AI Algorithms

11689 AIDI-1002-02

Analysis of classification algorithms and building best solution

Final Report

Purnima Yadav(100800499)

18 December 2020

Marcos Bittencourt

Business Statement

Due to the COVID-19 pandemic, many schools around the globe have been closed and have adapted online learning platforms. With school closures around 188 countries, about 1 billion students are at a risk of failing because they couldn't cope up with the new advancements and school closures due to the spread of COVID-19. Children in poorer houses may not have the resources such as proper internet, personal computers, TV or even radio, which have made them lag in terms of studies. Some of the key points to be noted:

- About 31%(463 million) children do not have internet or broadcast services.
- With the use of television, we could reach to about 62% students globally.
- 16% students can be reached by using radio-based systems worldwide.

Considering all this data, there should be a common platform for remote education in order to reach *all* children.

Rationale Statement

The objective of this project is to classify each student into different categories according to their family income, household condition, access to computer or internet, and therefore build a common platform that is accessible to all children. The vision is to provide education to every student in all the parts of the world.

The techniques that will be used in developing such a system are clustering, k-means classification, multiclass logistic regression and much more.

Problem Statement

The problem is to analyze different algorithms such as clustering or logistic regression in order to predict the best solution, therefore the to evaluate these algorithms there are many metrics available for e.g. Davies-Bouldin Index, Dunn Index for clustering and ROC curves, Confusion Matrix for logistic regression.

The problem is split into two parts- first to identify best algorithm to classify all students into their respective category and second to understand and find out the best solution for online education.

Data Requirements

Since we are using classification algorithms, we need to explore huge datasets that have both numerical and categorical data.

We want a data that contains children from different countries and when their schools were closed, family income, household condition etc. Additionally, we need the data that contains the broadcast services that available to children in different countries. To achieve a high accuracy model, we need to extract features from the input data.

Constraints/Limitations

- Numerical data contains many outliers that can produce inaccurate results if not properly managed.
- Scaling with number of dimensions. With the increase in number of dimensions, a distance-based similarity converges to a constant value. Dimensionality can be reduced using PCA.
- The limitation of Logistic Regression is the assumption of linearity between the dependent and independent variables.
- Logistic Regression requires average or no multicollinearity between independent variables.
- Another limitation in machine learning algorithms is overfitting and underfitting.

Assumptions

Logistic Regression:

- There's is no need for a linear relationship between the independent and dependent variables.
- There is no need for residuals to be normal.
- There is no need to meet the homoskedasticity assumption.

Data Sources

1. This dataset contains the countries where schools have been closed and the date they were closed. It contains columns like CountryCode, CountryName, IfClosedWhen, IncomeLevel, Enrolment of senior secondary, primary children. https://www.worldbank.org/en/data/interactive/2020/03/24/world-bank-education-and-covid-19
    a. country_code : string
    b. short_name: string
    c. currency_unit: euro
    d. region: string
    e. income group: string
2. This dataset shows the number of fixed telephones, fixed broadband, mobile cellular subscriptions per 100 people over 264 countries between 1960 and 2015. https://www.kaggle.com/taniaj/world-telecommunications-data

Test Process to guarantee the quality of the work

Clustering

The following are evaluation metrics for clustering algorithms:

1. Clustering Tendency- Using Hopkins test, a statistical test for spatial randomness of a variable can be used to measure the probability of data points.
2. Number of Clusters- Getting the optimal number of clusters is extremely important in the analysis. If K is too high or low, the algorithm can produce inaccurate results.
3. Clustering Quality- Ideal clustering is characterized by the distance between each cluster and the distance between each data point amongst themselves.
    a. Extrinsic Measures require ground truth labels such as Adjusted Rand Index, Fowlkes-Mallows scores.
    b. Intrinsic Measures do not require ground truth labels such as Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

Logistic Regression

1. Confusion Metrics: a tabular representation of Actual vs Predicted values to find the accuracy of the model.
    a. True Positives (Predicted 1 & Actual 1)
    b. True Negatives (Predicted 0 & Actual 0)
    c. False Positives (Predicted 1 & Actual 0)
    d. False Negatives (Predicted 0 & Actual 1)
2. ROC curve: it's called Receiver Operating Characteristic Curve. The Area Under Curve(AUC) of the ROC provides an overall measure of the fit of the model.
    a. It shows the increase of sensitivity will be accompanies by a decrease in specificity.
    b. The closer the under comes to the 45-degree diagonal of the ROC space, the less accurate the test.

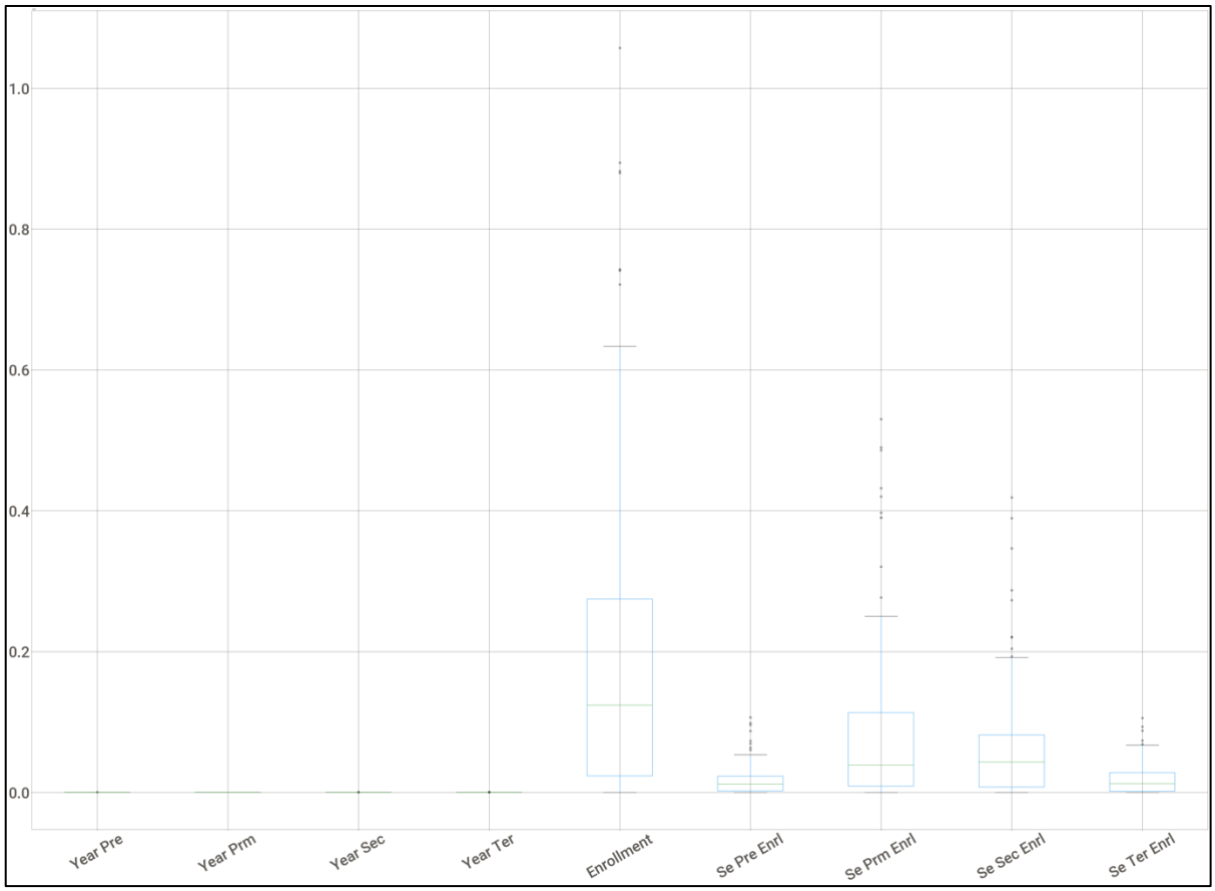
Random Forest

The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

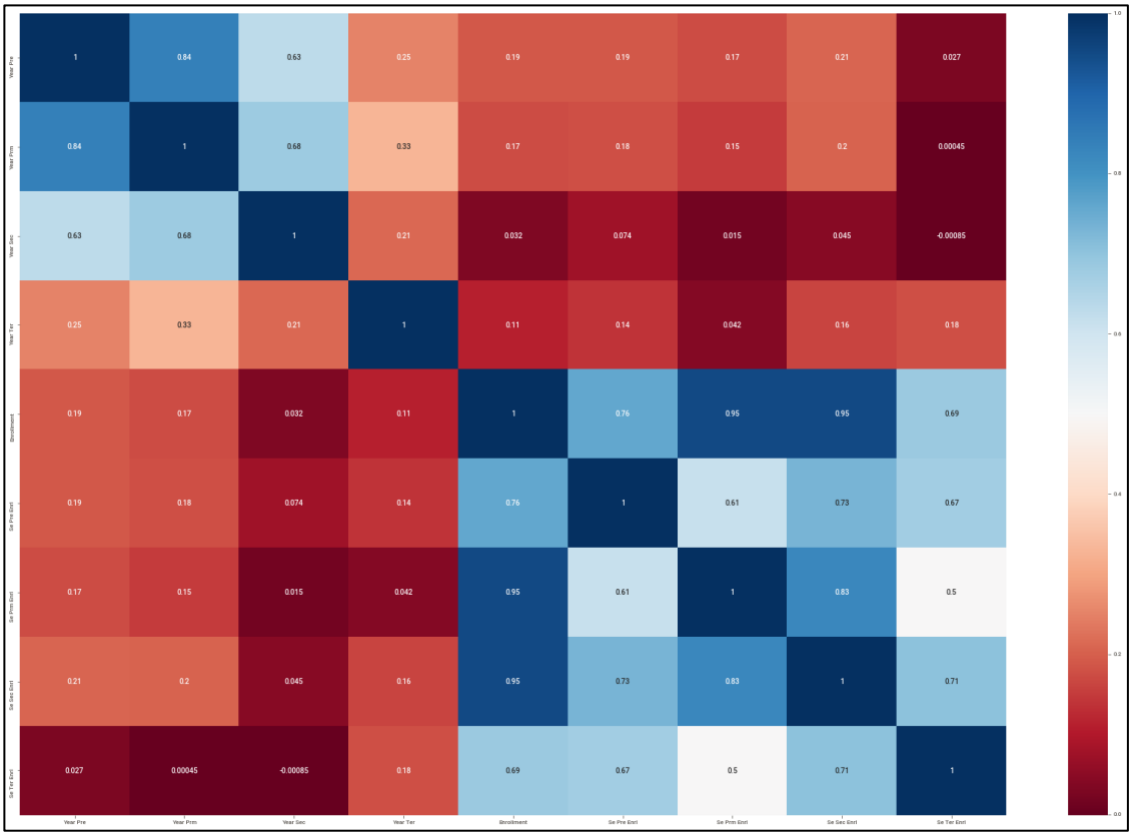
Exploratory Data Analysis

- World Data
  - There were certain missing values in some of the columns, I decided to fil them with (-999).
  - There were no duplicate rows.
  - There were outliers which made significant difference to the data, therefore I decided to remove them.
  - There was some correlation between some of the variables - Year Pre, Se Pre Enrl etc.

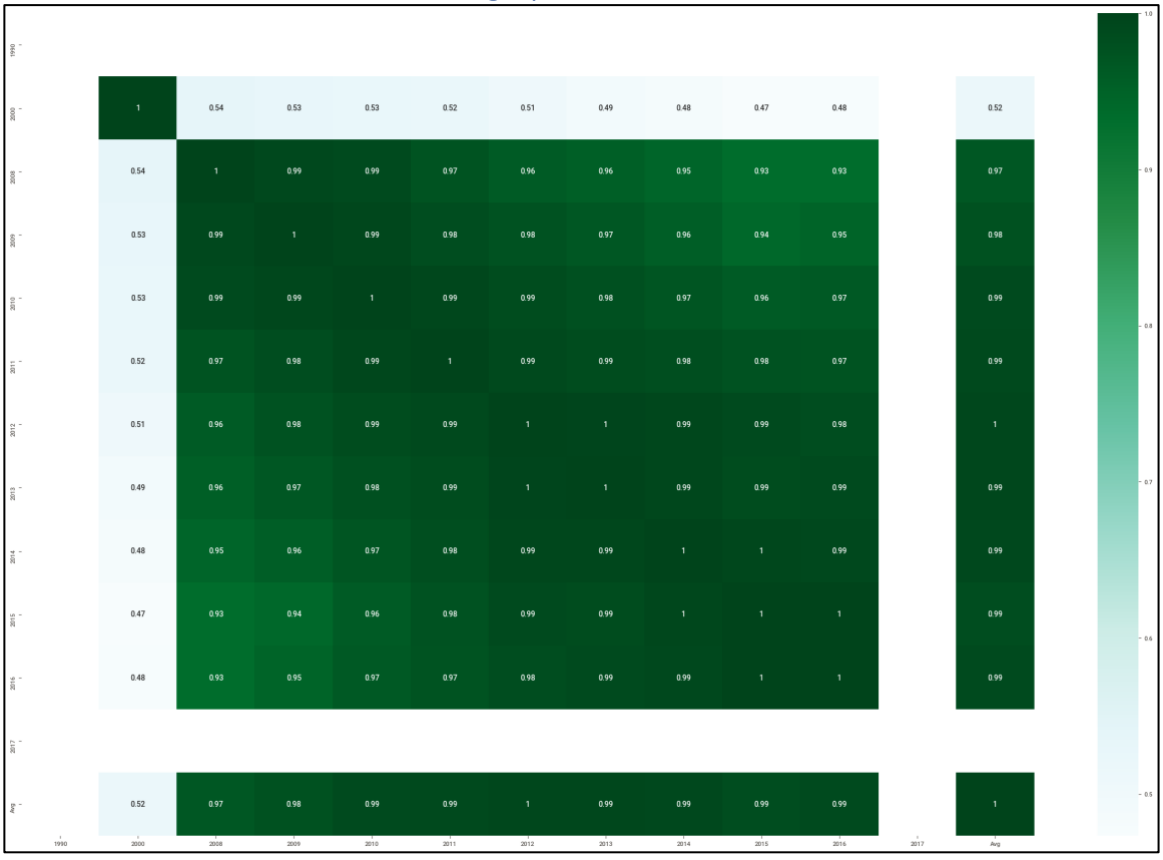
Outlier Analysis



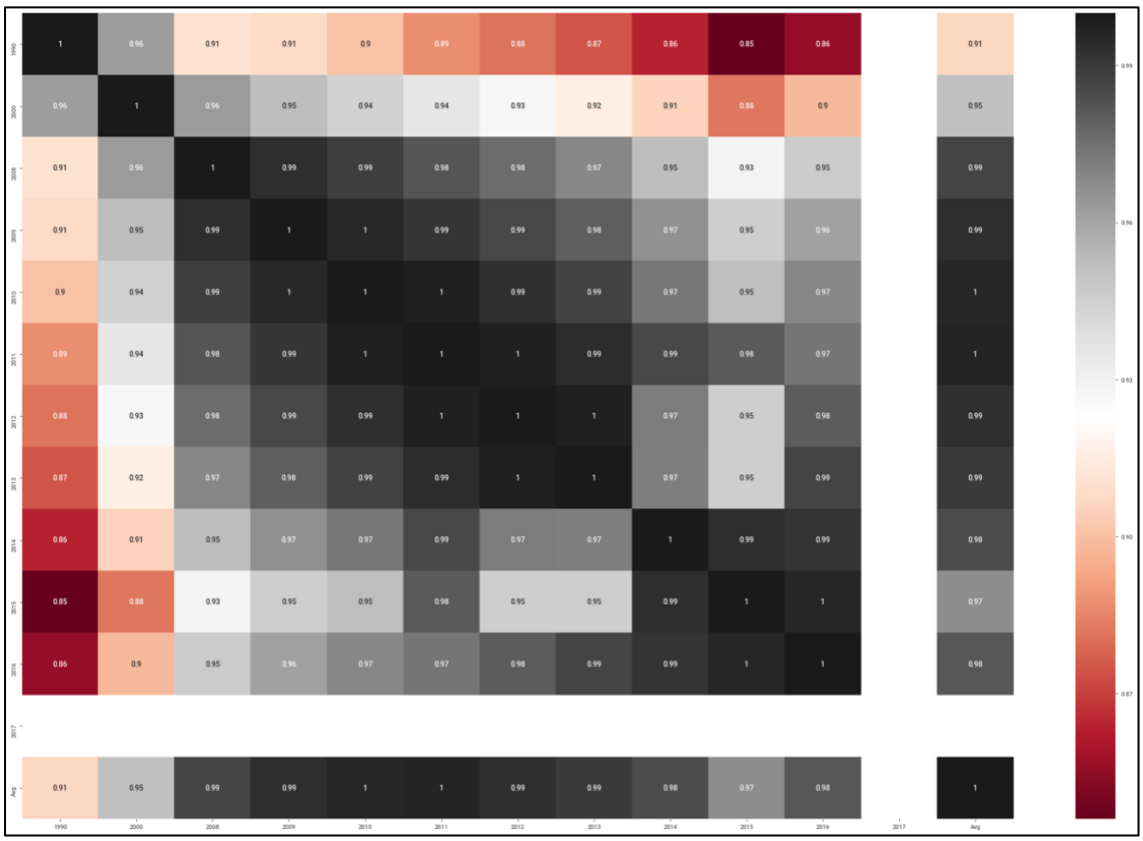
Correlation graph of World Data



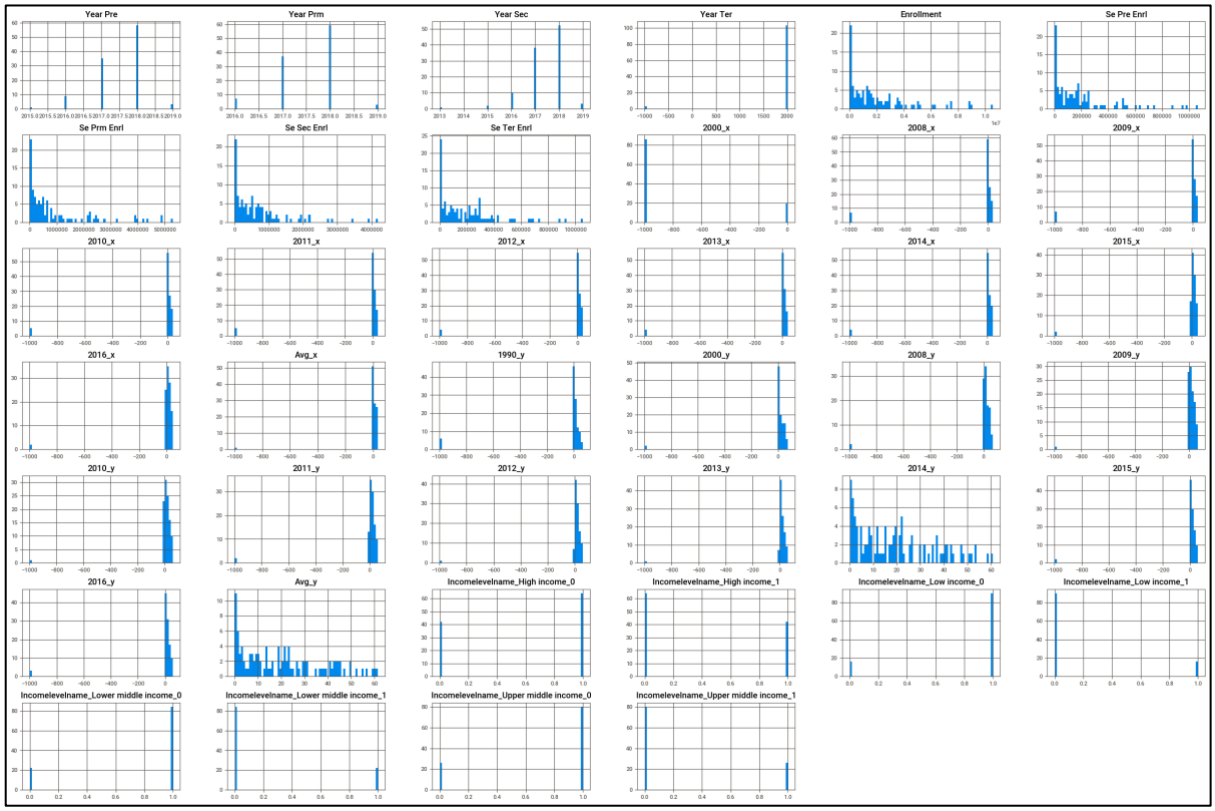
Correlation graph of Broadband data:



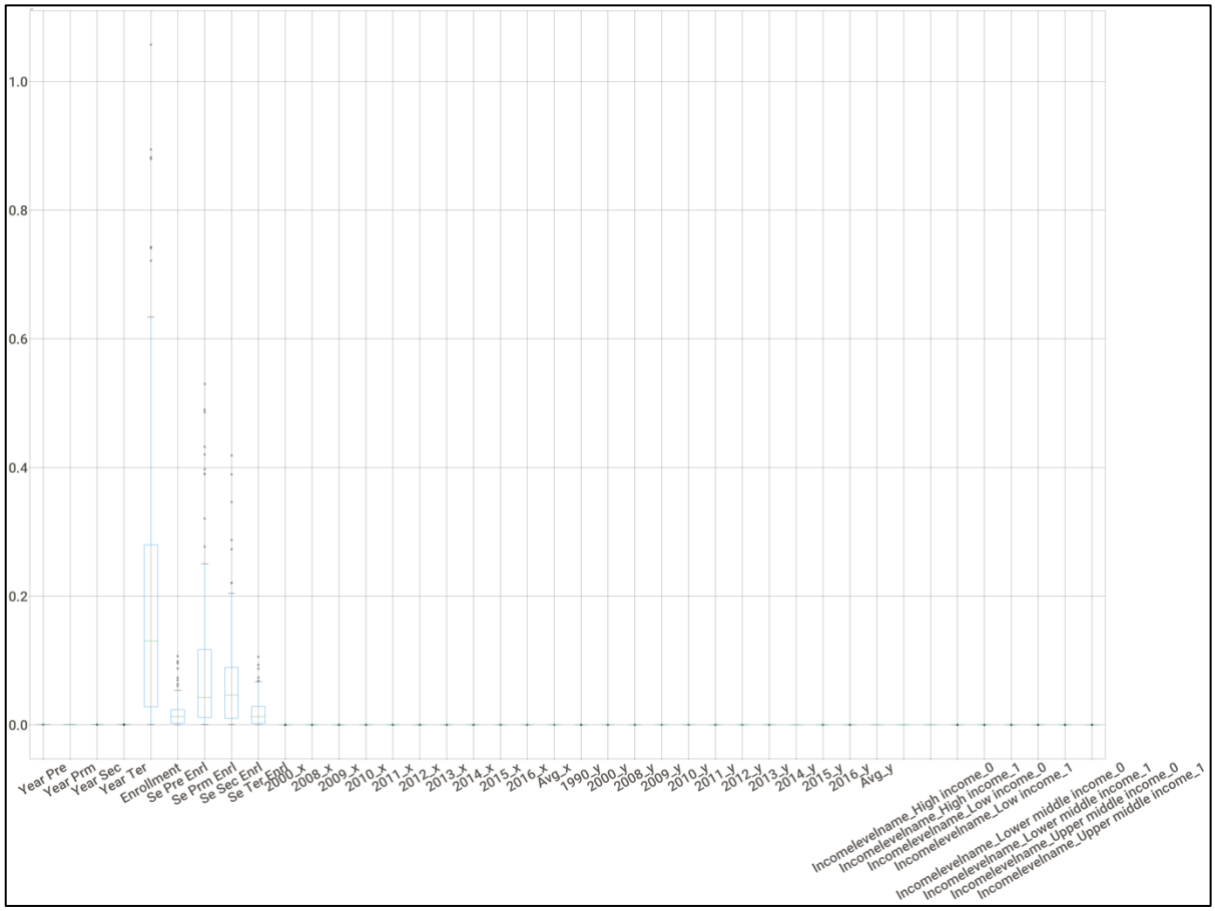
Correlation graph of Telephone data:



Histograms for final data



Outlier Analysis for final data



Sweetviz report shows Incomelevel is highly correlated with total enrolment and primary enrolments.

Data Manipulation

- There are some columns that needed to be manipulated.
    - Categorical columns converted to dummy values.
    - All dates converted to same format.
    - All numerical values rescaled to same format.
    - Date columns converted to true DateTime columns.
    - Used feature scaling and ensemble methods to ensure all data values are in the same format.
    - Firstly, I filled NaN values with (-999) so that all values remain in same float format.

Feature Engineering

- For each dataset, I need to extract best features that can help in predicting accurate results.
- In world_data, I need countries, ifclosedate, incomeLevel, schoolstatus, and number of students enrolled in each year- Se Pre Enrl, Se Prm Enrl, Se Sec Enrl, Se Ter Enrl.
- In telephone and broadband dataset, I want to see average number of connections in each country over the years, for which I will calculate average from each row for every country.

Model Evaluation

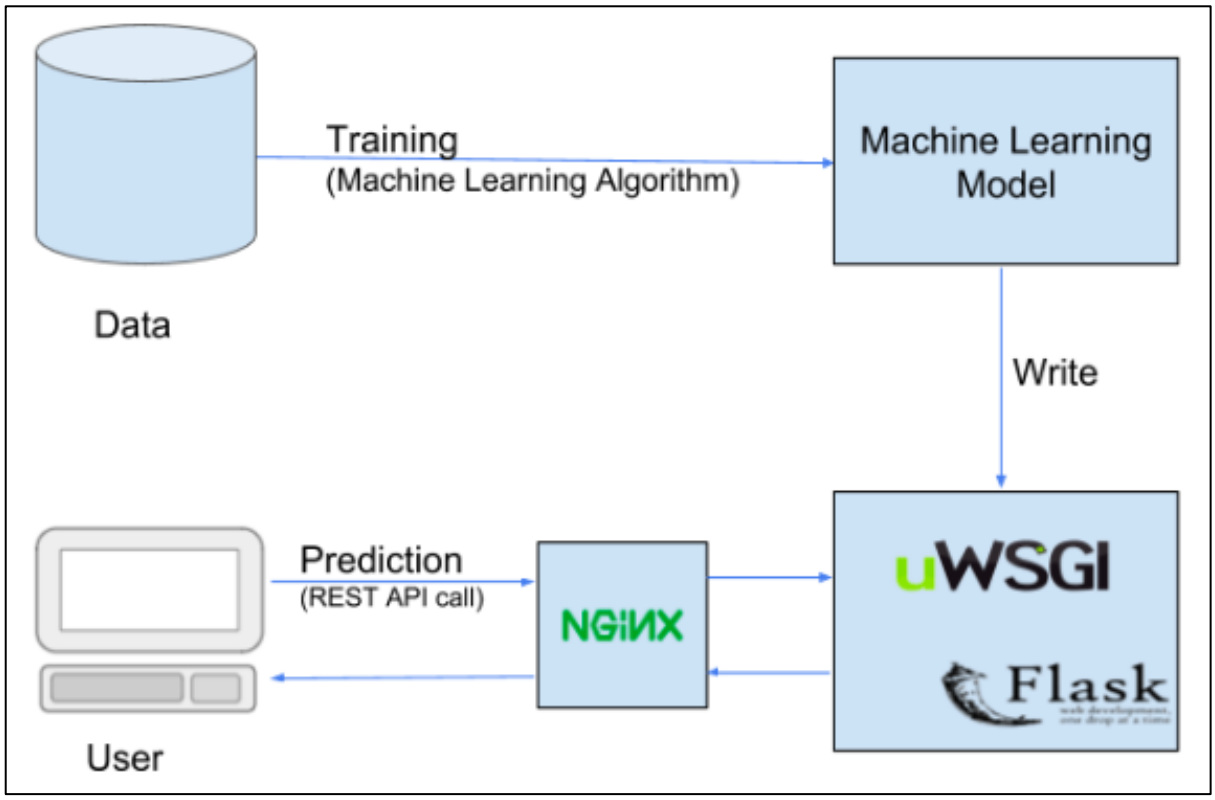
- Target variable is 'lower middle income' type. We need to predict how many enrollments have been in this income type.
- I decided to evaluate these columns and produce best accuracy and hence pick the best model to build the solution:
    - Random Forest Classifier
    - Logistic Regression
    - Decision Tree Classifier
- There are many algorithms to classify students based on their income level.
- Then I first used Random Forest Classifier to train the training Dataset. It showed the accuracy of 95%.
- After that I trained Logistic Regression Model which showed the accuracy of 90%.
- Lastly, I trained Decision Tree Classifier which was able to predict target variable successfully with an accuracy of 100%, therefore it was overfitting. I tried changing the depth to 6,10 etc, but it couldn't improve the overfitting.

Prediction

The model that produced the best result was Random Forest with the highest accuracy. Logistic Regression was a good model with moderate accuracy, and Decision Tree couldn't improve its overfitting model.

Deployment

- Researching into how to use Flask to integrate our ML model on a website and host it on the cloud.
- Preferred cloud services- GCP, AWS.
- I also have experience in HTML, CSS and PHP to work on the backend of the website.



References

1. https://data.unicef.org/topic/education/covid-19/
2. http://pubdocs.worldbank.org/en/451001601558649180/Education-Sector-Brief-September-25.pdf
3. https://www.worldbank.org/en/data/interactive/2020/03/24/world-bank-education-and-covid-19
4. https://www.kaggle.com/taniaj/world-telecommunications-data
5. https://towardsdatascience.com/back-to-basics-assumptions-of-common-machine-learning-models-e43c02325535
6. https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc
7. https://medium.com/@ODSC/assessment-metrics-for-clustering-algorithms-4a902e00d92d
8. https://towardsdatascience.com/how-to-easily-deploy-machine-learning-models-using-flask-b95af8fe34d4
9. https://towardsdatascience.com/improving-random-forest-in-python-part-1-893916666cd