# PREDICTING AND MODELLING THE SCHOOL PERFORMANCE FOR 2020

BY SABELO

DATE: 2020-08-25

# IMPORTANCE OF EDUCATION IN SA

**Motivation and rationale**

- Education is a basic right of every Human on this Planet, and the governments of all countries ensure to spread Education. It is a weapon to improve one's life and certainly determines the quality of an individual's life. As it teaches the value of discipline to individuals and enables individuals to express their views clear manner.

- Educated people are quite likely to convince people to their point of view, off which is desirable trait from employment and entrepreneurship. Furthermore, it helps in spreading and transferring knowledge from one generation to another in our societies.

While education improves one's knowledge, skills and develops the individual personality and attitude, there is a lack of quality Teachers amongst the school in South Africa. Then question arise:

- **If most of our quality Teachers are in Urban schools, does this mean the children in poor areas do not deserve quality education?**
- **And what is the relation between Quintile and school-performance?**

# DATA RESOURCES

The capital City of the EC province addresses is extracted from Google Maps API reverse geocoding**.** The rest of the data contains school performance info, district names, latitude and longitude, suburb and other unrelated info., as details below:

- ***School location Data***: To get the neighbouring information about the school. This include the nearest towns, municipality around. The size of the data is 6256, furthermore, it indicates whether the school is public or private, can be found http://ecdoe.co.za/files/resources/resource_260.xlsx.

- ***School Performance Data***: contains the historic performance data from 2017. It outlines the number of Quintile, students wrote, achieve and failed, together with the school pass rate , the size is 877 for EC province. entries. https://www.sace.org.za/assets/documents/uploads/sace_54588-2020-01-10-2019%20NSC%20School%20Performance%20Report.pdf

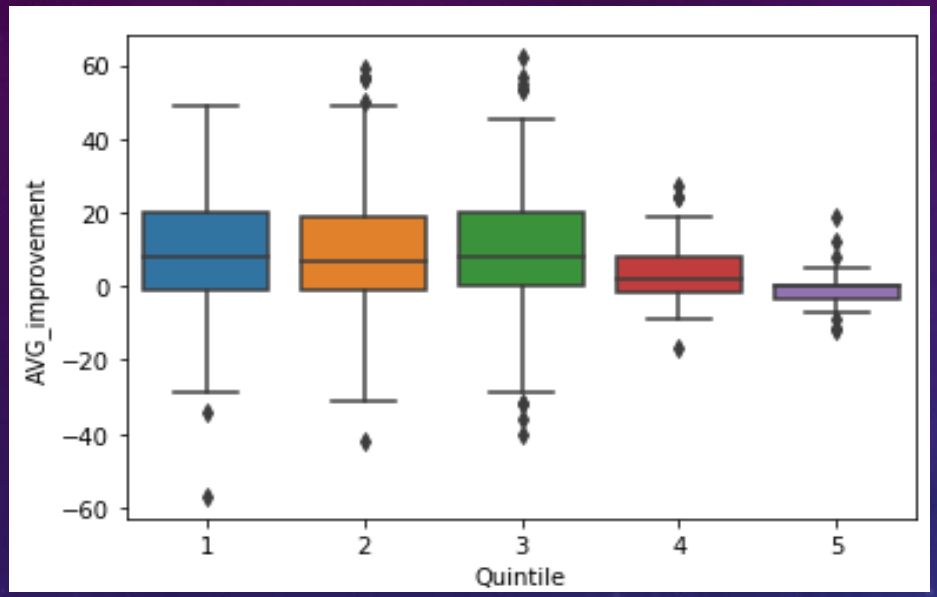
- we use Foursquare REST APIS services and Folium to analyse and create clusters.

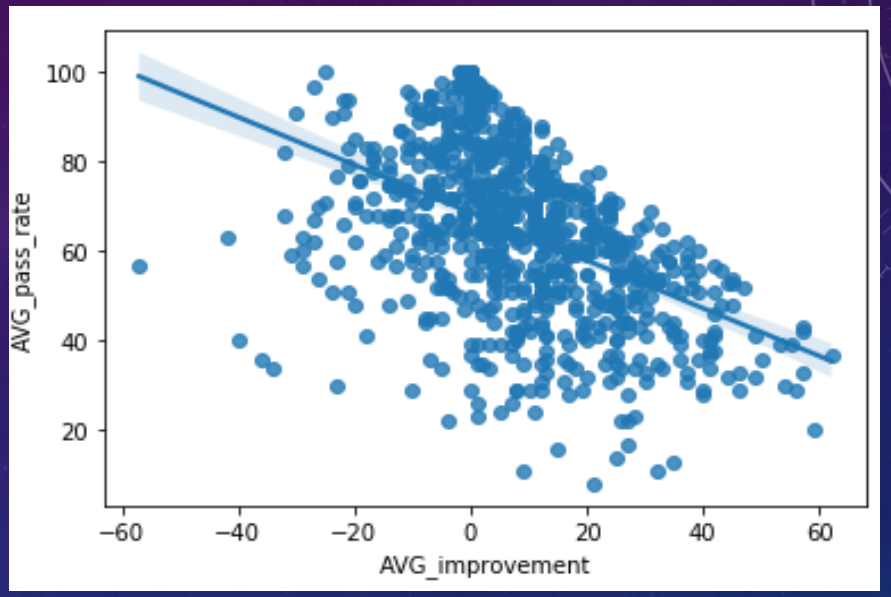
# OUR APPROACH

- We used different data sources to gather information as outlined in the previous section. After the processing of data, we determined the important features for modelling suing both classifying and Linear modelling algorithms.
- We used Linear regression, Ridge, Rando Forest and Boost as the linear models. While Support Vector Machine, Decision Tree, K-Nearest Neighbour, Logistic regression were used as classifiers.
- To achieve the project objectives and drawn solution to the problems identified, we used a hybrid of the *Sample, Explore, Modify, Model, and Assess (SEMMA)* and CRISP data mining approach.

| SEMMA | CRISP-DM |
|---|---|
| ---------- | Business understanding |
| Sample | Data Understanding |
| Explore | |
| Modify | Data preparation |
| Model | Modeling |
| Assessment | Evaluation |
| ---------- | Deployment |

# SCHOOL-PERFORMANCE IMPROVEMENT DISTRIBUTION

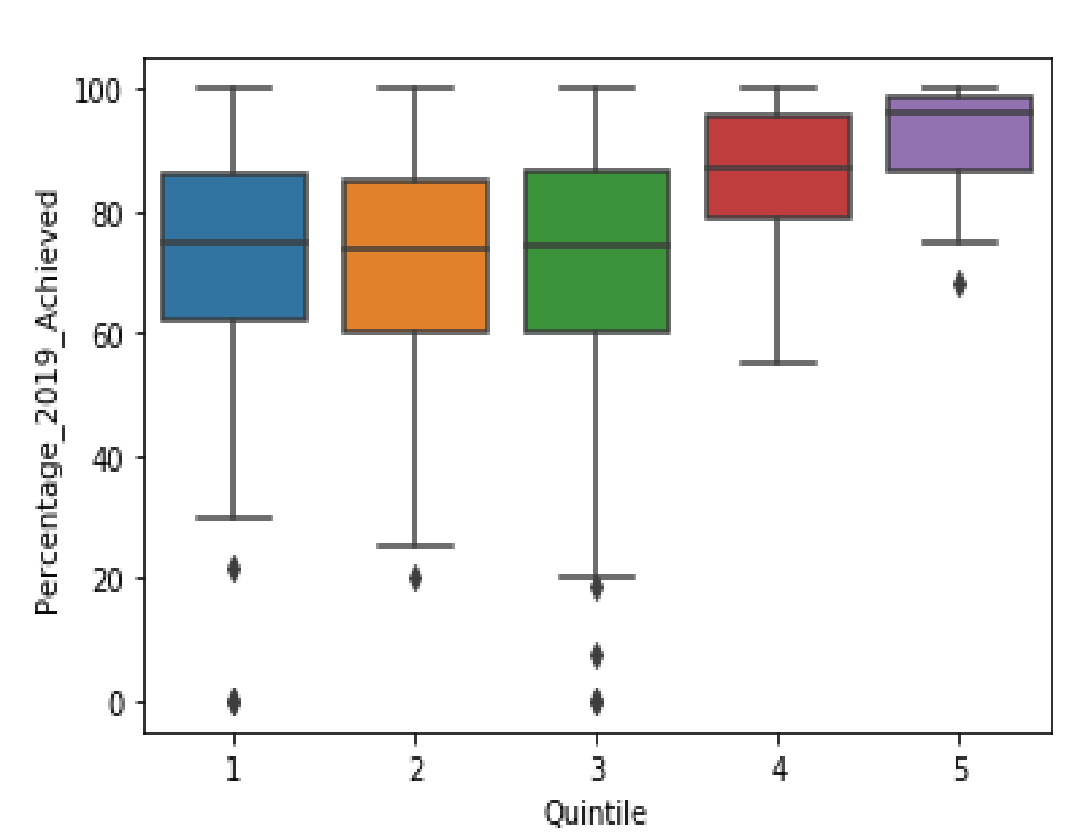


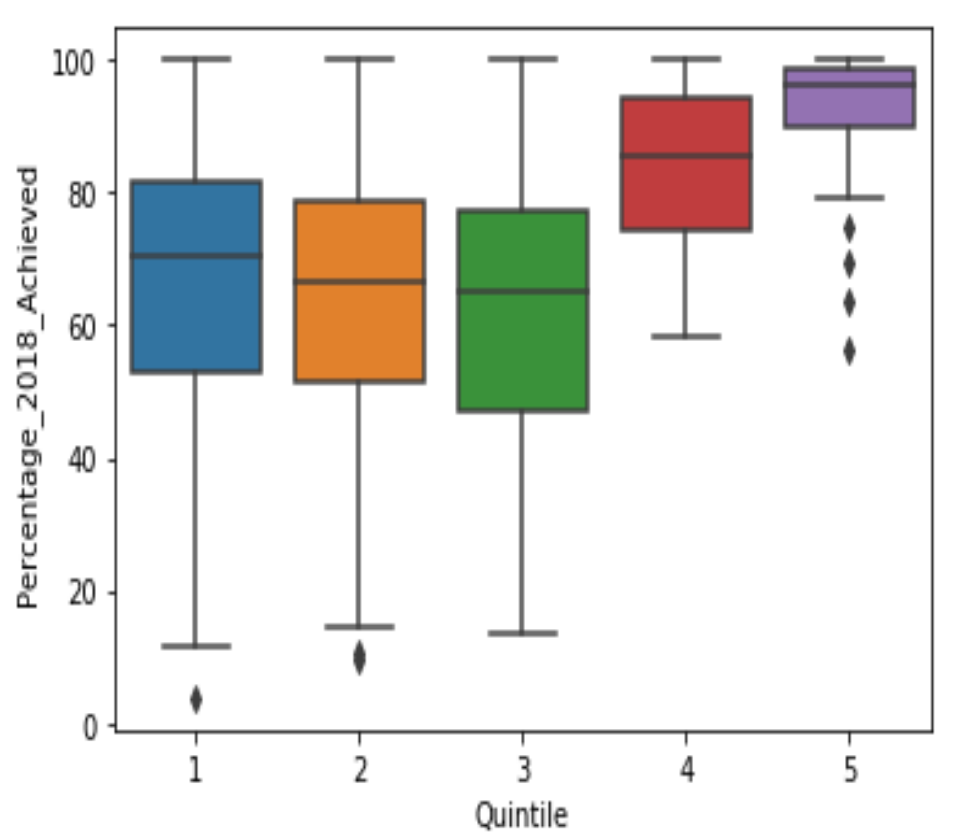




The *Status=1*. If the improvement in the historic data is greater or equal to zero , else the *Status=0,* if the average improvement is less than zero.

- The Status features is used as the target variable for supervised classification model outline in the methodology.
- To quantifying the magnitude of the improvement using linear models, we set the target variable as '*Percentage_achieved_2019*".

# LOWER QUINTILE PASS RATE < UPPER QUINTILE PASS RATE



# PROBLEM 1: IDENTIFY THE MAGNITUDE OF SCHOOL-PERFORMANCE IMPROVEMENT.

- Use supervised learning  linear models. Use the training data to Fit and predict the target variable (*Percent_2019_performance*). Use 2019 data to test the model.

| | Linear Regression | Ridge | Random forest | XGBOOSt |
|---|---|---|---|---|
| **R-Squared** | 54.3 | 54.7 | 57 | 57 |
| **RMSE** | 12.04 | 12.04 | 12.04 | 12.04 |

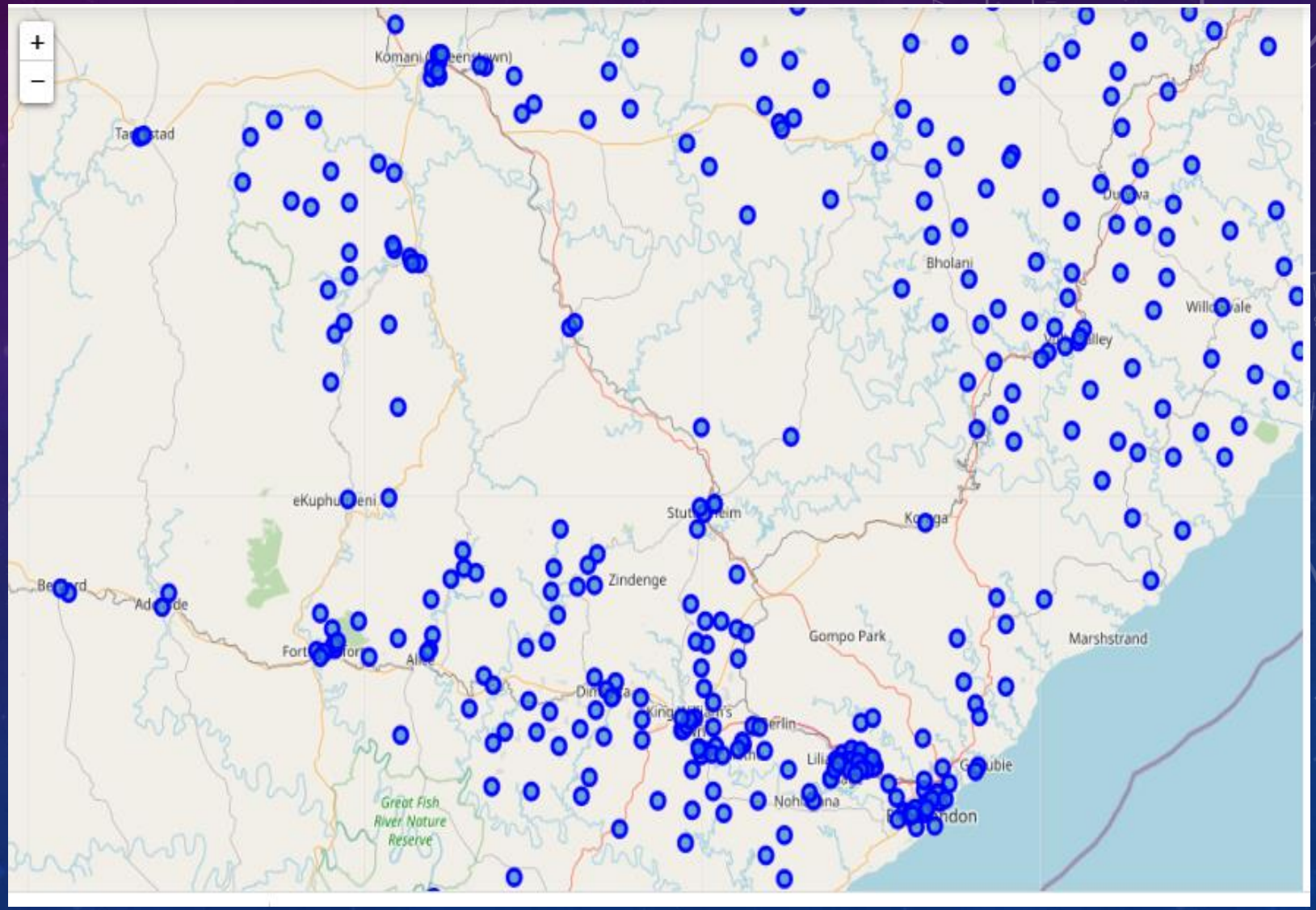
# PROBLEM 2: IDENTIFY THE MAGNITUDE OF SCHOOL-PERFORMANCE IMPROVEMENT.

- Use supervised learning classification models. Use the training data to Fit and predict the target variable (*Status*). Use 2019 data to test the model. NB: Status=0 if improvement <0 and 1 if improvement is >=0.

| | KNN | Decision Tree | SVM | Logistic |
|---|---|---|---|---|
| **F1-score** | 76% | 71% | 75% | 74 |
| **Jaccard index** | 0.75 | 0.71 | 0.75 | 0.76 |

# PROBLEM 3: CLUSTER NEIGHBOURHOOD SCHOOLS

- use K-means algorithm to create cluster on the Map. The 3 cluster were created using the school-performance results from above.



# CONCLUSION

- Purpose of this project was to identify investigate and explore the relationship between the school neighbourhood and school performance.
- The necessary data was sourced and cleaned. Various linear models and classifiers we used to achieve the project objective. The prediction accuracy were discussed and mapped to address the problems statement.
- The results showed that the classifier model had an average of 74% which meant better chance of classifying the improvement, while linear model had an average of 54% which just more than half of the schools we could determine the improvement magnitude for.
- This project serves as starting point for a proactive improvement of school performance in EC.

# FUTURE ACTION

- With the pandemics like COVID19, the use of data to determine and assess the quality of education high school student receive is important. With more that and resources, we recommend the use the above model with teacher's information included and school subject info included, the number of students to write in 2020, so that Tertiary institutions can be aware of what type of first year student they are dealing with.

- More can be done on improving the accuracy of the model, by getting more information about the factors affecting school performance, such the average school attendance, teacher qualifications, yearly intake of new students.