CSE 545 Project Report

# Investigating Quality Education from School Surveys

Team Members: Aakash Deep, Harshit, Rishabh Goel, Shikha Gupta

## Introduction

Educational opportunity for all has always been a defining progressive value, but in recent years, as the economic benefits of education have skyrocketed, living up to this ideal has taken on greater urgency. Studies have shown some surprising facts like 262 million between ages 6-17 were still out of school in 2017, Percentage of trained primary school teachers stagnant at 85 since 2015, Study Hours **highest** for UAE but learning outcomes are **poor,** Study Hours **lowest** for Finland but student performance **high.**

Hence, in our project, we are focussing on the SDG**-4** goal- ensure quality education accessible to all by 2030. To reach any goal, the government makes policy and allocates resources then deploy it. There are multiple companies that conduct surveys to know the current education status of schools. In this project, we are trying to conclude that using feedback/survey data, the government can know in which area it should focus more on, to successfully achieve the SDG goal.

## Background and Motivation

Several studies in the past have been done to improve education quality from analysis of students' and teachers' responses. Multinomial regression[1] analysis was conducted on a sample of 432 students to identify the characteristics of students which make their perception about the quality of higher education dissimilar. "Discovery"-based approaches [2] have produced very positive outcomes in classes taught by exceptional and highly committed educators. However, these studies analyzed a very small set of data and were focused on a few schools.

PISA[8] also published insights from assessment data for each year on how students are performing in different nations. Research published by Charles University in Prague[6] examines how education inequality can be measured from student performance. In order to find clusters of schools similarity based clustering used was motivated by the Chan-Zuckerberg initiative's [7] application towards clustering in RNA-sequencing.

## Data

To investigate feedback importances on SDG 4 indicators, two below datasets were used:

**1. UIS**

As the official statistical agency of UNESCO, the UIS produces a wide range of state-of-the-art databases to fuel the policies and investments needed to transform lives and propel the world towards its development goals.

- Size: 1GB; Number of Features: 3000; Feature Type: Numerical

| SDG_IND | Indicator | Country | TIME | Value |
|---|---|---|---|---|
| READ_LOWERSEC | Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in reading, both sexes (%) | Albania | 2018 | 47.76432 |
| MATH_LOWERSEC_NONNATIVE | Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in mathematics, immigrant background, both sexes (%) | Singapore | 2015 | 93.11794 |

Table 1: Sample Schema of UIS Dataset

**2. PISA**

PISA is the OECD's Programme for International Student Assessment. PISA measures 15-year-olds' ability to use their reading, mathematics, and science knowledge and skills to meet real-life challenges. Along with assessment, it collects responses from students, teachers, and schools on their background.

This survey data represents 31 million students from 79 countries. It includes student information (Representative data), student's answers on various qualitative questions (Feedback Data, ST[0-9]*), and their scores in reading, numeracy, science, etc (Subject Scores, PV*). Assessment is being conducted every 3 years starting from the year 2000.

- Size: 20 GB; Number of features: 1120; Number of observations: > 1 million
- Feature types: IDs, Rank, Numerical

| CNT | CNTSCHID | CNTSTUID | STRATUM | ST005Q01TA | ST006Q01TA | ST011Q09TA | ST006Q04TA | PV7RTML | PV8RTML | PV1MATH |
|---|---|---|---|---|---|---|---|---|---|---|
| JPN | 39200060 | 39203082 | JPN0101 | 2 | 2 | 2 | 1 | 673.665 | 612.816 | 679.482 |
| MNE | 49900039 | 49901439 | MNE0010 | null | null | 2 | 1 | 391.09 | 427.413 | 408.887 |
| GBR | 82600037 | 82603966 | GBR1314 | 1 | 1 | 2 | 2 | 562.021 | 533.748 | 467.462 |

Table 2: Sample Schema of the PISA student Dataset

# Methods

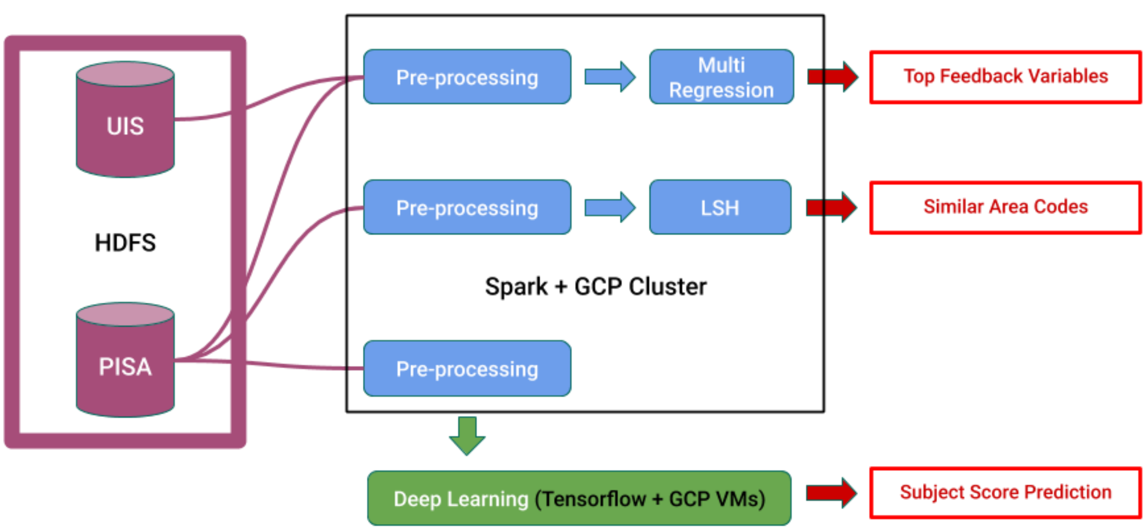



Fig 1: Project Workflow

**1. Multi-Hypothesis Testing -** To complete an SDG goal successfully, it is required to get feedback on existing policy timely, but for an existing policy, which features are more important to get feedback about, can help to track the progress for a given SDG goal. Thus, using Multi-Hypothesis Testing, positively or negatively correlated features can be identified. Here, **PySpark RDD** data pipeline and GCP Cluster **1 Master 2 Worker** are used for computation.

<u>Preprocessing</u>:-

- ➔ From UIS target goal variables were processed out for countries available in PISA [59 countries].
- ➔ From PISA, rank encoded feedback features were extracted for regression.

During preprocessing, null values for features were filled with median and removed from target. Then, Linear regression was performed on each feedback_variable feature to get its corresponding beta value and p-value (Bonferroni corrected) for the two sided significance test of feedback. Most positive and negative 20 correlated features were extracted as a result.

<u>Target SDG goals</u>-

- ➔ Target 1: Goal 4.1.1(b), Achieving at least a minimum proficiency level in reading at the end of lower secondary education
- ➔ Target 2: Goal 4.1.1(d), Achieving at least a minimum proficiency level in mathematics at the end of lower secondary education.

**2. Similarity Search**

To meet the SDG-4 goal, we need to improve schools that are lagging behind by taking help from schools that are excelling. With this motivation, we analyzed schools using similarity search using locality sensitive hashing from data collected from the student surveys. Data that we worked on have responses of students towards their perception of studies, school, and home environment. Since it was survey data, we preprocessed it to remove any null or missing values and replaced them with the value most occurring for that variable in that particular school.

| Shingle | CAN0654 | AUS0307 |
|---|---|---|
| (ST006Q01TA,2,100) | 1 | 0 |
| (ST006Q01TA,1,100) | 0 | 1 |
| (ST007Q01TA,1,80) | 1 | 0 |

Table 3: Characteristic Matrix

| CAN0654 | AUS0307 |
|---|---|
| minhash(batch-1) | minhash(batch-1) |
| minhash(batch-2) | minhash(batch-2) |
| ……. | …….. |
| minhash(batch-60) | minhash(batch-60) |

Table 4: Signature Matrix

The next step to apply LSH was to create shingles to build our characteristic matrix. We wanted to keep as much information as possible from survey data in our characteristic matrix so we defined a shingle as <questionID> + <option selected> + < percentage of students who selected option in a particular school>. We kept the percentage in buckets of size 20. For e.g, if 67% of students from CAN0654 school selected option 1 for question ST005Q01TA we added shingle "ST005Q01TA" + "1" + "80" for document CAN0654 as 67 falls in bucked 60-80. After creating the characteristic matrix, getting a signature matrix was easy. Finally, LSH is used to find candidate similar schools from a signature matrix. All analyses were conducted on GCP clusters using **HDFS (replication=2) and Pyspark** with 1-master(64 GB) and 2 worker nodes (32 GB) configuration with image Version: 1.4 (Debian 9, Hadoop 2.9, Spark 2.4).

1. Preprocessing
- $(areadID, fv_1, fv_2, \cdots)$
2. Create shingle
- $(areaID, fv_1^{20}, fv_1^{40}, ..., fv_2^{20}, fv_2^{40}, \cdots)$
3. Characteristic matrix
- $(fv_i^{b}, areaID_1, areaID_2, \cdots)$
4. Signature matrix
- $(h_x, areaID_1, areaID_2, \cdots)$
5. LSH
- $(\{areadID_a, areadID_b\}, \{areadID_z, areaID_y\}, ...)$

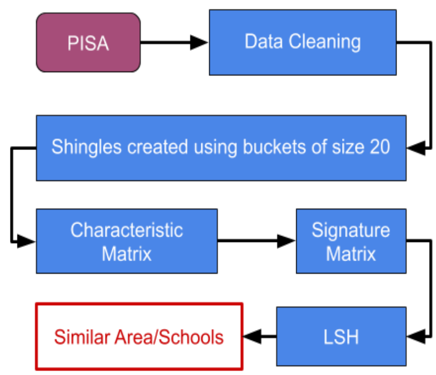



Fig 2: Steps for similarity search

**3. Deep Learning**

In search of further inferences on feedback from schools, performances of each student were predicted based on the feedback responses from them. A Deep Learning based approach was incorporated to investigate this inference. Different models were trained to obtain the best predictive performances.

PISA data was pre-processed on a distributed architecture with **PySpark Dataframe and HDFS** (on GCP cluster 1 Master - 2 Worker) to merge data available for different years. Then feedback (ST[0-9]*) and scores (PV[0-9]*) features with zeros-variance were removed and null values were filled with median for feedback and mean with performances. This is so because feedbacks were rank variables and scores were numerical variables. All variables were then standardized before feeding into the network.

Distributed training to map feedback variables to scores was done using **Tensorflow** 2x on GCP VMs with 2 NVIDIA T4. Tensorflow provided synchronous all reduce with below available strategies

- `tf.distribute.MirroredStrategy` (1 node multi GPUs)
- `tf.distribute.experimental.MultiWorkerMirroredStrategy` (multi-node multi GPU, nodes communicate via RPCs).

# Evaluation/Results

All methods were processed independently and obtained results as summarized below.

**1. Multi-Hypothesis Testing -**

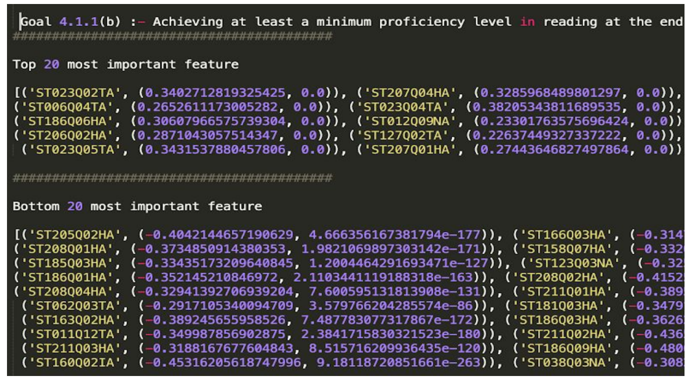
Top 20 positively and negatively correlated features for a given SDG goal were obtained.

For Goal 4.1.1(b), some positive correlated questions are "Which language do you usually speak with: My best friend", "Agree: I feel bad seeing other students bullied" and some negative correlated questions are "Thinking about yourself and how you normally feel: how often do you feel as described Proud?", "Think about your school, how true: It seems that students are competing with each other."

For Goal 4.1.1(d), some positive correlated questions are "When you have to read, does the teacher ask you to: Write a text related to what you have read", "Have you ever repeated a <grade>?" and some negative correlated questions are "Taught at school: How to detect phishing or spam emails", "How true for you: My goal is to completely master the material presented in my classes."

From the result, it is evident that as beta values and p values are quite high, so can say feedback variables are well correlated with SDG-4 for goal scores.

| Feedback Features | Beta | P-value |
|---|---|---|
| ST154Q10HA | 0.415 | 0.0 |
| ST207Q04HA | 0.29350 | 0.0 |
| ST166Q03HA | -0.23453 | 1.24e-55 |
| ST205Q02HA | -0.36735 | 6.4e-114 |



```
Goal 4.1.1(b) :- Achieving at least a minimum proficiency level in reading at the end
#############################################

Top 20 most important feature

[('ST023Q02TA', (0.3402712819325425, 0.0)), ('ST207Q04HA', (0.3285968489801297, 0.0)),
('ST006Q04TA', (0.2652611173005282, 0.0)), ('ST023Q04TA', (0.38205343811689535, 0.0)),
('ST186Q06HA', (0.30607966575739304, 0.0)), ('ST012Q09NA', (0.23301763575696424, 0.0))
('ST206Q02HA', (0.2871043057514347, 0.0)), ('ST127Q02TA', (0.22637449327337222, 0.0)),
 ('ST023Q05TA', (0.3431537880457806, 0.0)), ('ST207Q01HA', (0.27443646827497864, 0.0))

#############################################

Bottom 20 most important feature

[('ST205Q02HA', (-0.4042144657190629, 4.666356167381794e-177)), ('ST166Q03HA', (-0.314
('ST208Q01HA', (-0.3734850914380353, 1.9821069897303142e-171)), ('ST158Q07HA', (-0.332
('ST185Q03HA', (-0.33435173209640845, 1.2004464291693471e-127)), ('ST123Q03NA', (-0.32
('ST186Q01HA', (-0.352145210846972, 2.1103441119188318e-163)), ('ST208Q02HA', (-0.4152
('ST208Q04HA', (-0.32941392706939204, 7.600595131813908e-131)), ('ST211Q01HA', (-0.389
 ('ST062Q03TA', (-0.2917105340094709, 3.579766204285574e-86)), ('ST181Q03HA', (-0.3479
 ('ST163Q02HA', (-0.389245655958526, 7.487783077317867e-172)), ('ST186Q03HA', (-0.3626
 ('ST011Q12TA', (-0.349987856902875, 2.3841715830321523e-180)), ('ST211Q02HA', (-0.436
 ('ST211Q03HA', (-0.3188167677604843, 8.515716209936435e-120)), ('ST186Q09HA', (-0.480
 ('ST160Q02IA', (-0.45316205618747996, 9.18118720851661e-263)), ('ST038Q03NA', (-0.308
```

Fig 3: Beta and P-value for top correlated features; Raw results: Top features for goal 4.1.1(b) - Achieving at least a minimum proficiency level in reading at the end of lower secondary education

## 2. Similarity Search

We ran LSH for different values of band sizes and number of bands for 80% similarity and found band size = 5 and the number of bands = 12 produces optimal results. We got around 33% similar schools of the total pair of schools.

```
('"CHE0104"', '"PHL0007"')('"BGR0003"', '"TAP0216"')('"DNK0404"', '"ESP9010"')('"FRA0204"', '"MEX0011"')
('"CHL0203"', '"GBR1110"')('"CAN0547"', '"UKR0021"')('"MDA0003"', '"UKR0033"')('"SVN0007"', '"SVN0504"')
('"CZE1010"', '"QAZ0102"')('"MDA0007"', '"SRB0101"')('"ESP9006"', '"RUS2323"')('"ESP9021"', '"QRT8787"')
('"CAN0103"', '"LUX0103"')('"ARE0767"', '"NLD0008"')('"ESP9039"', '"GBR1109"')('"FIN0002"', '"KAZ0203"')
('"CAN0545"', '"MDA0017"')('"BIH0002"', '"CAN0657"')('"BRA0103"', '"CAN0986"')('"MDA0007"', '"SVN0102"')
('"IRL0002"', '"ISL0003"')('"CAN0982"', '"DOM0002"')('"CHL0204"', '"TUR0319"')('"CAN0434"', '"CAN0765"')
('"COL0412"', '"TUR0229"')('"ISL0015"', '"LUX0103"')('"IDN0105"', '"MDA0016"')('"ESP1430"', '"FRA0213"')
('"ESP9028"', '"MAR0003"')('"GEO0002"', '"LTU0116"')('"CAN1097"', '"ISL0002"')('"GBR1113"', '"MEX0005"')
```

Fig 4: Raw results: Similar areas

We found that schools from developing countries and developed countries are similar to schools in Finland. Hence, we believe progress for achieving the SDG-4 goal can be considerably improved if the former schools can take notes from the latter[8] on what they are doing right.

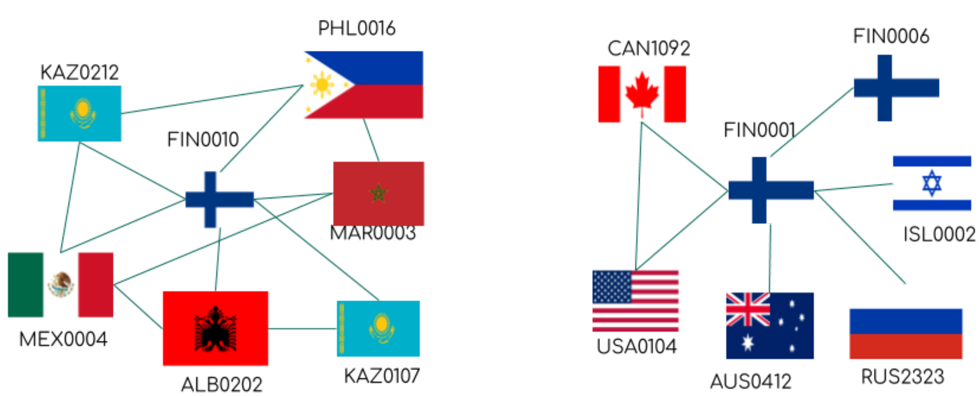



Fig 5: Area Clusters obtained after Similarity Search

**3. Deep Learning**

A three layer (layer 1 with 1024 neurons, layer 2 with 1024 neurons and layer 3 with 512 neurons) architectured was implemented after hyperparameter tuning. Training was done for 15 epochs with each epoch taking a little over a minute (little over 5 minutes for non-distributed architecture). Overfitting was tackled via batch normalization, still, low precision was obtained.

- Mean Absolute Error: 0.609 [standardized outputs].
- Subject PV1MATH score predicted for a sample was 577 with actual being 463.

**Inferences:** Low Precision; Scores cannot be accurately predicted only with feedback information.

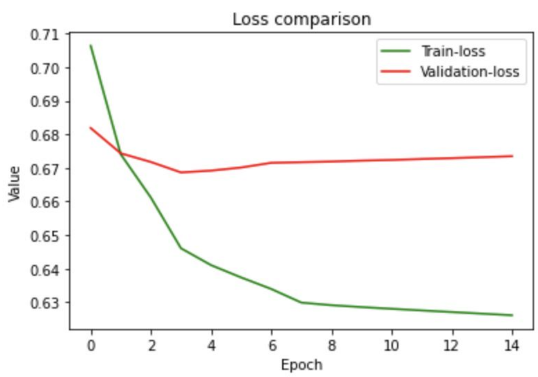



Fig 6: Training history

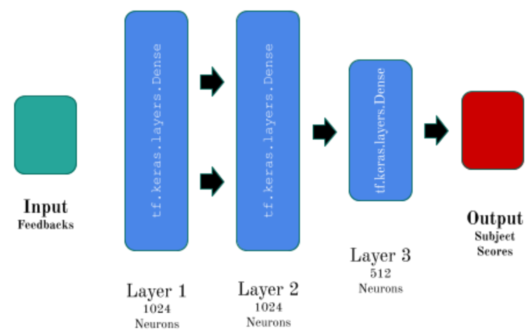



Fig 7: Network Design

# Conclusion

With the results of the above mentioned three concepts, it can be seen that feedback shows considerable importance in the assessment of education quality goals. Cross-nation similar area codes found can help to opt for similar policies, such as Amount of teaching Hours, Degree of interaction between students, etc. Inferences on feedback can help in policy designing to achieve the SDG-4 goals. For future work, we expect that with more semantic analysis on feedback, higher efficient inferences regarding policy improvement can be generated.

# References


1. Determinants of education quality: what makes students' perception different?
2. Quality of teaching and quality of education
3. UIS Statistics
4. Effects of students' and schools' characteristics on math achievement: findings from PISA 2006
5. Using PISA to examine educational inequality
6. Chan-Zuckerberg initiative used LSH for scalable clustering in single-cell RNA sequencing
7. Publications - PISA
8. Why Are Finland's Schools Successful? | Innovation
9. Education Policy Ideas for Progressives in 2018