Capstone Project Concept Note and Implementation Plan

Project Title: Sudanese Primary Schools Dataset Analysis and Classification based on Facility Availability

Team Members

1. Zainab Elfatih Mohamed Malik

Concept Note

1. Project Overview

- This project focuses on analyzing and classifying Sudanese primary schools based on facility availability in the schools, aligning with Sustainable Development Goal 4 (Quality Education), 5 (Gender Equality) and 9 (Industry, Innovation, and Infrastructure). The project aims to address the significant differences in educational facilities across regions in Sudan. By leveraging machine learning techniques.

2. Objectives

The specific objectives of our project include.

- Assessing the current status of primary schools in Sudan,
- classifying schools based on facility availability.
- To contribute in guiding stakeholders/government to address specific challenges faced by schools for optimal impact.

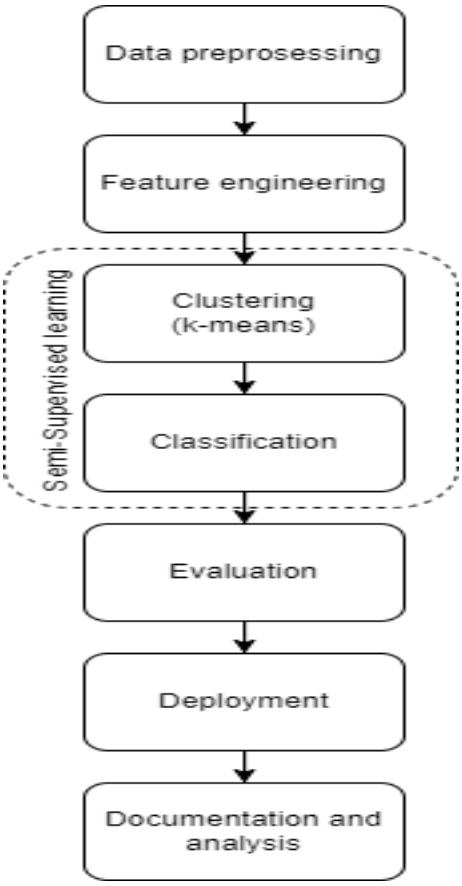
3. Background

- Sudan faces challenges in ensuring equitable access to quality education. While there have been efforts to address these issues, a comprehensive analysis using machine learning techniques can offer a detailed understanding of the disparities.

4. Methodology

- The project will employ machine learning techniques, specifically classification algorithms, to categorize primary schools based on facility availability. As the data do not have label features, clustering will be used as a first step for labeling and then classification will be done which is mainly semi-supervised learning method. For clustering, K-means will be used. For supervised learning classification models, support vector machine, random forest, k-nearest neighbor, neural network ... will be used and evaluate their performance using relevant metrics. The methodology will involve data preprocessing, feature engineering, model training, and validation.

5. Architecture Design Diagram



1. Data Preprocessing:

- Clean and preprocess the raw data to handle missing values, outliers, and inconsistencies.
- Standardize or normalize features to ensure uniformity.
- Categorical to numerical.

2. Feature Engineering:

- Identify relevant features that contribute to the classification task.
- Create new features or transform existing ones to enhance model performance.

3. Clustering (Unsupervised Learning):

- Apply K-means clustering to group schools based on similarity in facility availability.
- Assign cluster labels to the data as a preliminary step for semi-supervised learning.

4. Supervised Learning (Classification):

- Utilize various classification algorithms:
    - Support Vector Machine (SVM)
    - Random Forest
    - K-Nearest Neighbors (KNN)
    - Neural Network
    - …etc
- Train individual models on labeled data.

5. Model Evaluation:

- Assess the performance of each classification model using relevant metrics:
    - Accuracy, Precision, Recall, F1 Score, etc.
    - Compare and contrast the models to select the most suitable one.

6. Deployment:

- Integrate the selected model into the overall system.
- Deploy the system for real-time analysis and decision-making.

7. Analysis, Visualization and Documentation

6. Data Sources

- 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.

7. Literature Review

- This project aligns with previous studies emphasizing the significance of machine learning in educational analysis.
- Some papers used machine learning techniques in educational data to predict student achievement [2] while others used machine learning to evaluate facilities condition in school and also how facilities can affect the achievement. Paper [3] 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 [4] addressed data analysis report for New York State school facilities and student health, achievement and attendance. Similarly, [5] shows the relationship between the condition of school facilities and certain educational outcomes, particularly in rural public high schools in Texas.
- 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".

Implementation Plan

1. Technology Stack

- The technology stack includes Python as the primary programming language, popular machine learning libraries such as scikit-learn. Additionally, tools for data visualization and analysis, such as Matplotlib and Pandas, will be employed.

2. Timeline

Task	Description	Status/Deadline
Data Preprocessing	Clean and handle missing values. Identify and address outliers and inconsistencies.	Done
Feature Engineering	Identify relevant features for classification. Create new features or transform existing ones.	Done
Clustering	Apply K-means clustering to group schools based on facility availability. Assign cluster labels for semi-supervised learning.	Done
Supervised Learning (Classification)	Utilize SVM, Random Forest, KNN, Neural Network. Train models on labeled and unlabeled data.	In progress 29. Nov
Model Evaluation	Assess performance using metrics (Accuracy, Precision, Recall, F1 Score). Compare and select the most suitable model.	In progress 29. Nov
Deployment	Integrate the selected model into the overall system. Deploy for real-time analysis and decision-making.	11. Dec
Analysis, Visualization, and Documentation	Conduct in-depth analysis. Create visualizations to communicate results. Document methodologies, findings, and insights.	In progress 11.Dec

3. Milestones

- Utilize various classification algorithms.
- Train individual models on labeled data.
- Assess the performance of each classification model using relevant metrics.
- Compare and contrast the models to select the most suitable one.
- Integrate the selected model into the overall system.
- Deploy the system for real-time analysis and decision-making.
- Analysis
- Visualization
- Documentation

4. Challenges and Mitigations

- Challenges include issues related to data quality, model performance, and technical constraints. Mitigation strategies involve thorough data validation, model fine-tuning, and flexibility in the application of machine learning algorithms. Especially that the dataset is not labelled, and clustering method used in labeling it.

5. Ethical Considerations

- Prioritize data privacy,
- Address biases in the dataset,
- Carefully consider the potential impact on the target community.
- Transparency in the decision-making process and regular ethical reviews will guide the project approach.

6. References

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

[2] 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

[3] 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

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

[5] 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.