Balochistan University of Information Technology,
Engineering Management Sciences

Big Data Analytics

Projecct Report

SDG - 4

Quality Education

Author:

Manahil Sabir
52904

1 Problem Statement

Referencing the United Nations report on big data for sustainable development, I choose Quality Education as a challenge that can benefit from the analysis of large-scale datasets. The project will include analyzing patterns, identifying factors, and assessing the overall data.

2 Objective

This research project aims to gain proficiency in Apache Spark, get familiar with data analytics tools, and apply them to address quality education challenges related to sustainable development, as identified in the United Nations report on big data for sustainable development.

3 Introduction

The UIS collects education statistics in aggregate form from official administrative sources at the national level. Collected information encompasses data on the structure of national educational programs, access, participation, progression, teacherâs statutory compensation, school infrastructure, completion, literacy, educational attainment, and human and financial resources.

These statistics cover formal education in public (or state) and private institutions (early childhood education, primary and secondary schools, colleges, universities, and other tertiary education institutions), and special needs education (both in regular and special schools). UIS data can be accessed in the following ways:

- SDG 4 Data Explorer on SDG 4 indicators provides easy-to-navigate dashboards organized by country or indicator and the possibility to download a long-format database. There are tabs for regional averages and country profiles that include ISCED mappings.
- Global Education Observatory is an easy-to-use gateway to education data for the benchmark indicators1, including interactive visualizations that can be disaggregated by sex, region, and country.

4 Related Work

The integration of big data technologies in education has shown significant potential to transform various aspects of educational practices and outcomes. Research highlights the benefits of big data in enhancing academic productivity, teaching efficiency, and personalized learning experiences [8]. The use of big data technology in higher education management is seen to positively impact teaching quality and student development [10]. Moreover, the implementation of machine learning algorithms via Apache Spark and its MLlib library has been explored to handle large datasets efficiently, providing fast and accurate processing for various applications including intrusion detection systems and educational data analytics [2, 7]. UNESCO's efforts in collecting and analyzing educational data underscore the importance of data-driven decision-making in achieving Sustainable Development Goals (SDGs) related to education [12]. Studies also discuss the optimization of resource management in Apache Spark to enhance performance, ensuring efficient handling of big data in educational contexts [4]. Another paper discusses the challenges associated with Big Data, such as volume, velocity, variety, variability, and complexity[14]. It discusses Data Complexity, quality, silos, security, and skill shortage. Other research highlights the role of big data in transforming educational management and processes [1, 11, 5, 3, 13, 9, 15, 6]. Collectively, these studies highlight the transformative impact of

big data technologies and advanced analytics in the educational sector, driving improvements in efficiency, personalization, and strategic management.

5 Dataset

The project uses the Quality Education dataset by UNESCO. This dataset archive presents historical NON-CORE indicators for education, curated from a large dataset provided by the UNESCO Institute of Statistics (UIS). Released in February 2020 and extracted on August 21, 2020, this archive is designed to maintain high data quality while being accessible to a broad audience. The dataset is offered in a normalized CSV format with UTF-8 encoding, ensuring ease of use for both novice and experienced data professionals. For more information and tutorials, users can refer to the UIS Developer Portal.

The data consisted of 6 data files which were merged to create final dataset on the basis of Indicator ID.

1. Country
   This file lists all country codes and their descriptive labels:
   Country_ID, Country_name

2. Data National
   This file contains all the national data available for this dataset and includes the following fields:
   Indicator_ID, Country_ID, Year, Magnitude, Qualifier

3. Data Regional
   This file contains all the regional data available for this dataset and includes the following fields:
   Indicator_ID, Region_ID, Value, Magnitude, Qualifier

4. Label This file is a list of all indicator codes and their descriptive labels:
   Indicator_ID, Indicator_Label

5. Metadata
   This file contains all the metadata associated to the NATIONAL and REGIONAL data files above and includes the following fields:
   Indicator_ID, Country_ID, Year, Type, Metadata

6. Region
   This file lists all regions and the countries that belong to each region:
   Region_ID, Country_ID, Country_Name

5.1 Metadata

5.1.1 Indicator Metadata

Most indicators have a Glossary entry that can be accessed on the UIS website containing the indicator's definition, interpretation, purpose, quality standards, calculation, types of disaggregation, and limitation.

5.1.2 Magnitude

MAGNITUDE describes the NATURE of the data point. Possible values are:

- **NIL:** The value will be 0, and should be treated as NIL.
- **NA:** The value will be 0. This data point is NOT APPLICABLE for the submitting nation.
- **SUPP :** The value will be BLANK. The data point was SUPPRESSED at the request of the submitting nation.
- **LOWREL :** The value will be NUMERIC. The data point is of LOW RELIABILITY.
- **INCLUDED :**The value will be BLANK. This data is INCLUDED in ANOTHER data point.
- **INCLUDES** : The value will be NUMERIC. This data point INCLUDES data from another data point.

5.1.3 Qualifier

QUALIFIER describes the QUALITIES of the data point. Possible values are:

- **NAT_EST :** The value will be NUMERIC. This data point is a national estimate.
- **UIS_EST** : The value will be NUMERIC. This data point is an estimate produced by the UNESCO Institute for Statistics.

6 Methodology

The project focuses on studying Quality Education statistics using Spark and involves collecting, preprocessing, and analyzing the dataset using Spark's distributed data processing capabilities. A machine learning model is employed to extract insights and predict future trends in environmental indicators.

6.1 Data Collection

The dataset used in this study is the EDUNONCORE_DATA_NATIONAL dataset, which includes various environmental indicators across different countries and years.

6.2 Data Preprocessing

The data is preprocessed using Spark to handle missing values and ensure appropriate data types. Rows with missing values in critical columns are dropped, and missing values in other columns are filled with placeholders.

6.3 Exploratory Data Analysis (EDA)

EDA is performed to understand the distribution and trends in the data. Visualization techniques are used to identify key patterns.

6.4 Modeling

A Linear Regression model is built using Spark MLlib to predict future values of environmental indicators. The model is trained on historical data and evaluated using metrics such as Root Mean Squared Error (RMSE).

6.5 Tools and Technologies

The study utilizes Apache Spark for data processing and machine learning. Visualization is done using Matplotlib and Seaborn in Python.

6.6 Data Merge

The National and Regional data files can be linked with the label and metadata file using the indicator ID variable as the matching key. While merging the metadata file, remember that multiple metadata entries can match a unique data point from the data file or multiple rows for a specific INDICATOR_ID/COUNTRY_ID/YEAR combination in the metadata file.

7 Results and Findings

7.1 Trends in dataset

The summary of the results is given below and the trends in the dataset are visually shown and it visualizes the average values of an environmental indicator over the years from 1970 to 2020.

The blue line represents the actual average values of the environmental indicator over the years. The trend line shows significant fluctuations in the data, especially around the year 2000, the orange line appears briefly around the year 2000, indicating a subset of data points or predictions for that period, whereas the pink line at the horizontal axis appearing at zero, serves as a baseline for comparison, indicating where the average value is zero.

The plot in **fig1** shows dramatic changes in the indicator values over time. The spikes and drops could be due to significant environmental events, policy changes, or other influential factors during the time period.

7.2 Error (Actual vs Predicted Values)

The model's performance was evaluated using the Root Mean Squared Error (RMSE) metric. The scatter plot below shows the actual values and the predicted values. The blue points represent the actual values, while the red line represents the predicted values. The visualization provides a clear indication of how well the model's predictions align with the actual data.

From the visualization in **fig2**, it is evident that the model captures the general trend of the data, although there are some discrepancies between the predicted and actual values. These discrepancies could be due to various factors, including the simplicity of the linear regression model and the potential presence of outliers or noise in the data.

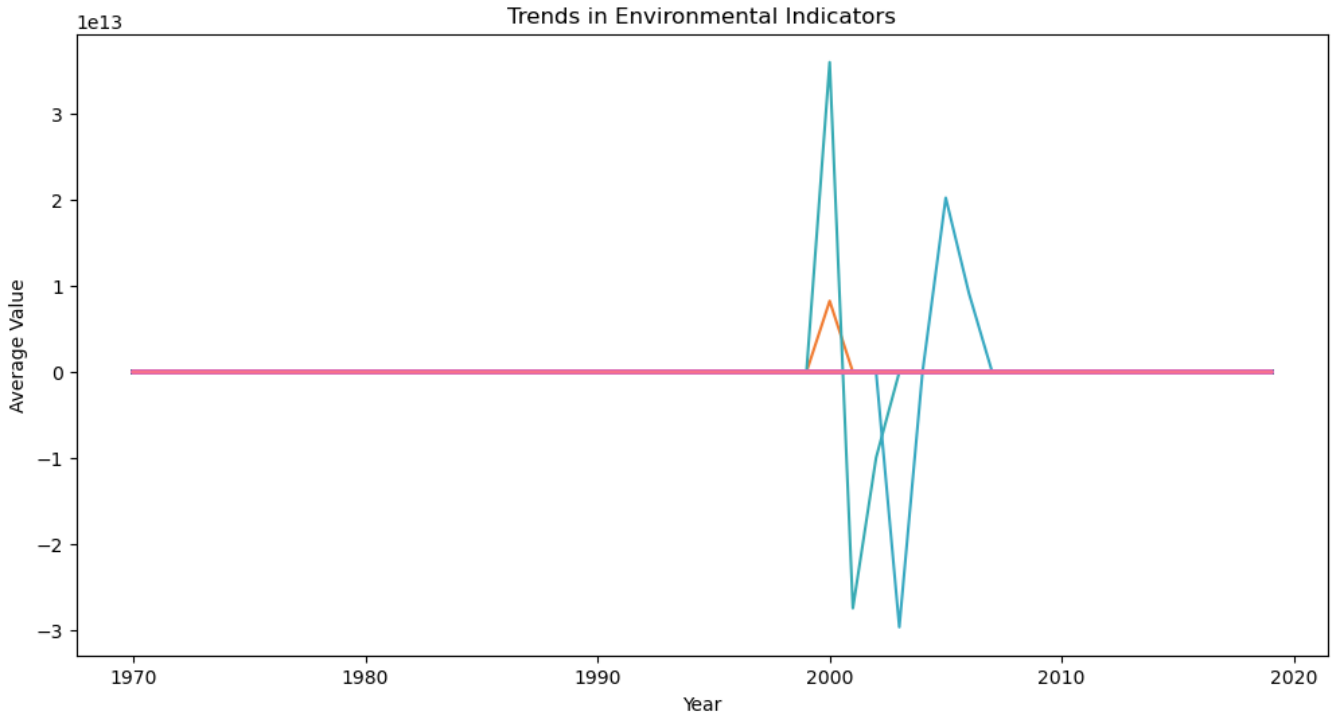

Figure 1: Trends in dataset

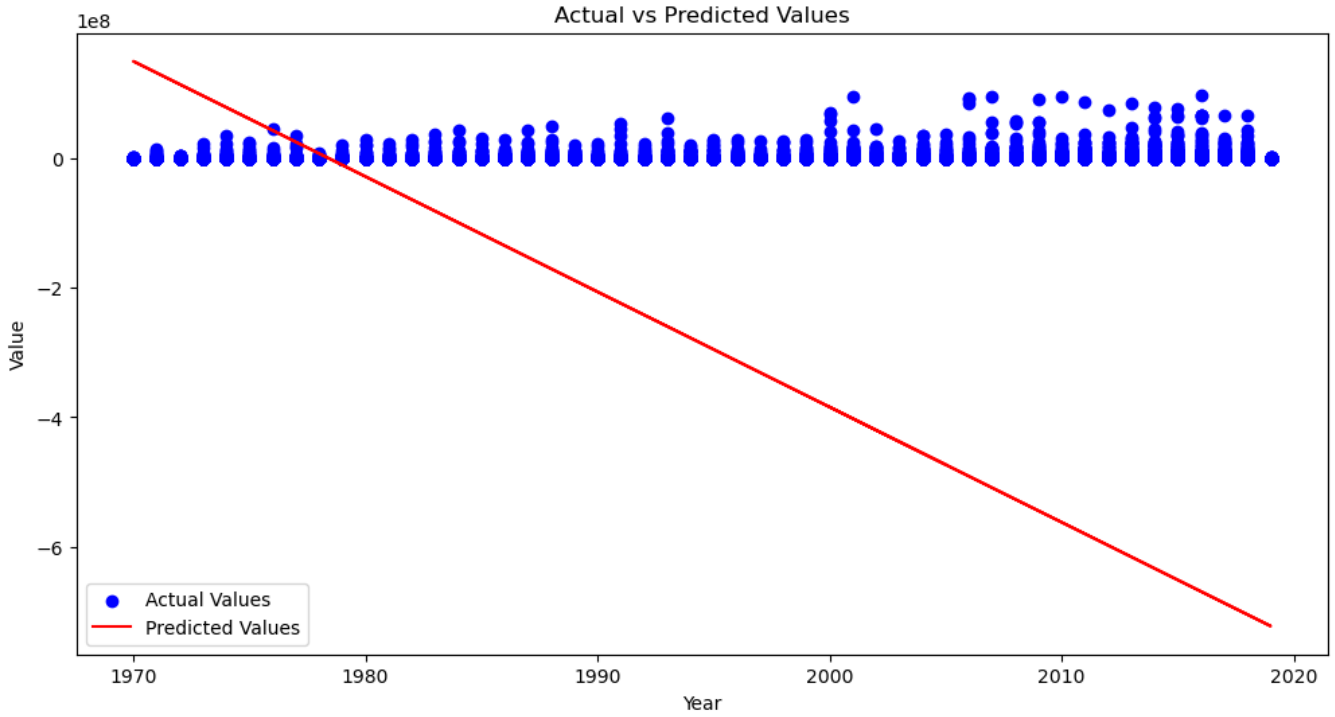

Figure 2: Actual vs Predicted Values

8 Conclusion and Future Work

The project provided an initial window to step towards learning a whole technology. It helped learning a lot with hands-on practice with enough space to explore and gain useful insights. Future work could focus on improving the model's accuracy and exploring different machine learning algorithms. We can also work to further enhance the accuracy and reliability of the predictions.

References

[1] Khatera Akrami, Mursal Akrami, Fazila Akrami, and Musawer Hakimi. Investigating the integration of big data technologies in higher education settings. *Indonesian Journal of Multidisciplinary on Social and Technology*, 2(2):296, 2024.

[2] Ramin Atefinia and Mahmood Ahmadi. Performance evaluation of apache spark mllib algorithms on an intrusion detection dataset. *arXiv.org*, 2022.

[3] Manuel Ayala-ChauvÃn, Boris Chucuri-Real, Pedro Escudero-Villa, and Jorge Buele. Big data as a tool for analyzing academic performance in education. In *Book Chapter*. 2024.

[4] Khadija Aziz, Dounia Zaidouni, and Mostafa Bellafkih. Leveraging resource management for efficient performance of apache spark. *Journal of Big Data*, 40537(19), 2019.

[5] Jiang Bian and Tao Yang. Application of big data technology in college music education. *International Journal of Web-based Learning and Teaching Technologies*, 2024.

[6] Malinka Ivanova, Valentina Terzieva, and Tat'yana Aleksandrovna Ivanova. The role of big data in intelligent educational platform: A functional architecture. In *Proceedings Article*, 2023.

[7] A.N.M. JayaLakshmi and K. V. Krishna Kishore. Performance evaluation of dnn with other machine learning techniques in a cluster using apache spark and mllib. *Journal of King Saud University - Computer and Information Sciences*, 1319(1578), 2018.

[8] Amina Khalid and Obeng Owusu-Boateng. The adoption of big data in the education sector. In *Book Chapter*. 2024.

[9] Keyu Liu and KimMee Chong. Impact of big data on the development of university education by computer software analysis. In *Proceedings Article*, 2023.

[10] Xueyuan Liu. The use of big data technology in higher education management. *World Journal of Education and Humanities*, 6(2):109, 2024.

[11] Naida O Omarova and A. A. Echilova. Big data technologies in the education system. In *Springer proceedings in business and economics*. 2024.

[12] Nimmi Maria Oommen and Suramya Mathai. Heuristic insights on gender education. *Ymer*, 21(2):16, 2022.

[13] Shakhzod Saydullaev. Exploring big data applications for knowledge management in higher education administration. *Yashil iqtisodiyot va taraqqiyot*, 1(11-12):374, 2023.

[14] Abdul Ghaffar Shoro and Tariq Rahim Soomro. Big data analysis: Ap spark perspective. *Global Journal of Computer Science and Technology: C Software & Data Engineering*, 15(1):7–14, 2015.

[15] Natalia Vitanova. Big data in the education. *Pedagogika*, 2023.

Dataset *Quality Education dataset by UNESCO
https://uis.unesco.org/bdds