CSE 545 (Big Data Analytics)
Project Presentation

**Authors : Rishabh Goel , Harshit
Aakash Deep, Shikha Gupta**

Investigating Quality Education from school surveys

Team Spark Champs

Contents

★ Introduction

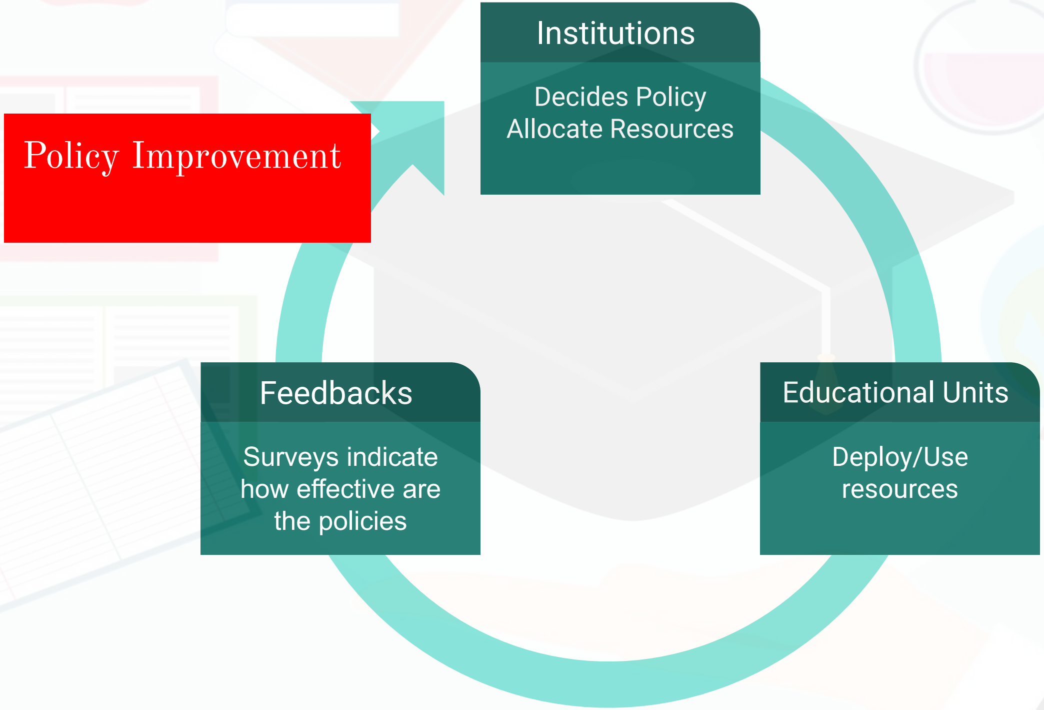
★ Past Work

★ Data

★ Methods and Evaluation

★ Conclusion

Introduction



Institutions

Decides Policy
Allocate Resources

Educational Units

Deploy/Use
resources

Feedbacks

Surveys indicate
how effective are
the policies

Policy Improvement

Why Care? SDG 4- Quality Education

- **SDG 4** - ensure quality education accessible to all by 2030

- 262 million between age 6-17 were still out of school in 2017.
- Percentage of trained primary school teachers stagnant at 85 since 2015.

- Ground reality of government measures can be understood by student and teacher **feedbacks**.
- Study Hours **highest** for UAE but learning outcomes are **poor**
- Study Hours **lowest** for Finland but student performance **high**

Image Credits : https://leverageedu.com/blog/best-education-system-in-the-world/

Why Big Data?

- Process large data (~20 GB)
- Use techniques to analyse and obtain inferenes
  - Similarity Analysis
  - Multi-Hyp Testing
  - Large Scale Machine Learning
- Need to use data pipeline to distribute the tasks on nodes and aggregate results
  - HDFS
  - PySpark
  - Tensorflow

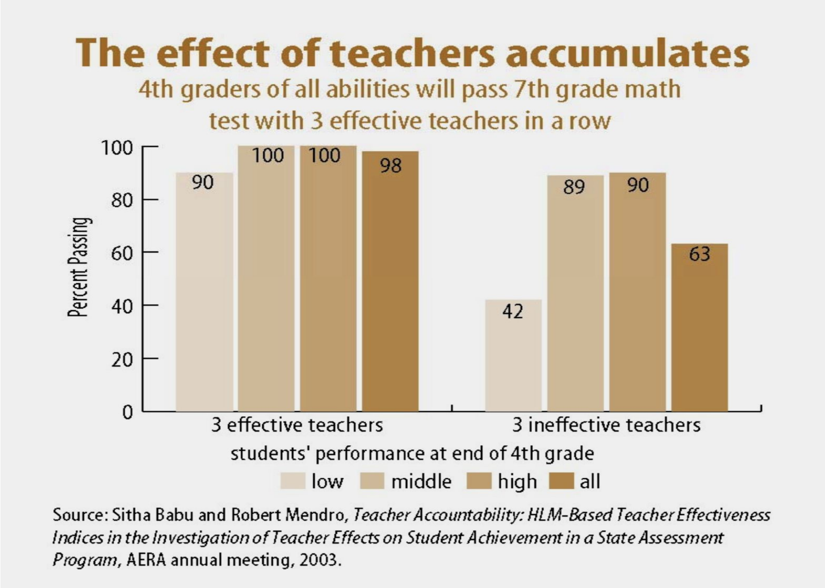
Keywords

- SDG Goal 4.1.1 ⇒ Achieving target proficiency level in Literacy and Numeracy
- Feedback features ⇒ Students' responses on survey questions [done by PISA]
- Subject Scores ⇒ Students' scores in different subjects [done by PISA]

Background

- Studies [Ref 2.] shows that school and family have only very little impact on academic achievement among pupils from disadvantaged backgrounds.

But Teachers are the most influential factor in student learning.



The effect of teachers accumulates

4th graders of all abilities will pass 7th grade math test with 3 effective teachers in a row

Source: Sitha Babu and Robert Mendro, *Teacher Accountability: HLM-Based Teacher Effectiveness Indices in the Investigation of Teacher Effects on Student Achievement in a State Assessment Program*, AERA annual meeting, 2003.

Image Credits: References 2

Past Work

- Multinomial regression analysis was conducted[Ref1] to identify characteristics of students
    - Students for scholarship
    - Extracurricular activities
    - Parents’ education
    - University they study in

  which make perception about quality of higher education dissimilar.

- Studies [Ref2] shows that “Discovery”-based approaches have produced very positive outcomes in classes taught by exceptional and highly committed educators.
    - However,considerable time and energy is required on the part of the individuals, average teacher is undoubtedly not in a position to contribute.

Data

- PISA [~20 GB] (https://www.oecd.org/pisa/)
  - Reading, mathematics and scientific literacy scores of more than 710,000 15 - year old students representing 31 million student from 79 countries. Along with general survey from teachers, parents school principal and students every 3 year since 2000
  - Number of features > 1120, observations > 1 million
  - Representative Data, Feedback Data (ST*), Subject Scores (PV*)

- UNESCO SDG [1 GB] (UIS Statistics)
  - Country wise quality indicators per year (1970-2019) (1GB)
  - Number of Features: >3000

Methods

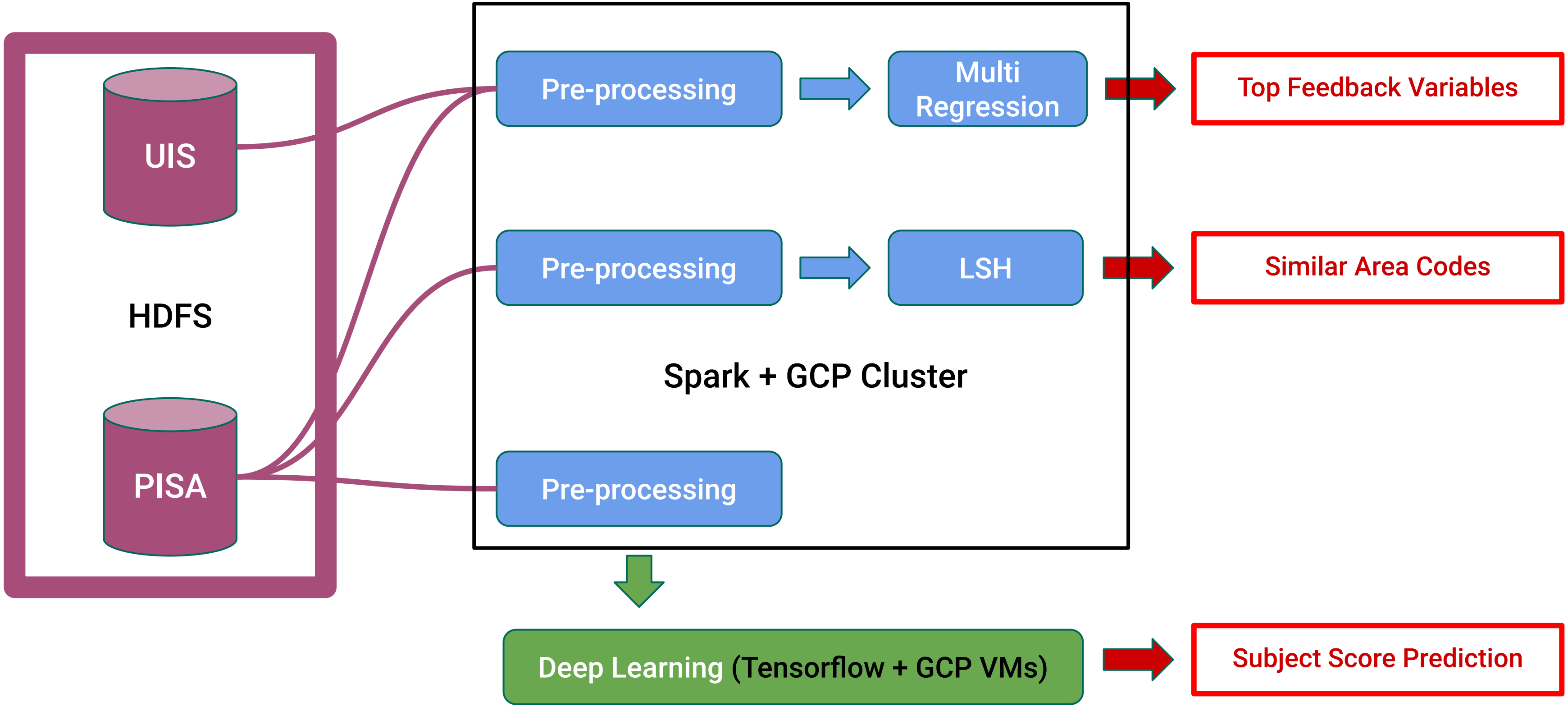


Fig: Code Analysis Pipeline

Methods - Hypothesis Testing

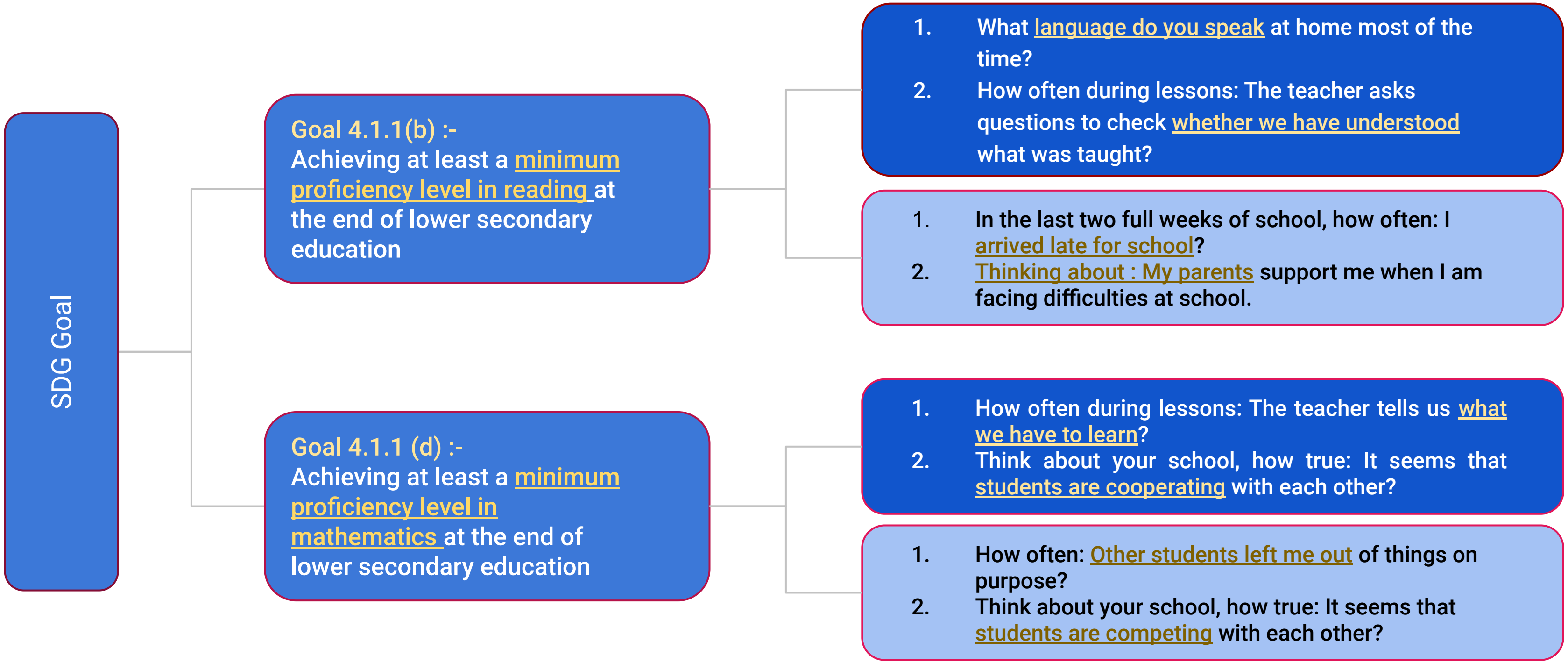
Task :- Find Correlated Feedbacks for the success of a SDG Goal

Approach :-

- Spark RDD, data ⇒ (Y = goal score, X = features) for each country
- 2 Target SDG goals
- Multi-linear regression ⇒ beta value and p value for each feature
- Top positively and negatively correlated features for each SDG goal

SDG 4.1.1 (b)	Feature	Beta Value	P value
1	ST153Q10HA	0.41500	0.0
2	ST207Q04HA	0.29350	0.0
-2	ST166Q03HA	-0.23453	1.24e-55
-1	ST205Q02HA	-0.36735	6.4e-114

Results - Hypothesis Testing



SDG Goal

Goal 4.1.1(b) :-
Achieving at least a minimum proficiency level in reading at the end of lower secondary education

1. What language do you speak at home most of the time?
2. How often during lessons: The teacher asks questions to check whether we have understood what was taught?

1. In the last two full weeks of school, how often: I arrived late for school?
2. Thinking about : My parents support me when I am facing difficulties at school.

Goal 4.1.1 (d) :-
Achieving at least a minimum proficiency level in mathematics at the end of lower secondary education

1. How often during lessons: The teacher tells us what we have to learn?
2. Think about your school, how true: It seems that students are cooperating with each other?

1. How often: Other students left me out of things on purpose?
2. Think about your school, how true: It seems that students are competing with each other?

Result: - **Most** or least correlated features for a given goal

Methods - Similarity Search

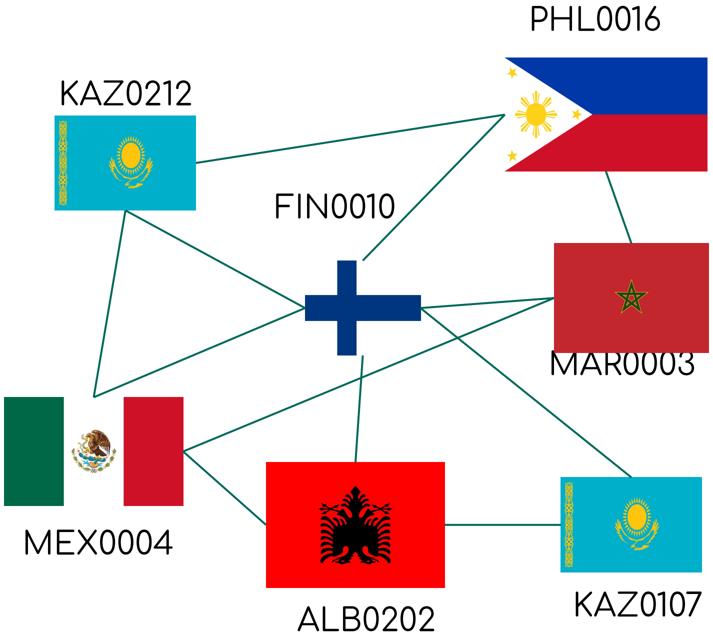
Clustering and Similarity Search using LSH

Task :- Find similar schools and areas from survey feedback collected from students

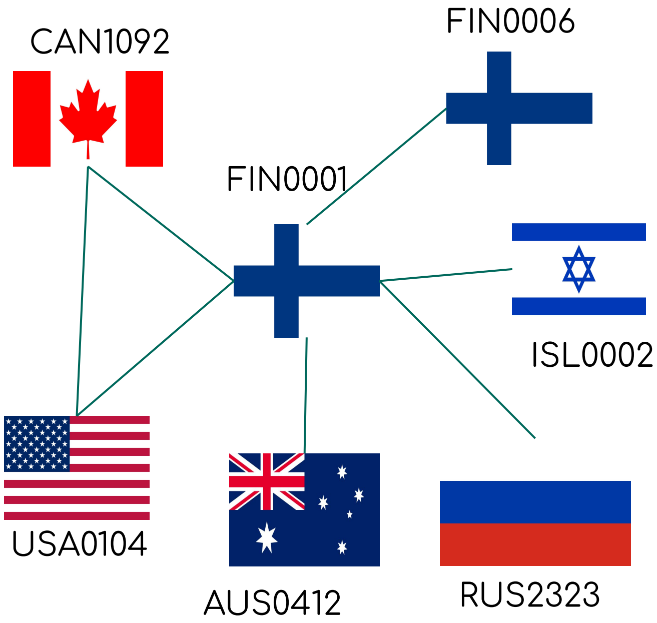
Approach :-

- Characteristic matrix ⇐ buckets [0-20%, 21-40%, 41-60%, 61-80%, 81-100%] for feedback responses
- Signature matrix (~1500 columns, ~60 rows) ⇐ Characteristic matrix (~1500 columns ~9000 rows).
- Analysis on band values- (b=5,r=12), (b=8,r=8), (b=12,r=5) [Sim > 0.8]
- Locality Sensitive Hashing ⇒ Similar area codes / schools (JSON)

Insights - Similarity Search



Schools from some developing countries matched with some Finland area schools



Schools from some developed countries matched with some Finland area schools

Image Credits : Google Images

Methods - Deep Learning

Task :- Predict Performances in Subject based on students feed_back data

Approach :-

- Distributed Preprocessing (using Spark Dataframes and HDFS, GCP Cluster 1 Master 2 Worker)
- Distributed training [TF2.0 Synchronous All Reduce]
  - `tf.distribute.MirroredStrategy` - single node multi-gpu, each gpu have their parameter and update each other synchronously
    - GCP Compute Engine VMs with 2 Nvidia T4
    - Training time ~ little more than a minute per epoch
  - `tf.distribute.experimental.MultiWorkerMirroredStrategy` - multi node multi-gpu, each node communicate with each other in round-robin using RPCs

Insights - Deep Learning

Design :-

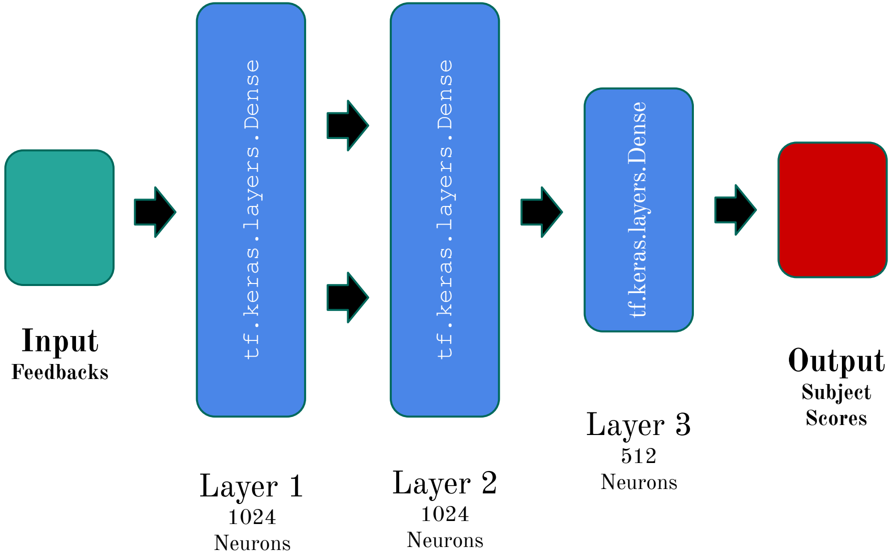


Fig :- Network architecture with 3 hidden layers

Inferences:

- MAE for prediction 0.609 [standardized]
- PV1MATH score for a sample
    - 463 (Actual)
    - 577 (Predicted)
- Low Precision

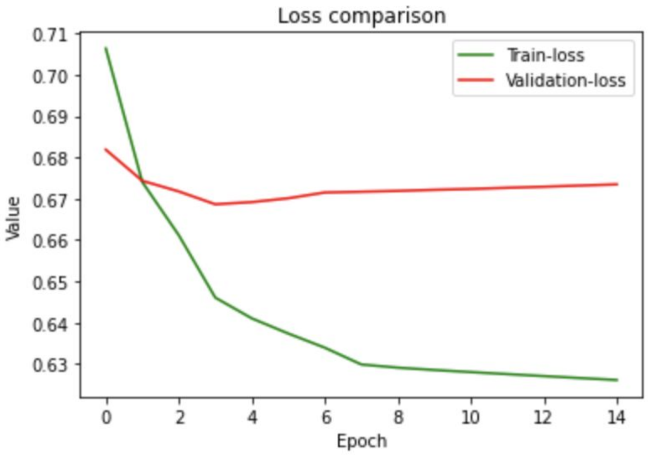


Fig :- Training History

Conclusion

- With our results ⇒ Feedbacks show a considerable importance in education quality
- Cross-nation similar area codes found can help to decide similar policies, such as
  - Amount of teaching Hours
  - Degree of interaction between students
- Institutions can decide to spend depending on type of students in particular region.
- Inference on feedbacks can help achieve SDG goals

- Future Work - more semantic analysis on feedbacks can produce higher efficient inferences regarding policy improvement.

Image Credits: https://myventurepad.com/software-education-something-constantly-revised/

References

1. https://www.tandfonline.com/doi/full/10.1080/23265507.2016.1155167
2. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.498.6274&rep=rep1&type=pdf
3. http://www.oecd.org/pisa/aboutpisa/
4. http://data.uis.unesco.org/#