific chart?

This histogram can easily show that in each district, the higher the deficiency, the higher will be the need to build schools and we can see that where it is deficiency the category is coming up to be red, which is correct.

## What is/are the insight(s) found from the chart?

Using this chart, we can see that the Mirabaud district is having the highest deficiency average hence it is in the most need for new schools to be built. We can also see that there are few districts that are having the average deficiency of more than 600, however rest of these are below 600.

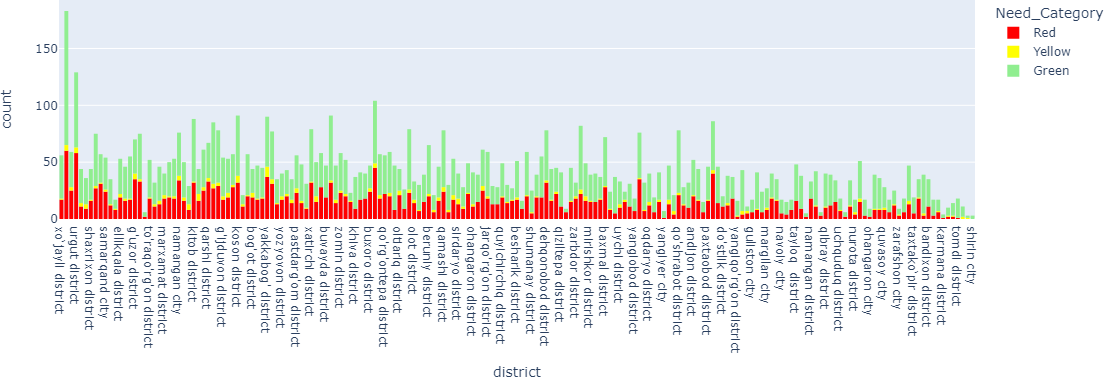
## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, this will help us to identify the utmost need of schools in each district, we can define a parameter on which ones to consider for each category

## Visualization chart - 11

**District vs Need Category**



## Why did you pick the specific chart?

This histogram is good to see the distribution of each category as per the district, it can help us to identify the need to schools in each district as per the count of each category in each district.

## What is/are the insight(s) found from the chart?

From this chart we can see that the Chiroqchi District and Urgut District are in the most need of new schools to be built. We can set a parameter to get the number of red categories, let's say if any district has more than 20 RED counts, we will take it as red in consideration.

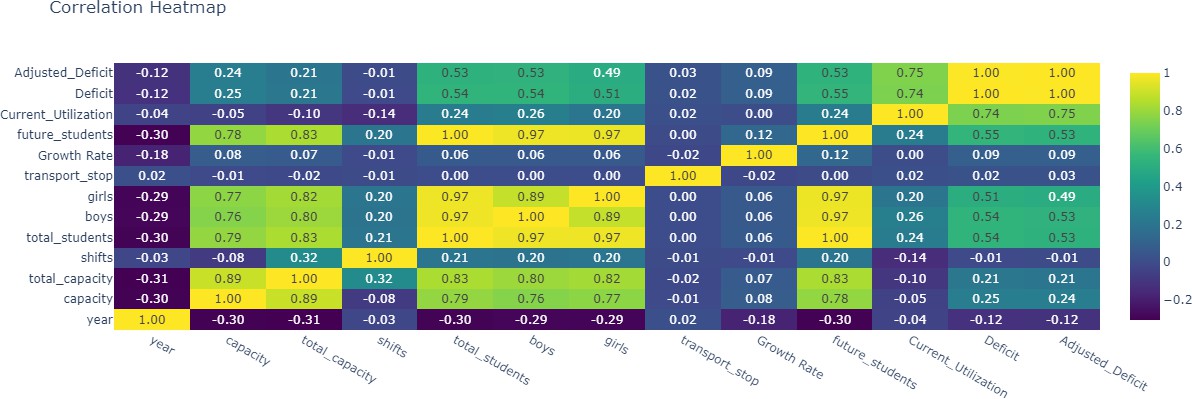
## Will the gained insights help in our prediction model?

Are there any insights that lead to negative growth? Justify with specific reason.

This insight is actually the resultant insights on how to define the results that we are getting and the parameter to define the collusion can be discussed.

## Visualization chart - 12

**Correlation Heatmap**



## Why did you pick the specific chart?

Here we can look for the correlation between all our numerical values, and we can see which values are closely related so that we can choose which values to consider.

# Feature Selection and Pre-Processing

## Which all features were found important and why?

The feature selection steps are as follows: -

 First of all, we removed the location names column because we don't want our model to be location specific.

 As we have nothing to do with the school type and the year it was built in this model, we removed that also.

 Our focus is only on the total capacity a place is having hence we removed the capacity and shifts columns and likewise we also removed boys and girl’s column.

 Now for all the other columns that we have used to get the target column are also removed as they have been derived from the original columns

except for the Growth Rate and the adjusted deficit as it can help our model to understand the relationship.

 We obviously need to separate the target column from our dataset as we don't want it to train from the actual data.

## What data splitting ratio have been used and why?

We have taken a random 70:30 split as our dataset is widely spread and there are no dates involved hence, we have taken it.

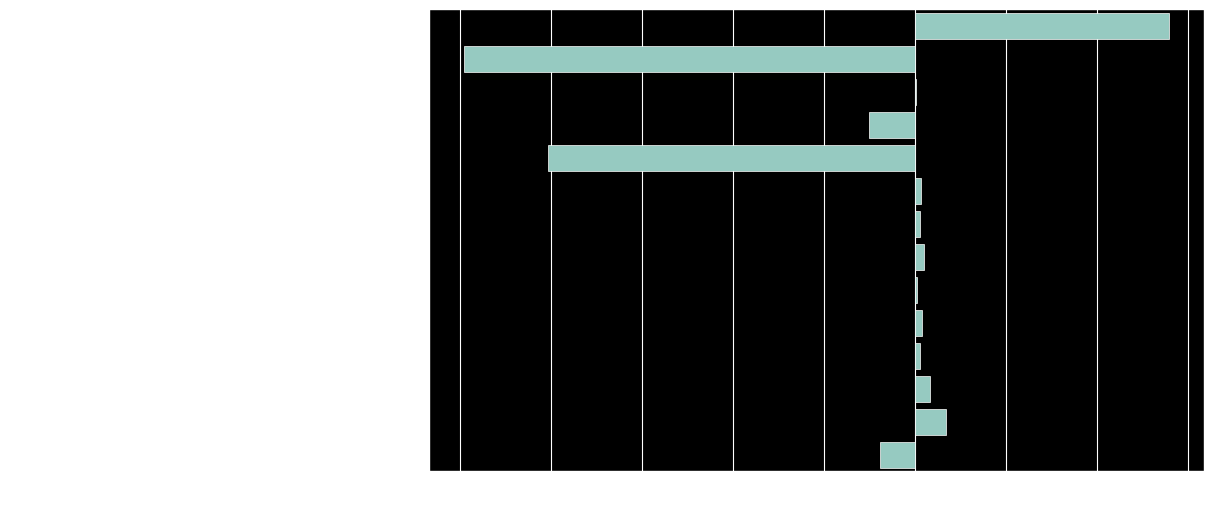
**Features taken for the model**

ml\_df = ml\_df[['total\_capacity', 'total\_students', 'water\_handwashing', 'drinking\_water', 'transport\_stop 'infrastructure', 'state\_of\_your\_school', 'Growth Rate

# ML Model Implementation

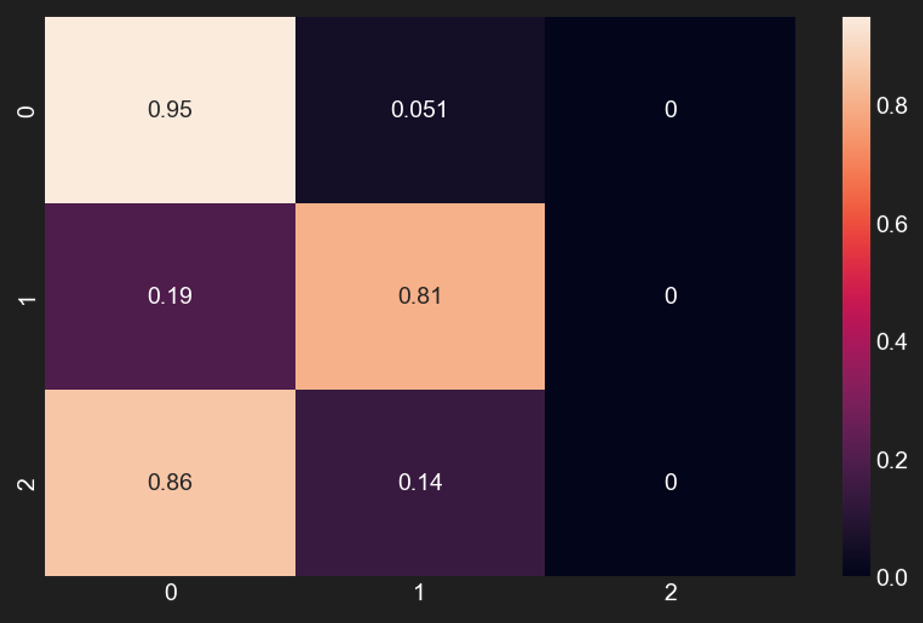
## ML Model - 1 - Logistic Regression

**Weights Given to the Features in the LR Model**

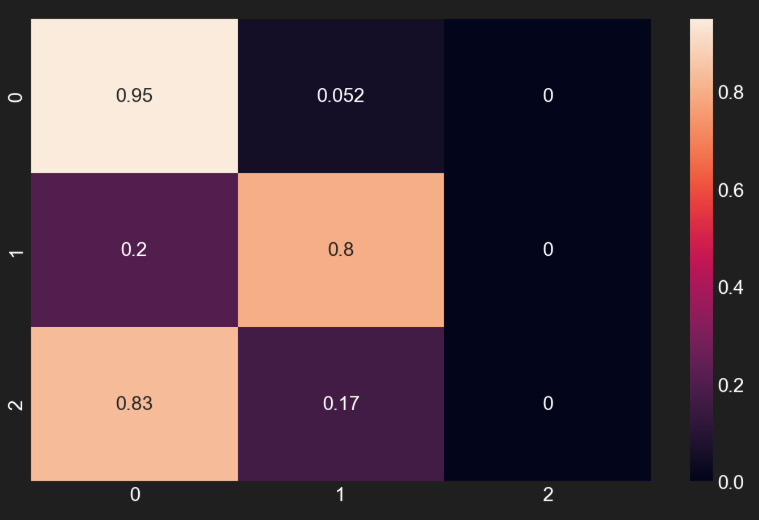


We can see that the most importance is given to the total students and then it is given to the total\_capacity, there are other features as well in which there was an amount of importance given, hence we can see in what direction our model is learning.

**Confusion Matrix - Train set**



**Confusion Matrix - Test set**



**Description on the overall performance of the model**

Here are the details of the model and its prediction.  Model Name: Logistic Regression

 Train Test split: 70% Train 30% Test

 Accuracy Score: 0.85 (test set) Which means our model is 85% Accurate on Test set.

 Accuracy Score: 0.86 (train set) Which means our model is 86% Accurate on Train set.

 Train Confusion Matrix:

 0.94 0.51 0

0.19 0.81 0

0.86 0.14 0

 Test Confusion Matrix:  0.94 0.05 0

0.20 0.79 0

0.83 0.17 0

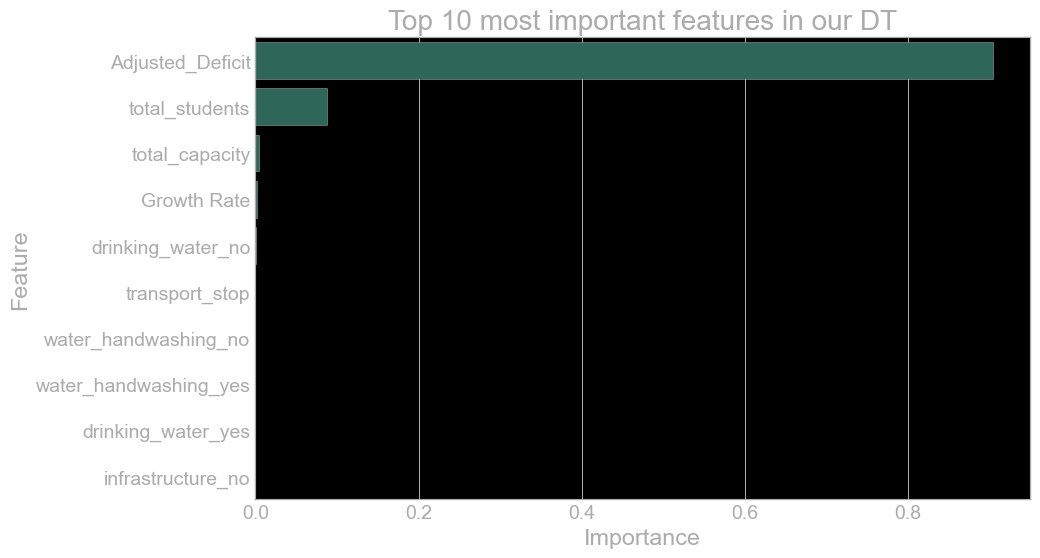
 Performance Report:

 precision recall f1-score support:

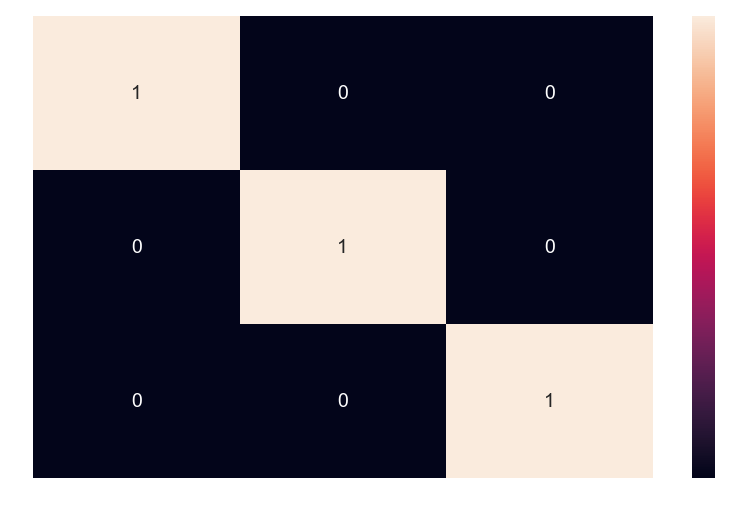
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| - Green | 0.84 | 0.95 | 0.89 | 1434 |
| - Red | 0.88 | 0.80 | 0.84 | 836 |
| - Yellow | 0.00 | 0.00 | 0.00 | 100 |

## ML Model - 2 - Decision Tree

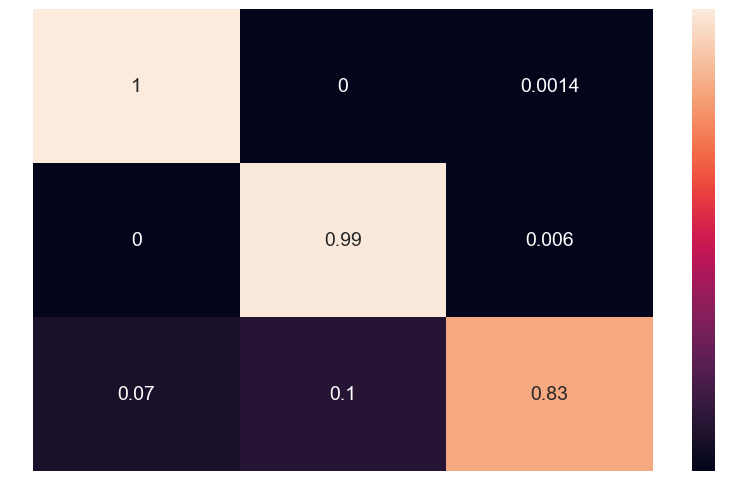
**Weights Given to the Features in the DT Model**



Here we can see that the importance given to the features in the Decision Tree is limited to the Adjusted Deficit and Total Students, few pointers of the importance are given to the total\_capacity and growth rate also. Which means that the importance is not that much divided among the features, however the results of this model were pretty good.

**Confusion Matrix - Train set**

**Confusion Matrix - Test set**



**Description on the overall performance of the model**

Here are the details of the model and its prediction.  Model Name: Decision Tree

 Train Test split: 70% Train 30% Test

 Accuracy Score: 0.99 (test set) Which means our model is 99% Accurate on Test set.

 Accuracy Score: 1.0 (train set) Which means our model is 100% Accurate on Train set.

 Train Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| - | [1 | 0 | 0] |
| - | [0 | 1 | 0] |
| - | [0 | 0 | 1] |

Test Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| - | [1 | 0 | 0.01] |
| - | [0 | 0.99 | 0.06] |
| - | [0.07 | 0.10 | 0.83] |

Performance Report:

precision

recall f1-score sup

port:

- Green

1.00

1.00

1.00

14

34

- Red

0.99

0.99

0.99

8

36

- Yellow

0.92

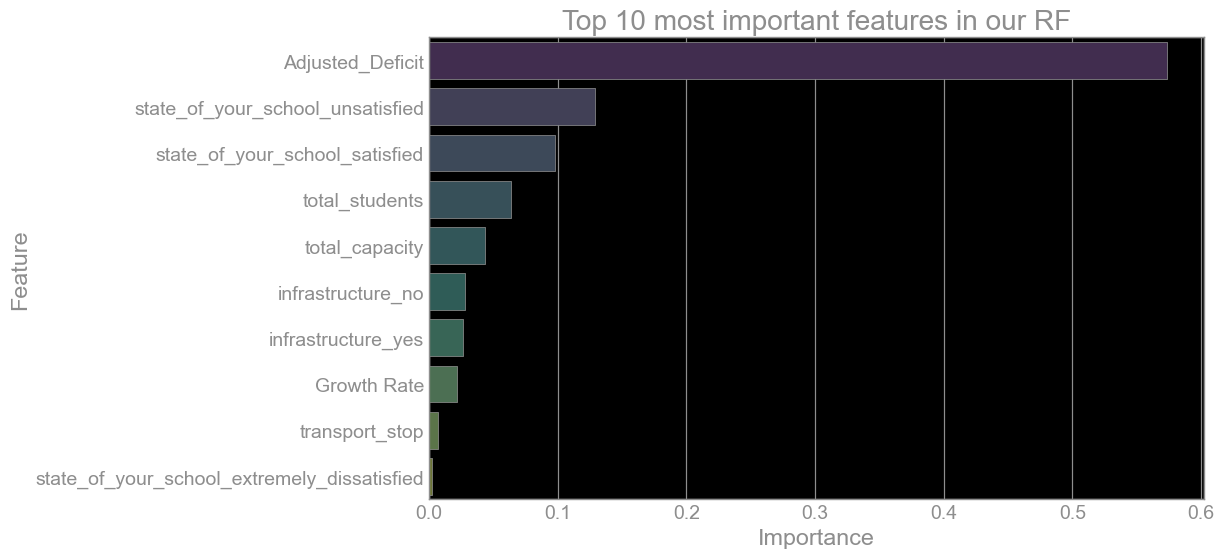
0.83

0.87

1

00

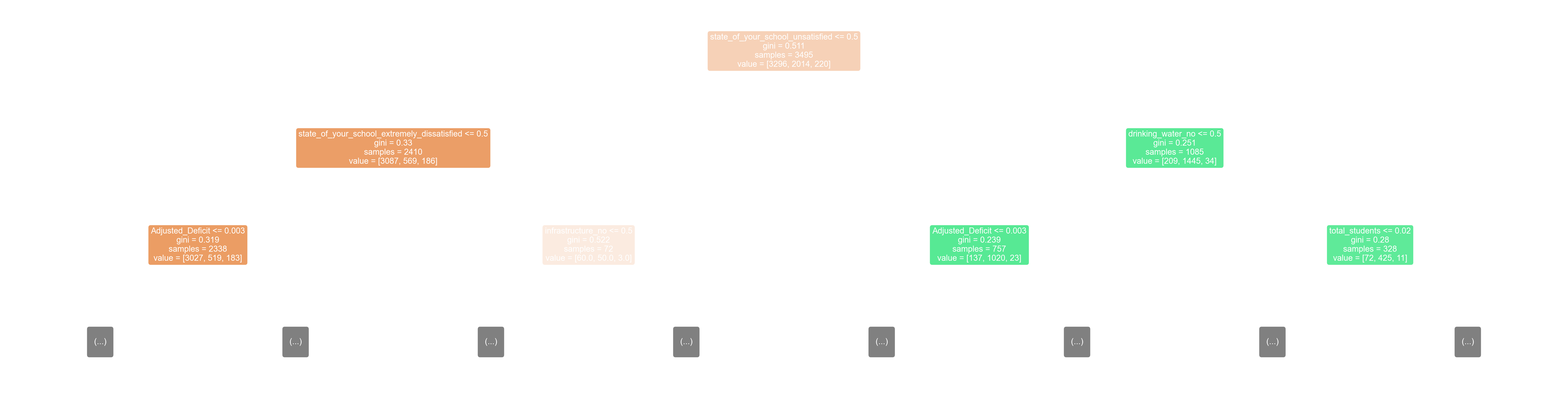
## ML Model - 3 - Random Forest

**Weights Given to the Features in the RF Model**

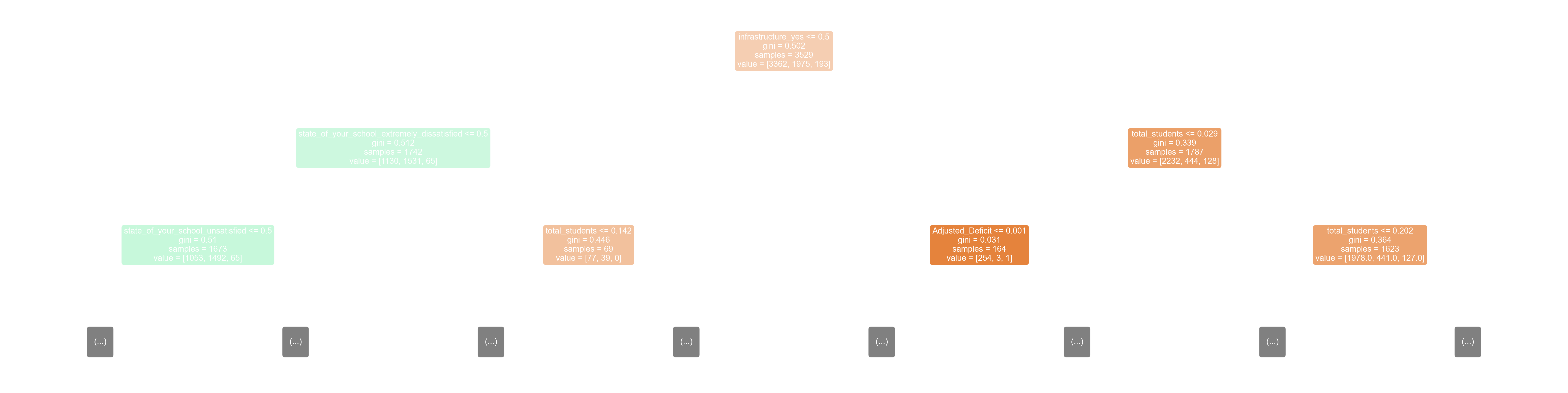
Now in this model we can see that the importance is divided among the features widely hence we can say this is more calculated results that has been given by the model.

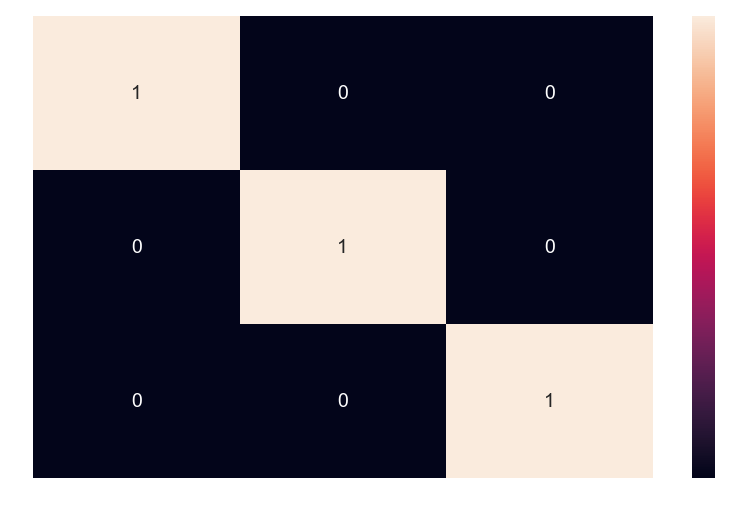
## Visualizing the Random Forest Estimators

**Plotting DT 0**

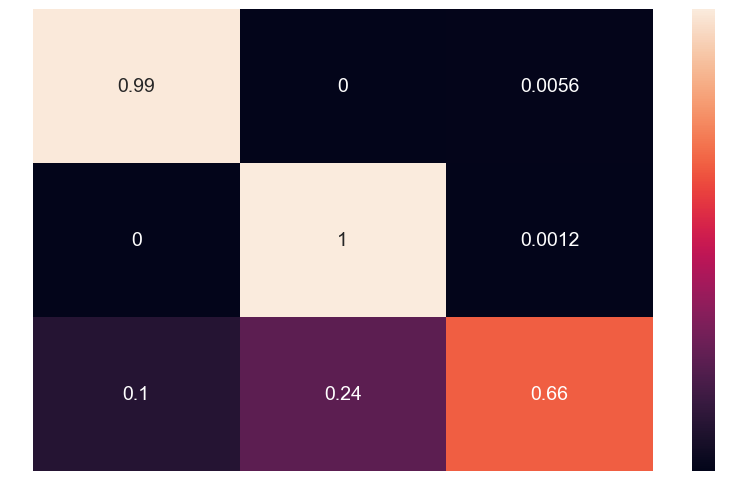


**Plotting DT 20**



**Confusion Matrix - Train set**

**Confusion Matrix - Test set**



**Description on the overall performance of the model**

Here are the details of the model and its prediction.  Model Name: Random Forest

 Train Test split: 70% Train 30% Test

 Accuracy Score: 0.98 (test set) Which means our model is 98% Accurate on Test set.

 Accuracy Score: 1.0 (train set) Which means our model is 100% Accurate on Train set.

 Train Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| - | [1 | 0 | 0] |
| - | [0 | 1 | 0] |
| - | [0 | 0 | 1] |

Test Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| - | [0.99 | 0 | 0.05] |
| - | [0 | 1 | 0.01] |
| - | [0.07 | 0.10 | 0.66] |

Performance Report:

- Matrix: precision recall f1-score sup

port

-

34

-

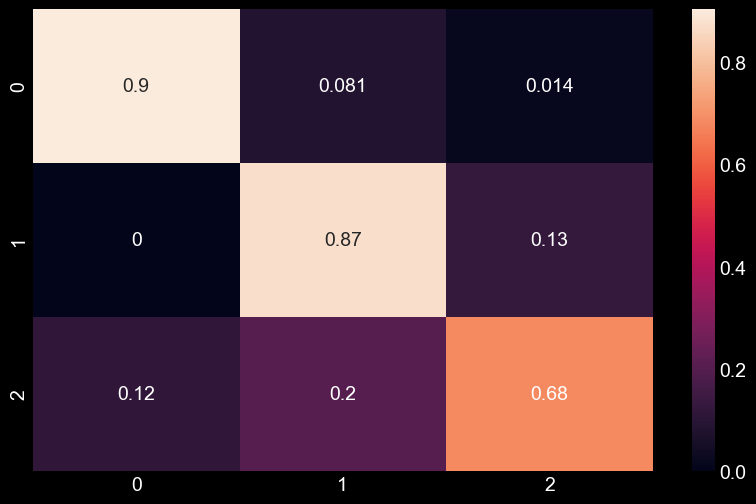
36

-

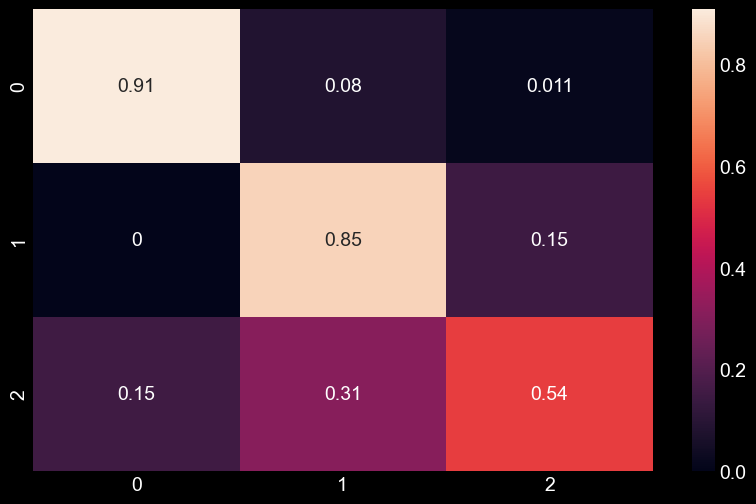
00

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Green | 0.99 | 0.99 | 0.99 | 14 |
| Red | 0.97 | 1.00 | 0.99 | 8 |
| Yellow | 0.88 | 0.66 | 0.75 | 1 |

**ML Model - 4 - Naive Bayes Confusion Matrix - Train set**



**Confusion Matrix - Test set**



**Description on the overall performance of the model**

Here are the details of the model and its prediction.  Model Name: Naive Bayes

 Train Test split: 70% Train 30% Test

 Accuracy Score: 0.87 (test set) Which means our model is 87% Accurate on Test set.

 Accuracy Score: 0.88 (train set) Which means our model is 88% Accurate on Train set.

 Train Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| - | [0.9 0.08 | 0.01] |
| - | [0.0 0.87 | 0.13] |
| - | [0.12 0.2 | 0.68] |

Test Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| - | [0.91 | 0.8 | 0.01] |
| - | [0.00 | 0.85 | 0.15] |
| - | [0.15 | 0.31 | 0.54] |

Performance Report:

- Matrix: precision recall f1-score sup

port

-

34

-

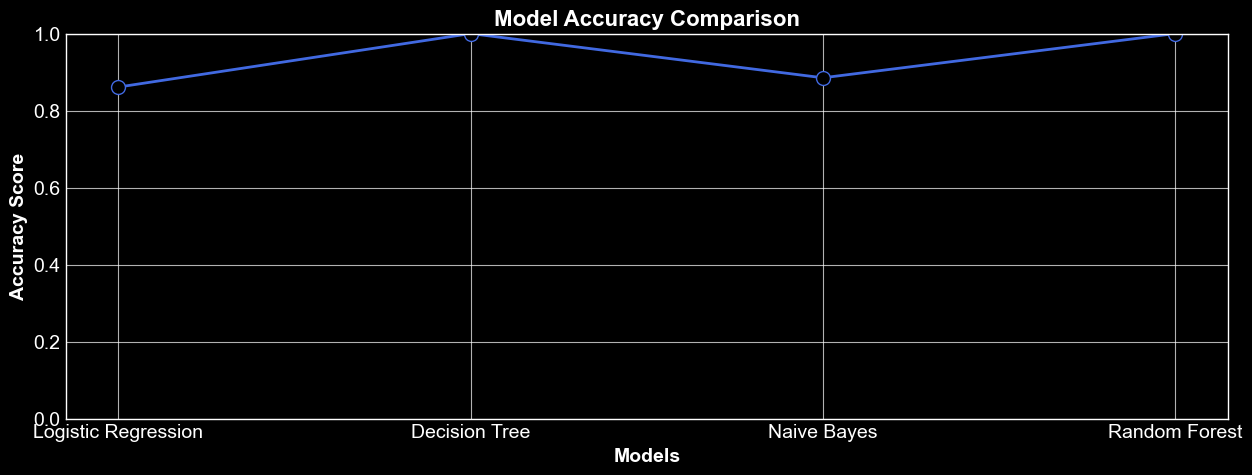
36

-

00

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Green | 0.99 | 0.91 | 0.95 | 14 |
| Red | 0.83 | 0.85 | 0.84 | 8 |
| Yellow | 0.28 | 0.54 | 0.37 | 1 |

# Overall Performance Graph



Here we can see the results of our models are all above 85%. However, the **Decision Tree** and the **Random Forest** are very close to 100% (99%, 98%) which means they are the best performing models. And if we consider those models, we can give an additional point to the Random Forest model as it is giving importance to the other features as well.

**Here is how we have divided the districts.**

We have divided the need categories as per districts. In each district, if the count of GREEN values is less than 50% then it will be in the RED zone, else if count is more than 50% and less than equal to 60% then they will be in the YELLOW zone, else if value is more than 60% then it would be in the GREEN zone only.

# How can these model be improved?

If we talk about the calculations, they were pretty much enough to establish a relation between the different features, however if we can have the data of the actual need of schools for every 1000 student population the accuracy in terms of reality can be very much authentic.

This prediction model is not location specific, and this model is made as per future perspective. We have taken how much student population will be there after 5 years. The 1 thing that is different is the change in the school population

as we are not having the data of on an average what population of students are passing out each year and how much are getting in, if we could get that

information we can consider the change in population more accurately.

Hence there are few of these data pointers that can improve the real time accuracy of these models.