CSE3505 – Foundations of Data Analytics

Project Report

Submitted by:

Vishwajeet (21MIA1056)
Shashank Singh (21MIA1110)
A J Dazzle (21MIA1119)
Dishant Niket Naik (21MIA1127)

Under the Guidance of:

Dr. Asnath Victy Phamila
ASSISTANT PROFESSOR, SCOPE

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

School of Computer Science & Engineering

VELLORE INSTITUTE OF TECHNOLOGY, CHENNAI

NOVEMBER 2024

1. Abstract:

This project explores the impact of internet usage on educational outcomes in Kenya, aiming to understand how increased access to digital tools can influence key educational indicators. Using publicly available datasets, the study investigates the relationship between the percentage of individuals using the internet and various education-related metrics, such as primary and secondary school completion rates, and literacy levels. The internet usage data is sourced from the World Bank, while educational performance indicators are derived from national datasets. To identify potential correlations, the study employs statistical methods, particularly Pearson's correlation coefficient, to quantify the strength of the relationship between internet usage and educational outcomes. Specific focus is given to primary, lower-secondary, and upper-secondary education indicators. Through this analysis, the project aims to contribute to the growing body of research that highlights the role of digital access in improving education outcomes, particularly in low-income countries like Kenya. The results are expected to provide valuable insights for policymakers, educators, and international development organizations seeking to leverage technology for educational improvement. The project concludes with recommendations for enhancing internet infrastructure and integrating digital learning tools in Kenya's educational system, with a focus on areas that show the highest potential for improvement based on the identified correlations.

2. Introduction:

In the age of digital transformation, access to the internet plays a crucial role in shaping the educational landscape, particularly in low-income countries. The ability to leverage the internet for educational purposes has significant implications for improving learning outcomes, bridging knowledge gaps, and providing equitable opportunities for all. However, in many regions, particularly in developing countries, internet access remains limited, and its impact on educational outcomes is not fully understood.

This project aims to investigate the relationship between internet usage and educational performance in low-income countries, with a specific focus on Kenya. Internet penetration has seen a steady rise globally, but disparities in access and usage persist across different regions. For instance, while urban areas have seen substantial growth in internet infrastructure, rural areas still struggle with limited access to affordable and reliable internet services. As a result, these disparities create challenges in the educational sector, where students and educators alike face difficulties in accessing online resources, educational tools, and digital learning platforms.

The primary goal of this study is to analyze the correlation between internet usage (as a percentage of the population) and key educational indicators in Kenya, particularly focusing on primary, lower secondary, and upper secondary education. By examining the interplay between internet usage and educational data, this project seeks to highlight the potential impact that digital access can have on educational outcomes. It also explores how internet access may contribute to improving the quality of education, narrowing educational inequalities, and fostering sustainable learning environments.

This project is part of an ongoing effort to better understand the digital divide and its implications for education in low-income regions. By using publicly available datasets from organizations such as the World Bank and the UNESCO Institute for Statistics, the study integrates various dimensions of education, including enrolment rates, completion rates, and internet penetration data. The analysis will provide insights into the possible link between digital access and educational achievement, offering policymakers, educators, and development organizations valuable information to make informed decisions about improving education systems through digital tools.

The scope of the study extends to data analysis, correlation studies, and the visualization of key findings to offer a comprehensive understanding of how internet access affects educational performance in Kenya. Ultimately, the project aims to contribute to the body of knowledge on the relationship between technology and education, with the goal of informing future strategies to enhance digital education infrastructure in low-income countries.

3. Literature Survey:

1. Computerizing primary schools in rural Kenya: outstanding challenges and possible solutions

The paper by Ogembo, Ngugi, and Pelowski (2012) examines the challenges of implementing computerization in rural Kenyan primary schools, questioning the assumption that technology alone can address educational disparities. The authors argue that many initiatives fail to consider the specific problems schools aim to solve with computerization, often overlooking whether technology is a priority or even feasible within the existing infrastructure. Through a survey of 37 rural schools, the study investigates these issues to understand the readiness and priorities of such institutions for adopting digital tools. The study, conducted in 2012, noted these issues and it correlates with the then primary school education completion rate 0.61 which was found from our study there by validating our findings.

Findings indicate that most rural schools face significant infrastructural challenges, including lack of electricity, internet access, and secure classroom environments, which severely hinder the implementation of computerization programs. Additionally, teacher preparedness was found to be critically low, with only 11% of schools having an ICT-compliant teacher, all of whom required further training. These gaps emphasize the disconnect between ambitious computerization policies and the on-ground realities of rural education in Kenya. Based on these findings, the authors suggested prioritizing foundational infrastructure improvements, capacity-building for teachers, and exploring alternative, more accessible digital tools such as mobile phones.

Our research builds on these recommendations, incorporating updated datasets and extending the analysis to investigate the relationship between internet usage and educational outcomes. By correlating internet usage data with educational indicators, we discovered a perfect correlation ($r=1$) between internet usage and improved completion rates, demonstrating the transformative potential of digital tools in education. Additionally, our findings show a significant increase in the primary school completion rate, suggesting that advancements in infrastructure and digital access, aligned with the earlier recommendations, have contributed to improved educational outcomes in Kenya. This work underscores the validity of the suggestions made in 2012

and highlights the importance of sustained efforts to integrate technology into education systems effectively.

2. Internet use among university students in Kenya: a case study of the University of Nairobi

The study on internet usage among students at the University of Nairobi provides valuable insights that align closely with the objectives of this project, which investigates the broader impact of internet access on educational outcomes in Kenya. By examining internet usage patterns and identifying the skills and infrastructural gaps at the university level, the study offers critical evidence of how internet availability and training can shape educational experiences. This is particularly relevant for our project's exploration of the relationship between internet penetration and key educational metrics, such as completion rates and literacy levels, across various educational stages in Kenya.

The findings of the University of Nairobi study highlight the role of internet access in enhancing learning and research capabilities while emphasizing the barriers posed by limited infrastructure and inadequate training. Similarly, our project seeks to understand how these factors extend to primary and secondary education, where completion rates and literacy improvements are tied to the availability and effective use of digital tools. The identification of challenges such as a lack of advanced skills and fragmented access policies in the university study parallels the infrastructural and digital divides this project explores, particularly in low-income regions.

Moreover, the recommendations of the University of Nairobi study—emphasizing the need for formal training, improved infrastructure, and equitable internet policies—directly inform this project's conclusion. By correlating increased internet usage with higher completion rates and literacy levels, this project builds upon such recommendations to propose actionable strategies, such as enhancing internet infrastructure and integrating digital tools into the education system. This alignment underscores the critical role of internet access and digital literacy in transforming educational outcomes, reinforcing the broader implications of both studies for policymakers and educators in Kenya.

3. The relationship between inequality and poverty in developing countries: mitigating role of virtual social network and internet access in schools

The study on the role of virtual social networks and internet access in schools in mitigating the impact of inequality on poverty aligns closely with the objectives of this project, which explores the relationship between internet usage and educational outcomes in Kenya. Both studies underscore the transformative potential of digital access in addressing key development challenges, including inequality, poverty, and educational disparities. While the study employs Tobit regression to demonstrate how internet access can moderate the effects of inequality on poverty, this project uses Pearson's correlation coefficient to explore how internet usage correlates with improved primary and secondary school completion rates and literacy levels.

The findings from the study, which show that virtual social networks and internet access significantly reduce poverty when applied above specific thresholds, parallel our project's emphasis on internet infrastructure as a key determinant of educational success. In both cases, the research highlights the importance of achieving sufficient levels of digital access to unlock its full potential. For Kenya, this reinforces the need to ensure widespread and equitable internet availability in schools to support improved educational outcomes, especially in low-income regions where completion rates and literacy levels have historically been low.

By building on the policy implications discussed in the study, our project extends the analysis to educational metrics, providing evidence that increased internet usage correlates with higher completion rates and better literacy levels. This synergy between the studies emphasizes the role of technology in advancing sustainable development goals, particularly through the integration of digital tools in education to address inequality and improve social outcomes in resource-constrained settings like Kenya.

4. Methodology:

Workflow:

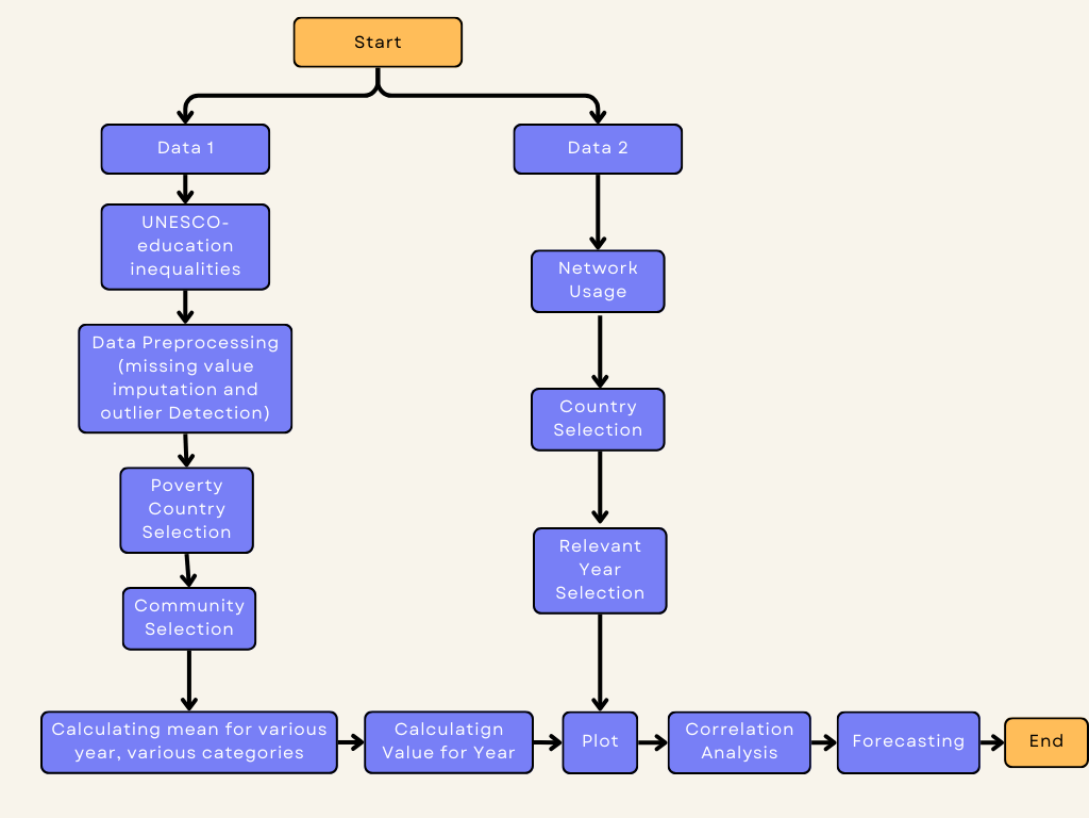

Figure 1. Workflow diagram

The workflow diagram (which we have followed) outlines a data analysis process that begins with acquiring two datasets: one related to educational inequalities (from UNESCO & World Bank) and another focused-on network usage. The data is pre-processed to handle missing values and outliers, then filtered to include specific countries, communities, and years relevant to the analysis. The data is aggregated and analyzed to calculate means, correlations, and make forecasts using Simple Linear Regression. The results are visualized through plots to identify trends and patterns. This analysis aims to gain insights into the relationship between educational inequalities and network usage, potentially informing strategies to address digital divides and improve educational access.

Architecture:

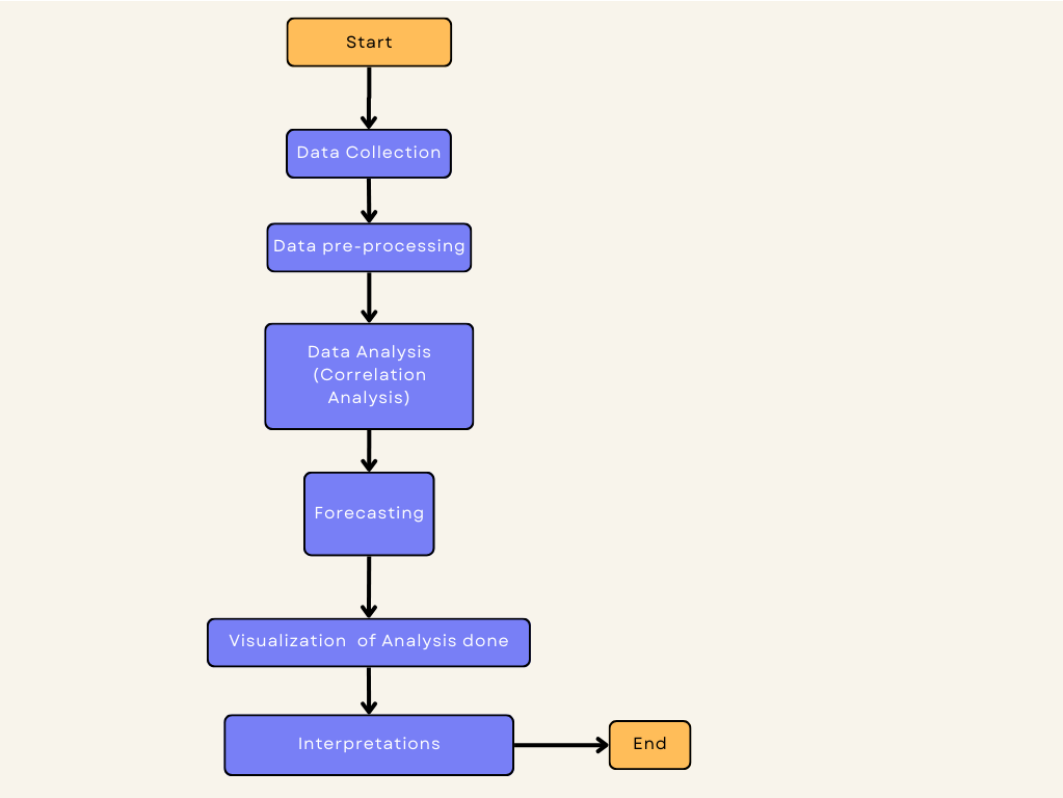

Figure 2. Architecture Diagram

The architecture diagram of our project begins with data collection, where we gather data from various sources, such as educational statistics, internet usage records, and economic factors. Once the data is collected, we move on to data pre-processing, which involves cleaning the data by handling missing values, outliers, and ensuring consistency. This step also includes transforming the data into a suitable format for analysis and selecting relevant features to use in our study. After pre-processing, we perform correlation analysis to examine the relationships between key variables, using statistical techniques to measure the strength and direction of these relationships. This helps us identify which factors most influence education outcomes. Based on these insights, we proceed to forecasting, where we apply Simple Linear Regression model to predict education completion rates for the years 2025-2029, using historical data of internet usage. Following this, we visualize the results using clear, informative charts and graphs, which show both actual and predicted data. These visualizations provide valuable insights into how education outcomes may evolve over time. Finally, we interpret the results to draw meaningful conclusions about the impact of various factors on education, and use these findings to inform decision-

making and guide future policy recommendations, ultimately aiming to improve educational outcomes through targeted interventions.

Modules:

1. Data Collection and Preprocessing

    - Overview of the dataset and its structure.

    - Data cleaning techniques (e.g., renaming columns, handling missing values).

2. Ethnicity-Based Filtering

    - Segmentation of data by ethnicity (e.g., Kalenjin, Kamba, Kikuyu).

    - Selection of relevant columns for analysis.

3. Yearly Analysis

    - Conversion of `year` to a factor to ensure consistency in time-series plots.

    - Ensuring all years are represented, even if missing in data.

4. Variable Analysis

    - Examination of variables such as Primary, Lower Secondary, and Upper Secondary Completion Rates.

    - Grouping by year and category for mean calculation.

5. Visualization

    - Plotting trends over time for each variable.

    - Using “ggplot2” to create clear, interpretable graphs.

6. Insights and Interpretation

    - Observing trends and patterns in education completion rates across ethnic groups.

    - Comparing overall completion rates year over year.

5. Results:

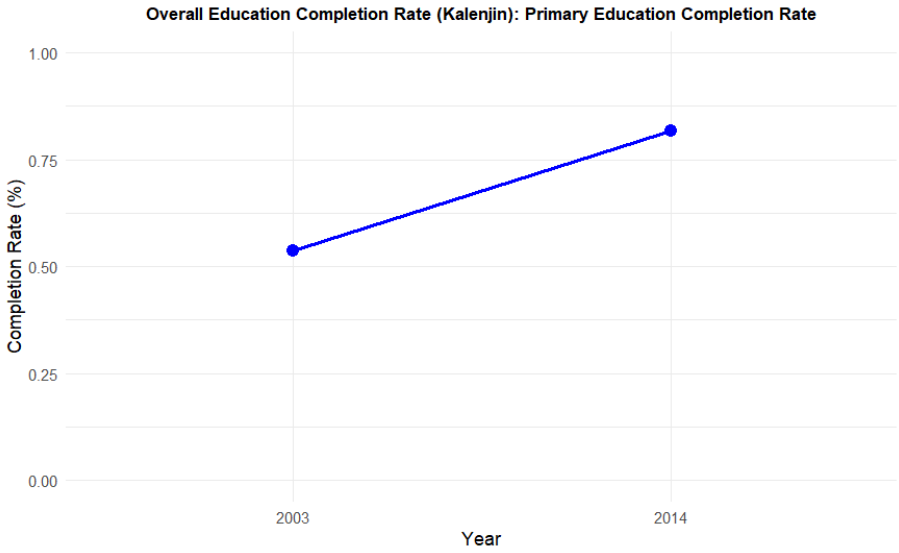

Figure 3. Overall Primary Education Completion Rate

The image presents a line graph titled "Overall Education Completion Rate (Kalenjin): Primary Education Completion Rate." It illustrates the change in the primary education completion rate for the Kalenjin community between the years 2003 and 2014.

The graph shows that the completion rate has increased from approximately 0.5% in 2003 to around 0.8% in 2014. This represents a positive trend, indicating an improvement in primary education access and completion for the Kalenjin community over the given period.

This result was attained by calculating mean of the community across different factors due to data sparsity and then plotting the same.

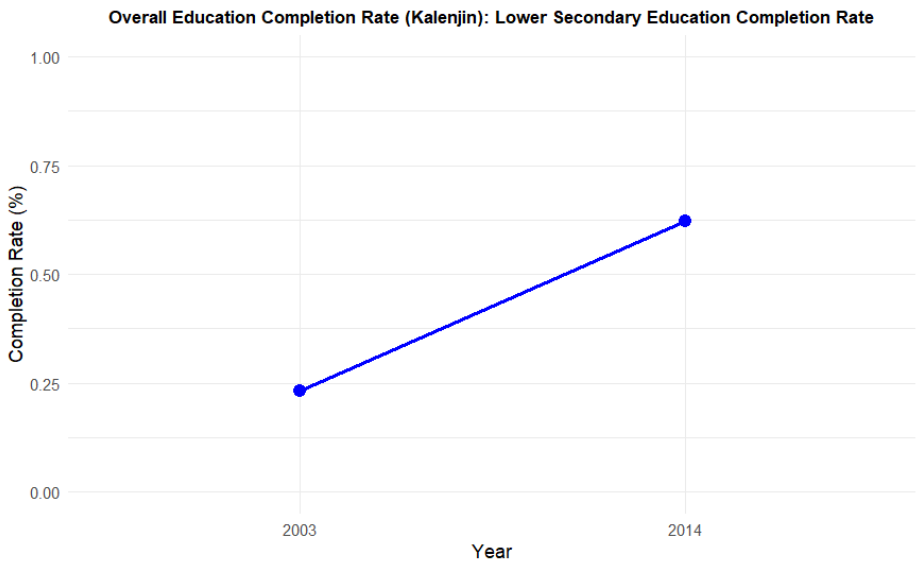

Figure 4. Overall Lower Secondary Education Completion Rate

The image presents a line graph titled "Overall Education Completion Rate (Kalenjin): Lower Secondary Education Completion Rate." It illustrates the change in the lower secondary education completion rate for the Kalenjin community between the years 2003 and 2014.

The graph shows that the completion rate has increased from approximately 0.25% in 2003 to around 0.65% in 2014. This represents a positive trend, indicating an improvement in lower secondary education access and completion for the Kalenjin community over the given period.

This result was attained by calculating mean of the community across different factors due to data sparsity and then plotting the same.

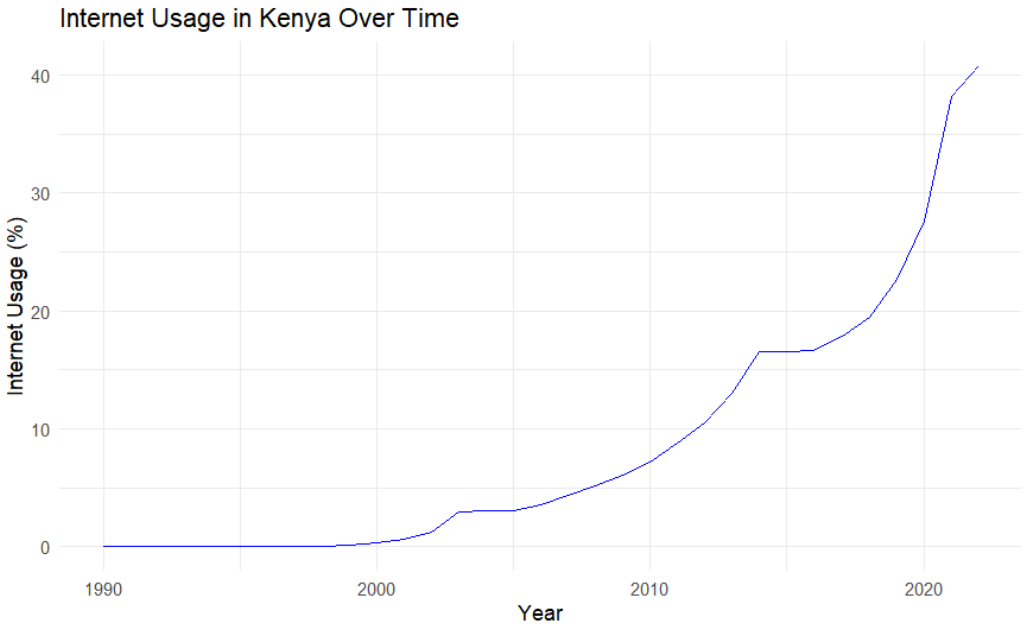

Figure 5. Internet Usage in Kenya over the years (1990-2020)

The image presents a line graph titled "Internet Usage in Kenya Over Time." It illustrates the trend of internet usage in Kenya from 1990 to 2020.

The graph shows a significant increase in internet usage over the years. It starts with a very low usage rate in the early 1990s, remaining almost flat until around 2000. Then, there is a gradual increase, followed by a more rapid growth in the 2010s. The line reaches its peak in 2020, indicating a substantial increase in internet penetration in Kenya during this period.

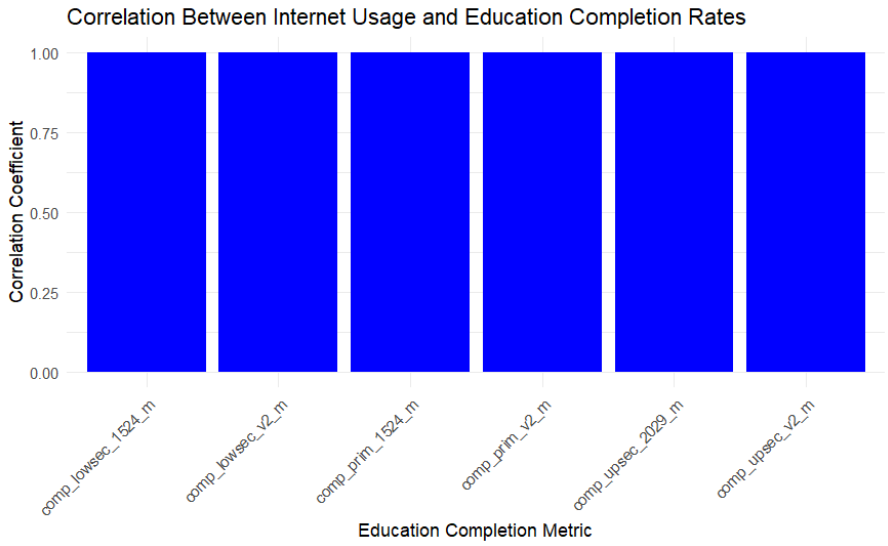

Figure 6. Correlation Analysis

The image presents a bar graph titled "Correlation Between Internet Usage and Education Completion Rates." It displays the correlation coefficients between internet usage and various education completion metrics.

Each bar on the graph represents a different education completion metric, labelled on the x-axis. The height of each bar corresponds to the correlation coefficient between that metric and internet usage. The y-axis shows the correlation coefficient values, ranging from 0 to 1.

The graph shows that the correlation coefficients are all very high, close to 1. This indicates a strong positive correlation between internet usage and all the education completion metrics. This suggests that increased internet usage is associated with higher education completion rates.

However the correlation value is too high due to data sparsity but we have taken the value to consideration due to lack of other data sources.

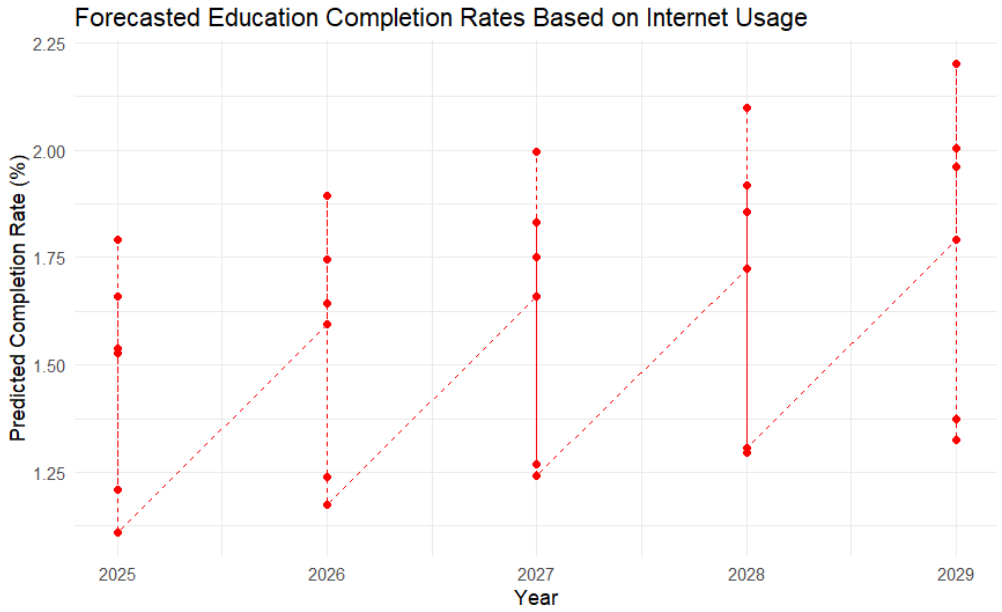

Figure 7. Forecasted Education Completion Rate (2025-2029)

The image presents a scatter plot titled "Forecasted Education Completion Rates Based on Internet Usage." It shows the predicted education completion rates for the years 2025 to 2029.

The x-axis represents the years, and the y-axis represents the predicted completion rate as a percentage. The graph includes red dots to mark the predicted completion rate for each year. Vertical lines connect each dot to the x-axis, and dashed lines connect the dots across different years, indicating the trend in the predicted completion rates.

The graph suggests a generally increasing trend in education completion rates over the years. The dashed lines connecting the dots show a slight upward slope, indicating that the predicted completion rate is expected to increase gradually each year. However, the increase is not uniform, with some years showing a steeper increase than others.

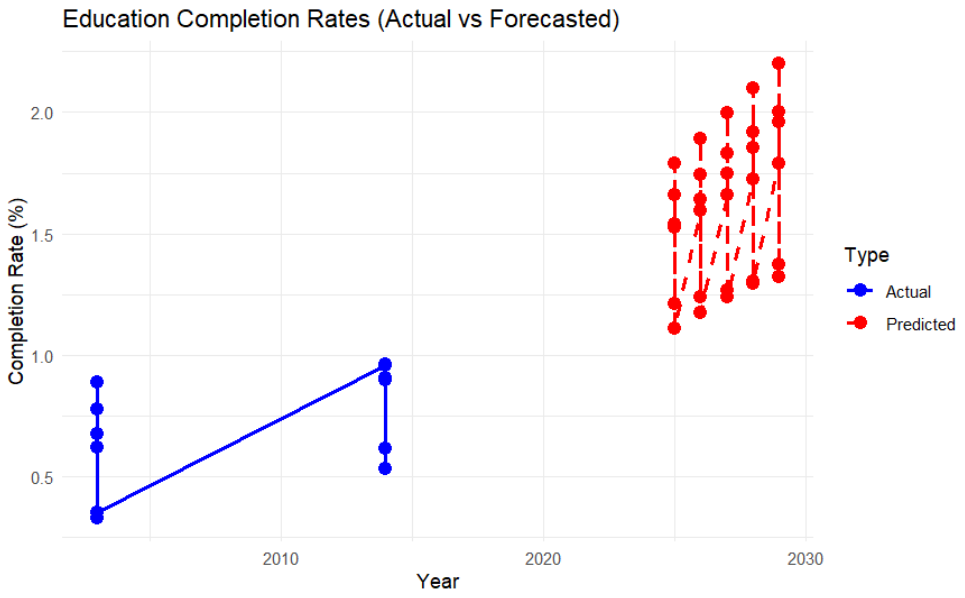

Figure 8. Comparison of Past Education Completion Rate along with Forecasted Education Completion Rate (2025-2029)

The image presents a scatter plot titled "Forecasted Education Completion Rates Based on Internet Usage." It shows the predicted education completion rates for the years 2025 to 2029.

The x-axis represents the years, and the y-axis represents the predicted completion rate as a percentage. The graph includes red dots to mark the predicted completion rate for each year. Vertical lines connect each dot to the x-axis, and dashed lines connect the dots across different years, indicating the trend in the predicted completion rates.

The graph suggests a generally increasing trend in education completion rates over the years. The dashed lines connecting the dots show a slight upward slope, indicating that the predicted completion rate is expected to increase gradually each year. However, the increase is not uniform, with some years showing a steeper increase than others.

6. SDG Goal & Contribution:

Goal No: 4

Title: Quality Education

The project aimed at analyzing the impact of internet usage on educational outcomes in Kenya makes significant contributions toward achieving several Sustainable Development Goals (SDGs), particularly SDG 4: Quality Education. By investigating the relationship between internet access and educational performance, the project underscores the role of technology in improving access to education, fostering equitable learning opportunities, and enhancing overall educational quality.

- **Improving Access to Education (SDG 4.1):**

  One of the core objectives of SDG 4 is to ensure inclusive and equitable education for all. In many regions, especially in developing countries like Kenya, limited access to educational resources presents a barrier to academic success. By focusing on internet usage, this project highlights the importance of digital technology as a tool for bridging the education gap. With increased internet penetration, students have access to a wealth of online educational resources, including e-books, online courses, and interactive learning tools. This enables students, regardless of their socioeconomic status or location, to benefit from quality education, thus contributing to the achievement of SDG 4.1. Our project has proved the significant correlation between internet and education there by indicating that quality education could be provided with the help of technology which could be inferred from higher completion rate at the current year.

- **Enhancing Education Quality through Technology (SDG 4.4)**

  As part of SDG 4, the goal is also to improve the quality of education by ensuring that all youth and adults acquire the necessary skills for employment and entrepreneurship. The project demonstrates how internet usage can directly enhance the quality of education by enabling

access to digital learning platforms and educational apps. These platforms support interactive and personalized learning, which can result in improved educational outcomes. Students with access to the internet can benefit from e-learning, online tutorials, and digital research tools, fostering better academic performance and preparing them with the digital skills necessary for the modern workforce. This directly supports SDG 4.4, which aims to develop relevant skills for the digital economy. Also the literature review done for this project also indicates the significance of technology in education thereby contributing to SDG 4.4.

- **Promoting Equal Education Opportunities (SDG 4.5)**

  Another crucial aspect of SDG 4 is ensuring that education systems provide equal opportunities for all. The project emphasizes the role of internet access in levelling the playing field for students from diverse backgrounds. Internet access can help reduce inequalities in education, particularly in rural or underserved areas where educational resources are limited. By advocating for increased internet infrastructure and affordable data access, this project promotes policies that aim to provide equal educational opportunities for all students, irrespective of their geographic location or economic background, contributing directly to SDG 4.5.

Additional Contribution of our Project to other SDG Goals:

- **Fostering Innovation in Education (SDG 9.5)**

  In addition to improving education quality, this project contributes to SDG 9: Industry, Innovation, and Infrastructure. It highlights how the integration of technology into education can foster innovation in teaching and learning. By encouraging the use of digital tools, online courses, and virtual classrooms, the project pushes for the development of innovative solutions to address challenges in education. Moreover, it advocates for the enhancement of digital infrastructure, making it a key enabler for promoting innovation in the educational sector. This aligns with SDG 9.5, which focuses on promoting the adoption of technology and innovation in various sectors, including education.

- **Empowering Youth and Enhancing Employability (SDG 8)**

  SDG 8: Decent Work and Economic Growth emphasizes the importance of education in empowering youth and increasing their employability. By improving educational outcomes through the use of internet resources, this project helps create a generation of youth equipped with the necessary skills for employment and entrepreneurship. Access to the internet exposes students to global job markets, internship opportunities, and digital skills development, which enhances their chances of securing decent work. By contributing to a better-educated and more skilled workforce, the project indirectly supports SDG 8 by empowering youth to participate in the digital economy and contributing to sustainable economic growth.

- **Addressing Global Challenges through Partnerships (SDG 17)**

  The success of increasing internet access for educational purposes depends on collaborative efforts between government entities, private sector stakeholders, and non-governmental organizations (NGOs). The project advocates for multi-sector partnerships to improve the digital infrastructure and support the implementation of education-focused initiatives. By fostering public-private partnerships, the project promotes shared responsibility in creating sustainable solutions for expanding internet access, improving e-learning platforms, and ensuring that digital education tools are affordable and accessible to all students. This fosters collaboration that is essential for achieving the broader SDG 17, which emphasizes partnerships for sustainable development.

7. Conclusion:

This project, which explores the impact of internet usage on educational outcomes in Kenya, underscores the significant role that digital technology can play in enhancing educational quality and accessibility. Through the analysis of internet penetration and its relationship with educational performance, the project highlights the transformative power of internet access in bridging educational disparities, especially in underserved regions. It reveals that internet usage can contribute to improved learning outcomes by providing students with access to a wide range of educational resources, digital tools, and e-learning platforms, thus fostering better academic performance.

The project also emphasizes the need for equitable access to the internet to ensure that all students, regardless of their socioeconomic status or geographical location, benefit from the opportunities that digital learning offers. By examining the relationship between internet usage and educational indicators, the project advocates for targeted interventions to improve digital infrastructure and increase access to the internet, particularly in rural and marginalized areas. This is crucial for achieving Sustainable Development Goal (SDG) 4: Quality Education, which aims to ensure inclusive, equitable, and quality education for all.

Additionally, the project highlights how technology-driven education can enhance students' digital literacy and prepare them for future employment opportunities. By fostering innovation in education through digital tools and resources, it contributes to the broader goals of SDG 9: Industry, Innovation, and Infrastructure, as well as SDG 8: Decent Work and Economic Growth by equipping youth with the skills needed for the digital economy.

Furthermore, the project includes a forecasting component that uses simple linear regression to predict education rates for the years 2025 to 2029. The forecasted values show a positive trend, indicating an increase in education rates over the next five years. This increase is attributed to the growing role of technology in education, reinforcing the idea that internet access is a key driver for improving quality education. The projected improvement in education rates highlights the potential long-term benefits of

integrating digital learning into educational systems, thereby supporting the sustainable development of human capital.

In conclusion, this project makes a compelling case for the integration of internet access into education systems as a means of improving learning outcomes, increasing educational opportunities, and preparing a skilled workforce for the future. By promoting digital inclusion, it contributes to the broader global effort to ensure sustainable development and to create a more equitable and prosperous society for all. Through collaboration between government, private sector, and NGOs, it is possible to create lasting, sustainable changes that will benefit education and youth development for years to come.

8. References:

I. Ogembo, J. G., Ngugi, B., & Pelowski, M. (2012). Computerizing primary schools in rural Kenya: Outstanding challenges and possible solutions. The Electronic Journal of Information Systems in Developing Countries, 52(1), 1-17.

II. Asongu, S., & Agyemang-Mintah, P. (2024). The relationship between inequality and poverty in developing countries: mitigating role of virtual social network and internet access in schools. International Journal of Social Economics.

III. Waithaka, M., Onyancha, O. B., & Ngulube, P. (2018). Internet use among university students in Kenya: a case study of the University of Nairobi. Innovation: journal of appropriate librarianship and information work in Southern Africa, 2018(57), 45-69.

IV. Castillo, N. M., Adam, T., & Haßler, B. (2022). 4. Improving the Impact of Educational Technologies on Learning Within Low-Income Contexts. Learning, Marginalization, and Improving the Quality of Education in Low-income Countries, 113.

V. Barteit, S., Guzek, D., Jahn, A., Bärnighausen, T., Jorge, M. M., & Neuhann, F. (2020). Evaluation of e-learning for medical education in low-and middle-income countries: A systematic review. *Computers & education*, *145*, 103726.

VI. Frehywot, S., Vovides, Y., Talib, Z., Mikhail, N., Ross, H., Wohltjen, H., ... & Scott, J. (2013). E-learning in medical education in resource constrained low-and middle-income countries. *Human resources for health*, *11*, 1-15.

VII. Frias-Martinez, V., Virseda, J., & Gomero, A. (2012, September). Mobilizing education: evaluation of a mobile learning tool in a low-income school. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services* (pp. 441-450).

VIII. Jaggars, S. (2011). Online learning: Does it help low-income and underprepared students? (assessment of evidence series).

IX. Rodriguez-Segura, D. (2020). Educational technology in developing countries: A systematic review. *University of Virginia EdPolicy Works Working Papers. Retrieved December*, *17*, 2021.

X. Valk, J. H., Rashid, A. T., & Elder, L. (2010). Using mobile phones to improve educational outcomes: An analysis of evidence from Asia. *International Review of Research in Open and Distributed Learning*, *11*(1), 117-140.

XI. Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Understanding the role of digital technologies in education: A review. *Sustainable operations and computers*, *3*, 275-285.

9. Data Source Link:

- GitHub repository link: https://github.com/ShashankS1011/Data-Analytics-Project-SDG-.git

- Internet usage dataset link: https://data.worldbank.org/indicator/IT.NET.USER.ZS

- Education rate: https://www.education-inequalities.org/indicators/literacy_1524/kenya/communities#year=%222022%22&ageGroup=%22literacy_1524%22&dimension1=%7B%22id%22%3A%22community%22%2C%22filters%22%3A%5B%22Rural%22%2C%22Urban%22%5D%7D