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, a country with a rich cultural heritage, is grappling with challenges in providing equitable access to quality education. Despite concerted efforts to address these issues, disparities persist, hindering the nation's progress towards achieving Sustainable Development Goal 4 (Quality Education). The Sudanese education system faces challenges such as unequal distribution of resources, varying infrastructure standards, and disparities in gender access, all of which contribute to an uneven educational landscape [6].
- To comprehend and tackle these challenges comprehensively, a thorough analysis is a must. While previous initiatives have made strides in improving education, a more nuanced and data-driven approach is needed to gain deeper insights into the complexities of the educational landscape [6]. This project, driven by machine learning techniques, seeks to provide a granular understanding of the current state of primary schools in Sudan.
- The World Bank and UNESCO emphasize the significance of quality education in fostering economic development and societal progress. The comprehensive analysis

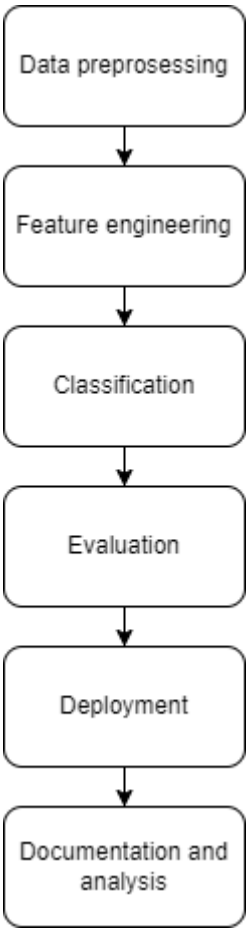
undertaken in this project aligns with the global commitment to Sustainable Development Goal 4, aiming to ensure inclusive and equitable quality education for all [7]. By specifically delving into the facility availability in Sudanese primary schools, the project addresses the multifaceted challenges hindering progress toward this goal.

- Furthermore, Sustainable Development Goal 5 (Gender Equality) and Goal 9 (Industry, Innovation, and Infrastructure) are inherently intertwined with the educational landscape. Gender disparities in access to education persist, and addressing these issues requires a multifaceted approach that considers not only the availability of facilities but also the inclusivity and accessibility of education for all genders. Additionally, fostering innovation and improving infrastructure in schools are pivotal components in advancing the educational sector and preparing students for the demands of the modern world [7].

4. Methodology

- The project will employ machine learning techniques, specifically classification algorithms, to categorize primary schools based on facility availability. As the data does not have label features, new features will be generated from given ones then apply manual labelling according to that. 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. 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.

4. 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.

5. Deployment:

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

6. 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
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

- Identify relevant features for classification using feature importance ranking.
- Create new features to help in classification.
- Utilize SVM, Random Forest, KNN, Neural Network.
- Assess model performance using metrics (Accuracy, Precision, Recall, F1 Score).
- Integrate the selected model into the system.
- Deploy for real-time analysis and decision-making.
- Conduct in-depth analysis.
- Create visualizations.
- Document methodologies, findings, and insights.

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.

[6] United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Retrieved from https://sdgs.un.org/2030agenda

[7] World Bank. (2020). World Development Report 2018: Learning to Realize Education's Promise. Retrieved from https://www.worldbank.org/en/publication/wdr2018