Literature Review:

1. **Introduction:**

The importance of this project lay in conducting a comprehensive data analysis of Sudanese schools by focusing on critical aspects such as facilities, student number and gender, and teachers' numbers ...etc. The findings from this analysis carry the potential to guide stakeholders/government to address specific challenges faced by schools for optimal impact. Below are the most important research questions that needed to be addressed and analyzed.

- What are the key facilities available in Sudanese schools?
- How do these facilities vary across different regions or types of schools within Sudan?
- Are there significant disparities in the availability of facilities between urban and rural schools?
- How do schools in economically disadvantaged areas compare to those in more affluent regions in terms of educational facilities?
- Are there specific facilities that are consistently lacking or in need of improvement across the country?
- Is there gender-based differences in the availability of facilities?

2. **Related work**

This part will focus on related literature review that addresses the use of artificial intelligence and machine learning in sustainable development goals (SDGs). In addition to that, papers that addressed the working with educational dataset using data mining and machine learning as well.

Paper [2][15][19] presents the use of machine learning for SDGs. In Paper [2], a machine learning approach is employed to prioritize Sustainable Development Goals (SDGs). The results reveal that SDG3 "Good health and well-being", SDG4 "Quality education" and SDG7 "Affordable and clean energy" exhibit the highest synergy and can be done together. The findings are valuable for decision-makers and enables them to implement strategic initiatives and allocate resources more effectively by prioritizing goals that demonstrate significant synergy. Paper [15] reviews the applications of machine learning in many sectors like agriculture, education, greenhouse gas reduction, and environmental tax reform. The paper highlights the role of machine learning and data mining in supporting the United Nations' SDGs by providing decision makers with insights, predictions and objective recommendations to achieve it for a better life. Similarly, paper [19] addressed the role of artificial intelligence in achieving SDGs.

On the other hand, some papers used machine learning techniques in educational data to predict student achievement [3] while others used machine learning to evaluate facilities condition in school and also how facilities can affect the achievement. Paper [17] uses

unsupervised machine learning approach to evaluate sports facilities condition in primary school. The paper concludes that the proposed method effectively differentiates sports facilities in primary schools. In addition to that, primary schools with more grades of students are equipped with more types and sizes of sports facilities.

Research [11] addressed data analysis report for New York State school facilities and student health, achievement and attendance. Similarly, [21] shows the relationship between the condition of school facilities and certain educational outcomes, particularly in rural public high schools in Texas.

4.   **Conclusion:**

To conclude that, it is clear that the artificial intelligent and machine learning have a good impact when using with SDGs and as many papers addresses the relation between facilities and student achievement this research will be valuable as it analyze the school dataset in Sudan with focusing in classifying them according to their facilities to get more detailed information and to help decision makers in finding a suitable solution for each class "group".

5.   **Citations:**

[1] Asadikia, A., Rajabifard, A., & Kalantari, M. (2021). Systematic prioritisation of SDGs: Machine learning approach. World Development, 140. https://doi.org/10.1016/j.worlddev.2020.105269

[2] Tunmibi, S., Okhakhu, D. O., & Okhakhu, D. (n.d.). Machine Learning for Sustainable Development. https://www.researchgate.net/publication/362697825

[3] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. In Nature Communications (Vol. 11, Issue 1). Nature Research. https://doi.org/10.1038/s41467-019-14108-y

[4] YILDIZ, M., & BÖREKCİ, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. Journal of Educational Technology and Online Learning, 3(3), 372–392. https://doi.org/10.31681/jetol.773206

[5] Xia, J., Wang, J., Chen, H., Zhuang, J., Cao, Z., & Chen, P. (2022). An unsupervised machine learning approach to evaluate sports facilities condition in primary school. PLoS ONE, 17(4 April). https://doi.org/10.1371/journal.pone.0267009

[6] New York State School Facilities and Student Health, Achievement, and Attendance: A Data Analysis report. (2005).

[7] Martin Eugene Sheets, by, & Hartmeister William Lan Fred Hartmeister, F. (2009). THE RELATIONSHIP BETWEEN THE CONDITION OF SCHOOL FACILITIES AND CERTAIN EDUCATIONAL OUTCOMES, PARTICULARLY IN RURAL PUBLIC HIGH SCHOOLS IN TEXAS.

Data Research

1. **Introduction:**

The project aims to analyze Sudanese schools' dataset and classify it according to the facilities to enhance the understanding of the current state of Sudanese schools. As known, Sudan one of the least development countries which still need many steps of development in many aspects but the most important is to provide well based education environment and suffers from proper dataset that can help in analyzing the situation and making good decisions. An impressive collaboration between the Ministry of Education, UNICEF, and OCHA Sudan have been done in 2021 in collecting the school's data [1] in order to ease the decision-making process and to know the actual statics of students, teachers and facilities availability as well.

3. **Data Description:**

The dataset type is .xlsx, containing 38 columns with more than 19000 rows for different schools in Sudan. The dataset contains the number of students in each class from class 1 to class 8 in addition to total number of students, total number of females, total number of males, and teachers as well. Moreover, the dataset contains facilities information. Including electricity, water, bathrooms, fence and number of classes that need restructuring.

4. **Data Analysis and Insights:**
   - Dataset first 5 rows

	School ID	school_name_arabic	school_name_english	state_name	STCODE	locality_name	LOCENG	LOCCODE	location	Type	...	school_feeding	kinde
0	53411301	ابن عباس بنين	Ibn Abbas for boys	West Kordofan	SD18	الاضية	Al Idia	SD18104	urban	Boys	...	no	
1	53411302	ابن عمر بنين	Ibn Omer for boys	West Kordofan	SD18	الاضية	Al Idia	SD18104	urban	Boys	...	yes	
2	53411303	الصباغ بنين	Asabag for boys	West Kordofan	SD18	الاضية	Al Idia	SD18104	rural	Boys	...	no	
3	53411304	عريس بنين	Arees for boys	West Kordofan	SD18	الاضية	Al Idia	SD18104	rural	Boys	...	no	
4	53411305	الحميراء بنات	Alhomayraa for girls	West Kordofan	SD18	الاضية	Al Idia	SD18104	urban	Girls	...	no	

5 rows × 38 columns

   - Dataset features type

```
: School ID              int64        grade4              int64
  school_name_arabic     object       grade5              int64
  school_name_english    object       grade6              int64
  state_name             object       grade7              int64
  STCODE                 object       grade8              int64
  locality_name          object       Total_Classrooms    float64
  LOCENG                 object       Permanent           float64
  LOCCODE                object       Needs Rehabilitation float64
  location               object       Fence               object
  Type                   object       store               object
  Status                 object       school_feeding      object
  teachers               int64        kindergarten        object
  students_f             int64        kinder_level1       float64
  students_m             int64        kinder_level2       float64
  students_total         int64        electricity         object
  grade1                 int64        Potable_Water_source object
  grade2                 int64        Latrines            object
  grade3                 int64        Latrine_male        float64
                                      Latrine_female      float64
                                      Latrine_common      float64
                                      dtype: object
```

- Missing data

School ID	0	grade4	0
school_name_arabic	16	grade5	0
school_name_english	12	grade6	0
state_name	13	grade7	0
STCODE	12	grade8	0
locality_name	13	Total_Classrooms	52
LOCENG	16	Permanent	53
LOCCODE	16	Needs Rehabilitation	52
location	2	Fence	46
Type	4	store	44
Status	4	school_feeding	44
teachers	0	kindergarten	506
students_f	0	kinder_level1	5
students_m	0	kinder_level2	7
students_total	0	electricity	44
grade1	0	Potable_Water_source	44
grade2	0	Latrines	45
grade3	0	Latrine_male	67
		Latrine_female	75
		Latrine_common	85
		dtype: int64	

- Dataset statistic review

	count	mean	std	min	25%	50%	75%	max
✗ School ID	19379.0	4.163286e+07	1.546511e+07	11101301.0	31401323.5	41713349.0	53202313.5	317033467.0
teachers	19379.0	9.959647e+00	7.189632e+00	0.0	5.0	9.0	14.0	144.0
students_f	19379.0	1.418883e+02	1.895088e+02	0.0	0.0	70.0	198.0	2806.0
students_m	19379.0	1.610883e+02	1.926406e+02	0.0	0.0	98.0	238.0	3667.0
students_total	19379.0	3.029766e+02	2.267904e+02	0.0	129.0	268.0	425.0	3667.0
grade1	19379.0	4.943083e+01	3.725515e+01	0.0	27.0	45.0	65.0	1020.0
grade2	19379.0	4.496089e+01	3.257442e+01	0.0	23.0	41.0	62.0	561.0
grade3	19379.0	4.353919e+01	3.298433e+01	0.0	20.0	40.0	61.0	519.0
grade4	19379.0	4.178528e+01	3.435616e+01	0.0	17.0	37.0	60.0	580.0
grade5	19379.0	3.726276e+01	4.219738e+01	0.0	12.0	33.0	54.0	3424.0
grade6	19379.0	3.280959e+01	3.414505e+01	0.0	8.0	28.0	49.0	2040.0
grade7	19379.0	2.900609e+01	2.886437e+01	0.0	2.0	24.0	44.0	478.0
grade8	19379.0	2.418200e+01	2.613285e+01	0.0	0.0	20.0	37.0	552.0
Total_Classrooms	19327.0	7.903451e+00	3.239467e+00	0.0	7.0	8.0	8.0	80.0
Permanent	19326.0	5.095312e+00	3.975517e+00	0.0	2.0	5.0	8.0	80.0
Needs Rehabilitation	19327.0	2.049309e+00	2.592868e+00	0.0	0.0	1.0	4.0	22.0
kinder_level1	19374.0	7.028130e+00	3.366818e+01	0.0	0.0	0.0	0.0	4025.0
kinder_level2	19372.0	6.589562e+00	2.402465e+01	0.0	0.0	0.0	0.0	2111.0
Latrine_male	19312.0	1.669894e+00	2.713348e+00	0.0	0.0	1.0	2.0	80.0
Latrine_female	19304.0	1.856455e+00	2.662501e+00	0.0	0.0	1.0	3.0	55.0
Latrine_common	19294.0	2.144190e-01	9.679450e-01	0.0	0.0	0.0	0.0	40.0

- Main key facilities

Fence, Store, School feeding, Electricity, water, latrines

- Object features explanation

```
Unique classes in location: ['urban' 'rural']
Unique classes in Type: ['Boys' 'Girls' 'Mixed']
Unique classes in Status: ['normal' 'nomadic' 'nongovernmental' 'special needs' 'quranic'
 'complementary' 'displaced']
Unique classes in electricity: ['No' 'Solar energy' 'Generator' 'Public network']
Unique classes in Potable_Water_source: ['Well' 'Other' 'Pump' 'No' 'Public network']
```
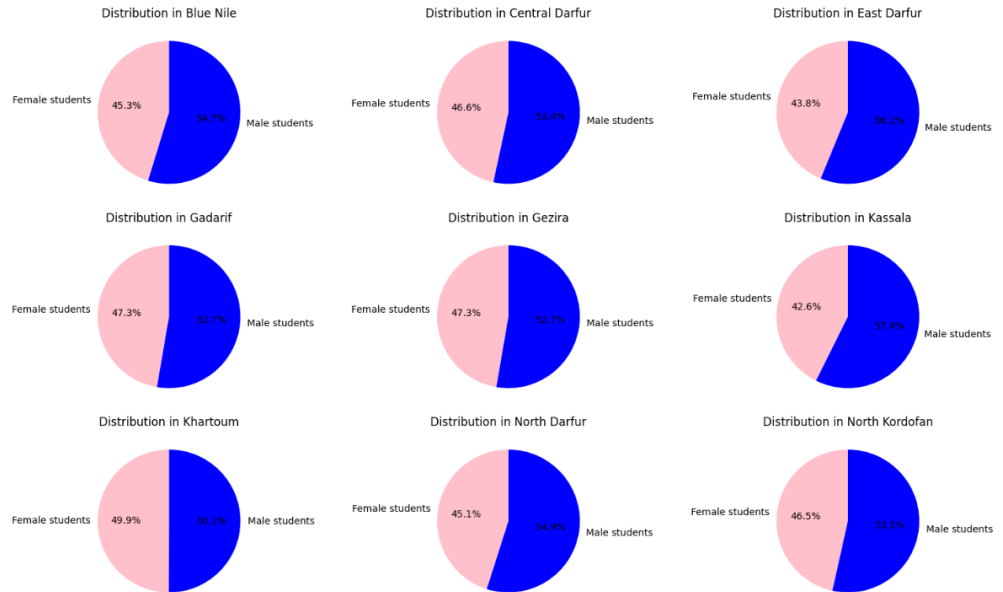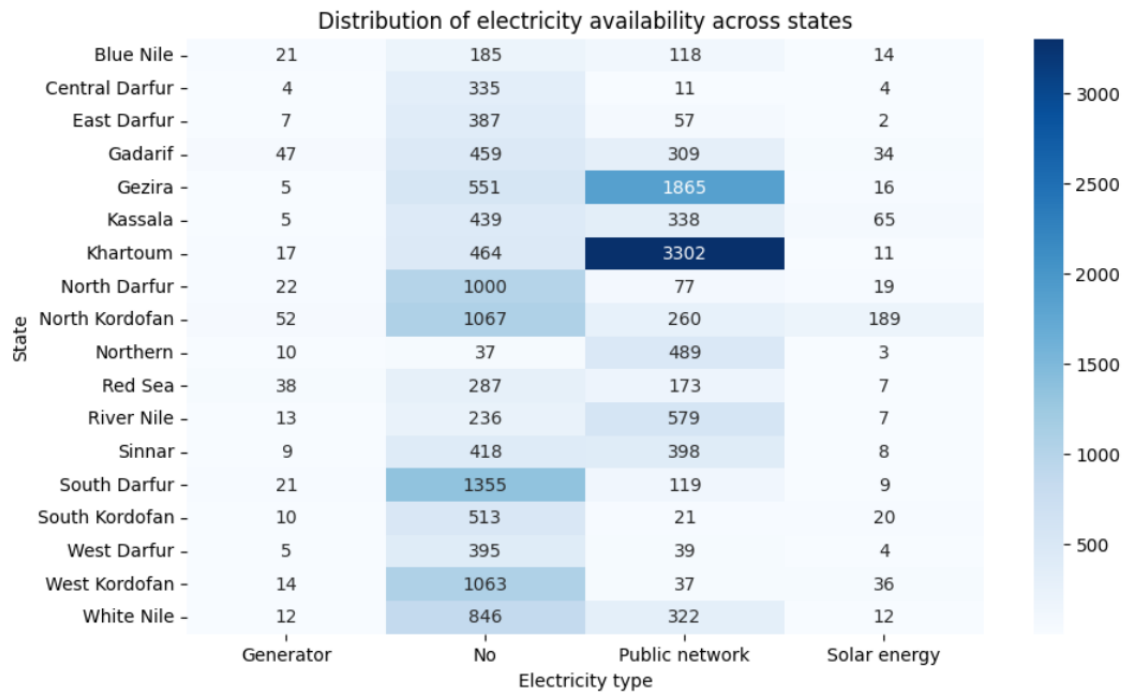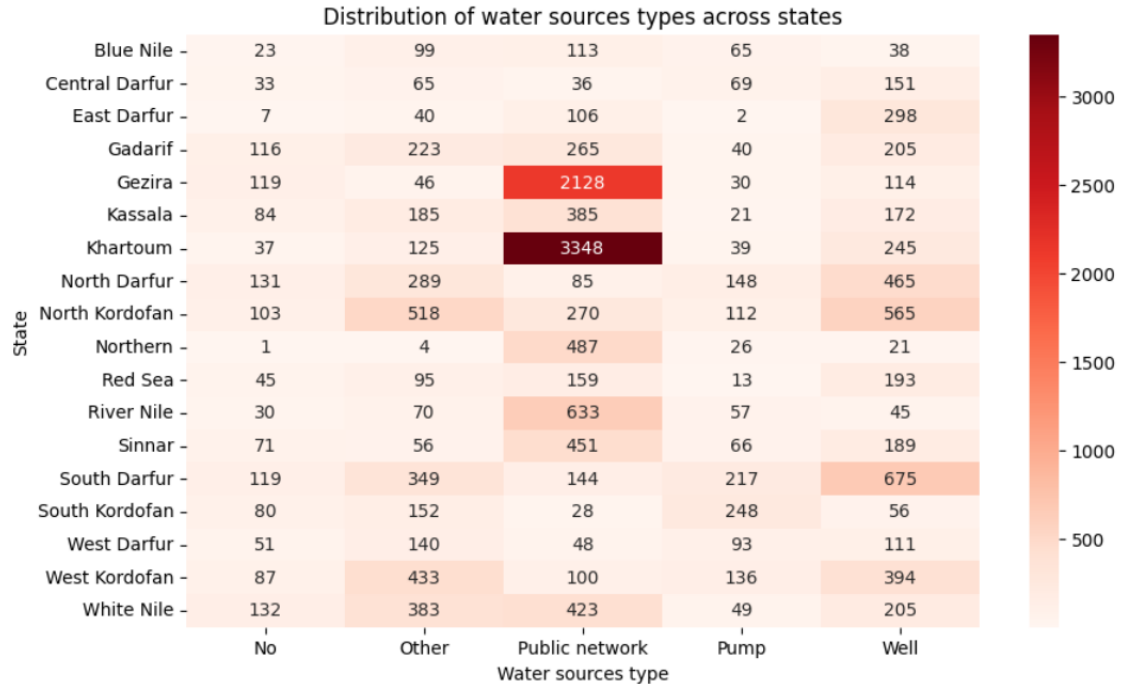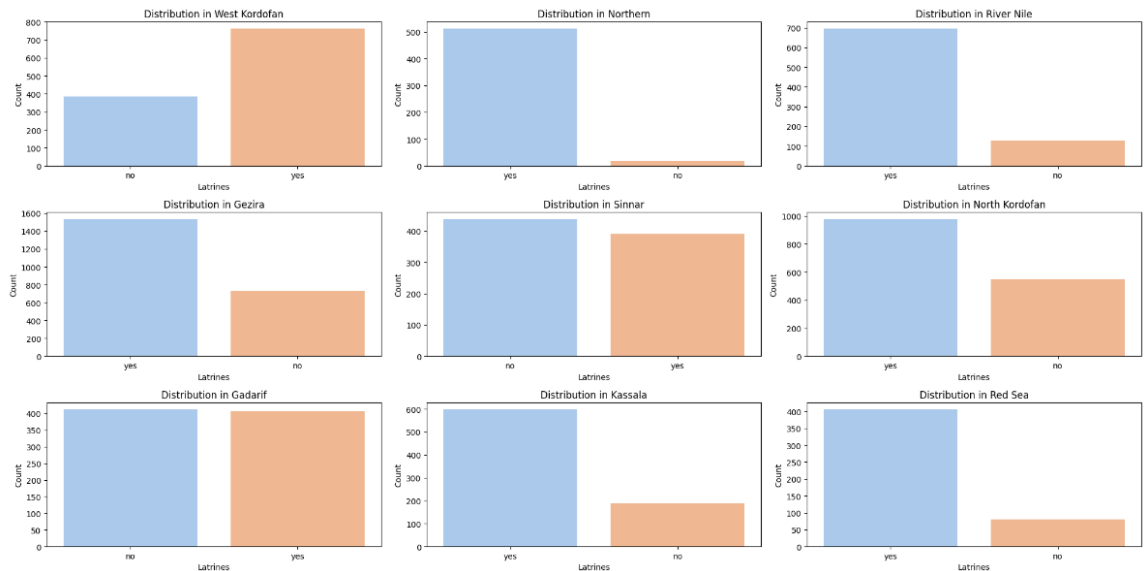
- # Gender distribution across some states

Distribution in Blue Nile	Distribution in Central Darfur	Distribution in East Darfur
Female students 45.3% / Male students 54.7%	Female students 46.6% / Male students 53.4%	Female students 43.8% / Male students 56.2%
Distribution in Gadarif	**Distribution in Gezira**	**Distribution in Kassala**
Female students 47.3% / Male students 52.7%	Female students 47.3% / Male students 52.7%	Female students 42.6% / Male students 57.4%
Distribution in Khartoum	**Distribution in North Darfur**	**Distribution in North Kordofan**
Female students 49.9% / Male students 50.1%	Female students 45.1% / Male students 54.9%	Female students 46.5% / Male students 53.5%

- # How facilities vary across regions

Distribution of electricity availability across states

State	Generator	No	Public network	Solar energy
Blue Nile	21	185	118	14
Central Darfur	4	335	11	4
East Darfur	7	387	57	2
Gadarif	47	459	309	34
Gezira	5	551	1865	16
Kassala	5	439	338	65
Khartoum	17	464	3302	11
North Darfur	22	1000	77	19
North Kordofan	52	1067	260	189
Northern	10	37	489	3
Red Sea	38	287	173	7
River Nile	13	236	579	7
Sinnar	9	418	398	8
South Darfur	21	1355	119	9
South Kordofan	10	513	21	20
West Darfur	5	395	39	4
West Kordofan	14	1063	37	36
White Nile	12	846	322	12

Electricity type

Distribution of water sources types across states

State	No	Other	Public network	Pump	Well
Blue Nile	23	99	113	65	38
Central Darfur	33	65	36	69	151
East Darfur	7	40	106	2	298
Gadarif	116	223	265	40	205
Gezira	119	46	2128	30	114
Kassala	84	185	385	21	172
Khartoum	37	125	3348	39	245
North Darfur	131	289	85	148	465
North Kordofan	103	518	270	112	565
Northern	1	4	487	26	21
Red Sea	45	95	159	13	193
River Nile	30	70	633	57	45
Sinnar	71	56	451	66	189
South Darfur	119	349	144	217	675
South Kordofan	80	152	28	248	56
West Darfur	51	140	48	93	111
West Kordofan	87	433	100	136	394
White Nile	132	383	423	49	205

- Latrines availability across some states

5. **Conclusion:**
   As shown the dataset is very informative and a lot of works can be done in result of analyzing it. The importance of this data analysis relays in empowering the LDC with effective and workable analysis that ease decision-making.

6. **Citations:**
   [1] Sudan schools dataset,
       https://data.humdata.org/dataset/sudan-schools

Technology Review:

1. **Introduction:**

   The aim of technology review is to address the methodology steps in addition to project implementation steps. Machine learning and data analysis techniques have been widely used in many applications, especially for achieving SDGs as mentioned in literature review. In this project machine learning and data analysis will be applied to our dataset to get a well classified school dataset according to facilities availabilities.

2. **Technology Overview:**

   Machine learning is a subset of artificial intelligence (AI) that enables systems to learn from experience. Data science and data analysis as a part of it is also a cross with artificial intelligence field.
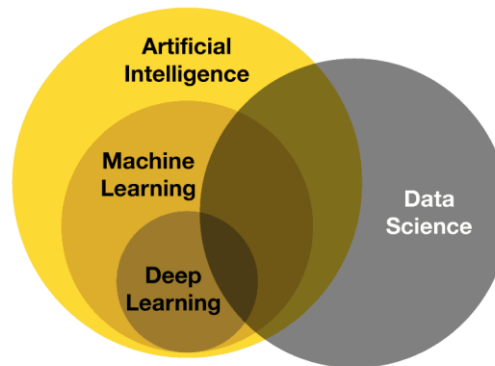
- Types of Machine Learning:
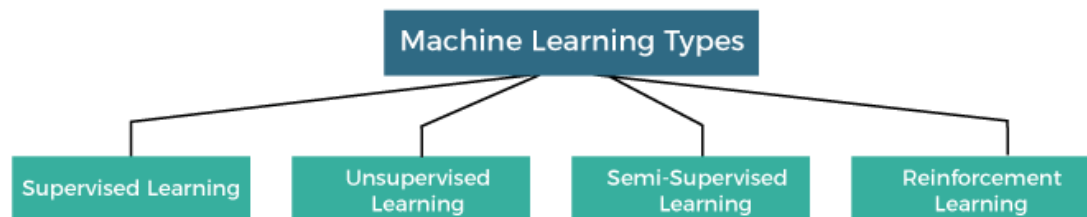
   a) Supervised Learning:
   - Trained based on labeled datasets.
   - Models learn to map input data to corresponding output labels.
   - Useful for classification and regression tasks.
   b) Unsupervised Learning:
   - Works with unlabeled datasets.
   - Emphasizes discovering hidden patterns and relationships within the data.
   - Common applications include clustering and dimensionality reduction.
   c) Semi-Supervised Learning:

- Utilizes both labeled and unlabeled data for training.
- Leverages a limited amount of labeled data alongside a more extensive pool of unlabeled data.
- Practical when acquiring labeled data is resource intensive.

  d) Reinforcement Learning
  - Depends in action => (rewards/ punishment)
  - Used in robotics.

3. **Relevance to Your Project:**
   In this project Semi-Supervised Learning will be used as the data don't have label. The first step will be labeling either by clustering (k-means) or manual labeling depending on features provided.

- Applications of Semi-Supervised Learning:

  - Valuable in various domains such as natural language processing, image recognition, and healthcare.
  - Efficiently addresses challenges where acquiring labeled data is costly or time-consuming.
- Advantages of Semi-Supervised Learning:
  - Combines benefits of both labeled and unlabeled data.
  - Offers a practical and efficient solution to real-world challenges.
  - Enhances generalization capabilities by leveraging a limited amount of labeled information.

4. **Workflow:**
   - Data preparation and preprocessing including
     - dealing with missing data,
     - convert categorical data to numerical,
     - choosing features
   - Data analysis which aims to answer research questions.
   - Label dataset according to facilities availabilities in schools
     - Manual labeling, according to availability of facilities
     - Using k-means clustering model
   - Prepare dataset for machine learning classification models.
   - Classification using different models (support vector machine, random forest, k-nearest neighbor, neural network ..)
   - Comparison between models according to performance metrices such as accuracy, f1-score, precision and recall.
   - Python programming language will be used for machine learning and for data visualization as well.
   - Cost: ??

5. Use Cases and Examples:

Educational dataset is informative dataset and allows many works on it. According to literature review some machine learning methods have been used in educational dataset. Table 1 explained the different methods used beside their advantages, disadvantages and application. Scenarios contains, knowledge tracing, undesirable student detection, performance prediction, personalized recommendation.

TABLE I: Deep Learning Algorithms in Educational Scenarios [1]

Algorithm Classification	Method	Advantage	Disadvantage	Application Scenario
Supervised Learning	CNN	Features can be extracted automatically Good ability to avoid over-fitting	Large computation Poor interpretation	Knowledge tracing Student behaviors detection
	RNN	Good at processing sequence data	Prone to gradient vanishing or gradient explosion	Knowledge tracing Student behaviors detection
	Recursive NN	Able to handle NLP	Difficult to capture the hierarchy in the data.	Knowledge tracing Student behaviors detection Personalized recommendation
	LSTM	Effectively process data with long-term dependencies	Massive computation	Knowledge tracing Student behaviors detection
	GNN	Able to process graph structure data Strong expansibility	Noise sensitivity Large computation Poor interpretation	Student behaviors detection Personalized recommendation
	Attention	Parallelizable	Large computation Poor interpretation	Knowledge tracing Student behaviors detection Personalized recommendation
Unsupervised Learning	GAN	Data without labels can be used	Difficult to evaluate	Personalized recommendation
	DBN	Able to handle complex high dimensional data	High computational cost High data quality Requirement	Knowledge tracing Student behavior detection Personalized recommendation
	VAE	Excellent data process performance	Large data requirements Complex training process	Student classification Student behavior detection Personalized learning Resources generation
Reinforcement Learning	Value Function Approach	Personalized orientation Self-adaptive adjustment	Hard to capture all the details and factors	Personalized learning path Intelligent tutoring and feedback
	Policy Search Method	Suitable for multiple educational scenarios and tasks Able to handle continuous action space	Large computation High hyperparameters requirement	Personalized recommendation Learning environment design
	Actor-Critic Algorithm	Personalized orientation Real-time feedback	Poor interpretability Difficult to train Large sample demand	Personalized recommendation Intelligent tutoring system

6. Identify Gaps and Research Opportunities:

- There are not many researchers found in order to use machine learning in facility-based classification for schools in LDC.

7. Conclusion:

To sum up, there are many techniques that can be used to analyze and classify the school's dataset depending in a facility availability. Support vector machine, random forest, k-nearest neighbor, neural network will be implemented and compared according to performance.

8. Proper Citations:

[1] Lin, Y., Chen, H., Xia, W., Lin, F., Wu, P., Wang, Z., & Liu, Y. (2023). A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining. http://arxiv.org/abs/2309.04761