FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY DEPARTMENT OF COMPUTING AND TECHNOLOGY EASTER 2025 SEMESTER EXAMINATION - DATA SCIENCE LIFE CYCLE FINAL EXAMS

Efprem Okello - J25M19/001, Access Number - B31324[a,*]

[a] *Department of Computing and Technology, P.O. Box 4, Kampala, Uganda*

Abstract

This report explores three key themes: Sentiment Analysis, AI in Education, and Financial Inclusion in East Africa, showcasing the power of data-driven insights to address modern challenges. Sentiment Analysis: Analyzes Amazon reviews to understand consumer emotions and preferences. Findings reveal balanced sentiments, highlighting product quality and usability, helping businesses improve customer relationships and products. AI in Education: Demonstrates how AI enhances higher education by analyzing student performance and improving assessments. Predictive models, like Logistic Regression, show AI's potential to boost academic success and teaching effectiveness. Financial Inclusion: Examines financial access in East Africa, where innovations like mobile money have expanded services. However, gaps remain for marginalized groups. Data analysis identifies digital payments and technology as key drivers, calling for targeted interventions to promote equity. In summary, this report highlights how data-driven approaches can transform understanding of human behavior, improve education, and advance financial inclusion, offering actionable insights for progress.

Keywords: Sentiment Analysis, AI in Education, Financial Inclusion

1. Theme 2: Human Behaviour - Sentimental Analysis / Natural Language processing of human text and opinions

1.1. Introduction

The growth of digital content creation and social media has led to an unprecedented increase in unstructured text data. Every day, people express their thoughts and opinions through reviews, comments, and feedback, generating vast amounts of user-generated content (Liu, 2012). This data holds valuable insights into human emotions, preferences, and behaviors. Sentiment analysis has emerged as a powerful tool for understanding these emotions (Pang and Lee, 2008). By analyzing text, it helps uncover the underlying feelings and thoughts people express. The ability to extract sentiment from unstructured data provides meaningful insights, making it indispensable for businesses, researchers, and anyone seeking to build a deeper connection with their audience (Cambria et al., 2013).

As one of the world's largest e-commerce platforms, Amazon receives a constant stream of customer feedback from product reviews to service ratings (Hu and Liu, 2004). This wealth of user-generated content holds valuable insights into consumer sentiment, shedding light on what people love, what frustrates them, and what they expect from their shopping experiences. By applying Sentiment Analysis to Amazon's customer feedback, we can uncover meaningful patterns in consumer behavior (Zhang et al., 2018). These

*Corresponding author

Email address: okelloefprem@gmail.com (Efprem Okello - J25M19/001, Access Number - B31324)

insights go beyond just understanding customer satisfaction and product preferences; they can also reveal potential areas for improvement and even broader societal trends, such as how online shopping impacts mental well-being and social dynamics.

1.2. Methodology

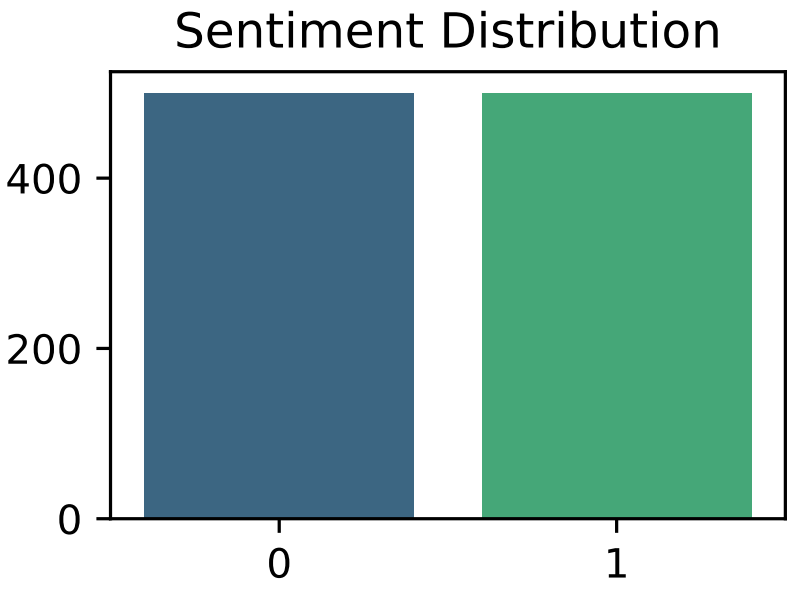
The analysis followed a structured approach, incorporating data collection, preprocessing, exploratory analysis, and modeling to extract insights from Amazon customer reviews.

- **Data Collection and Pre-processing:** Sentiment-labeled data set containing customer feedback was obtained from the UC Irvine Machine Learning Repository and imported into python for analysis. The text data underwent pre-processing, where it was converted to lowercase, special characters, stop-words, and short words were removed. This enhanced the relevance of the data and allowed for more accurate analysis.
- **Exploratory Data Analysis:**In the exploratory data analysis phase, we examined the sentiment distribution, revealing an even split between positive and negative reviews. A word frequency analysis was conducted to identify key themes in customer feedback. Words such as "phone," "great," and "good" were frequently mentioned, highlighting common sentiments. A word cloud visualization was created to graphically represent the most common terms, focusing on product quality and usability.
- **Sentiment Classification:** For sentiment classification, we applied TF-IDF vectorization to convert text data into numerical format. A Naïve Bayes classifier was then trained to predict sentiment, and the model was evaluated using accuracy scores, classification reports, and a confusion matrix. We also validated the model with hypothetical test samples, demonstrating its ability to predict sentiment.
- **Unsupervised Learning (Clustering):** To provide an unsupervised perspective, we implemented K-Means clustering, grouping the reviews into two sentiment-based clusters. This clustering approach helped us explore sentiment trends in the data without predefined labels.

1.3. Findings

1.3.1. Sentiment Distribution

The sentiment analysis reveals a perfectly balanced distribution between positive and negative feedback. Out of the total dataset, 500 reviews (50%) express negative sentiment (Sentiment = 0), while 500 reviews (50%) convey positive sentiment (Sentiment = 1). This even split suggests a diverse range of customer experiences, with equal representation of satisfaction and dissatisfaction.



1.3.2. Word Cloud Visualization

The text analysis reveals key themes in customer feedback, with "phone" being the most frequently mentioned word, appearing 168 times. This suggests that many reviews focus on mobile devices, likely discussing their performance, features, or overall satisfaction. The words "great" (99) and "good" (77) indicate a generally positive sentiment, implying that customers are largely satisfied with their purchases. "Product" (55) and "quality" (49) further reinforce this, highlighting that shoppers often comment on the overall value and craftsmanship of their items. Specific mentions of "headset" (48), "battery" (46), and "sound" (43) suggest that many reviews relate to audio devices, possibly wireless headphones or earphones. The frequent appearance of "works" (47) implies that customers often evaluate whether the product functions as expected. Additionally, "use" (41) indicates that usability is an important factor in customer experiences. Overall, the feedback suggests that customers are generally pleased with the quality and performance of their purchases, particularly in relation to phones and audio accessories.

```
## (-0.5, 799.5, 399.5, -0.5)
```

1.4. Build, Evaluate and Optimize models

1.4.0.1. Model Evaluation. The model demonstrates a solid performance with an overall accuracy of 81%, correctly predicting the class in 81% of the cases. It performs well across both classes, with Class 1 showing slightly better results. For Class 0, the model achieves a precision of 82% and a recall of 75%, indicating that it correctly identifies 82% of the predicted Class 0 instances, though it misses about 25% of actual Class 0 instances. For Class 1, precision stands at 80%, and recall is higher at 86%, meaning it identifies 86% of the true Class 1 instances but with a slightly lower precision. The F1-scores are balanced for both classes, with 0.79 for Class 0 and 0.83 for Class 1, reflecting a strong trade-off between precision and recall. The model's performance is consistent across both classes, as indicated by the macro and weighted averages of 81% for precision, recall, and F1-score, showing a well-rounded and effective classification model.

```
## MultinomialNB()
```

```
## <string>:3: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
```

```
## ==================================================
```

```
## Model Evaluation Metrics
```

```
## ==================================================
```

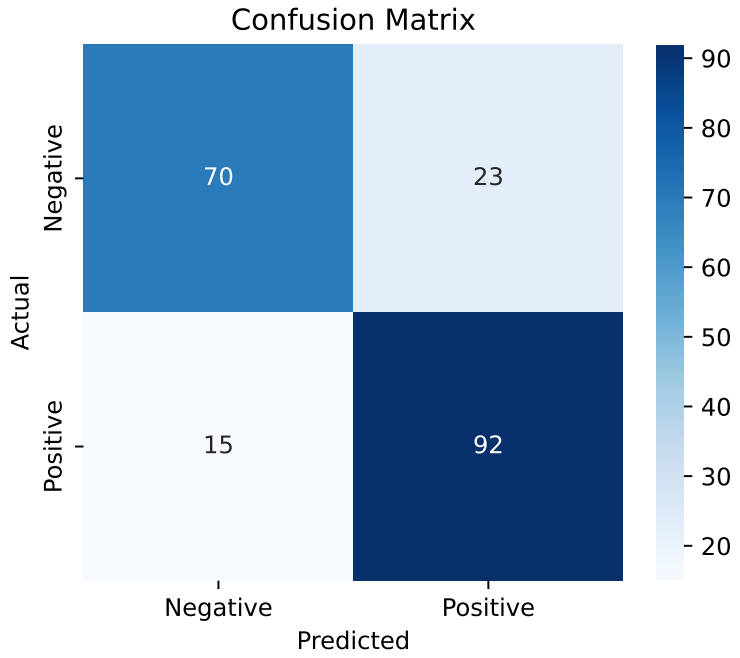
```
## Accuracy Score: 0.81
```

```
## Classification Report:
```

```
## +--------------+-----------+--------+----------+---------+
## |              | precision | recall | f1-score | support |
## +--------------+-----------+--------+----------+---------+
## |      0       |   0.82    |  0.75  |   0.79   |  93.0   |
## |      1       |    0.8    |  0.86  |   0.83   |  107.0  |
## |   accuracy   |   0.81    |  0.81  |   0.81   |  0.81   |
## |  macro avg   |   0.81    |  0.81  |   0.81   |  200.0  |
## | weighted avg |   0.81    |  0.81  |   0.81   |  200.0  |
## +--------------+-----------+--------+----------+---------+

## ================================================
```

1.4.0.1.1. Confusion Matrix The confusion matrix reveals how well the model predicts Negative and Positive cases. The model correctly identified 70 Negative cases and 92 Positive cases. However, it incorrectly predicted 23 Negative cases as Positive and 15 Positive cases as Negative. Overall, the model performs well, but there is room for improvement, particularly in reducing the number of False Positives and False Negatives.



2. Theme 4: Education: Artificial Intelligence (AI) in Higher Education

2.1. Introduction

Education is a key driver of national development, and improving the quality of higher education is essential for progress. One way to achieve this is by using faster methods to assess student performance and teaching effectiveness. Traditional assessment methods can be slow and limited (Bennett, 2011), but Artificial Intelligence (AI) offers a promising solution (Luckin et al., 2016). AI can analyze large amounts of data quickly and identify patterns that can help improve both teaching methods and student outcomes (Baker and Inventado, 2014). In higher education, machine learning algorithms, a type of AI, can be used to classify student performance and evaluate teaching approaches. By analyzing data from student assessments and other academic activities, these algorithms can provide valuable insights. The goal of this study is to find the most effective approach for improving student performance.

2.2. Methodology

This study aims to predict student performance, specifically determining whether a student will pass or fail based on various factors. The process involved several key steps, including data collection, preprocessing, exploratory analysis, model building, evaluation, and validation.

- **Data Collection and Preprocessing:** The dataset used in the study contains various student characteristics such as their previous grades, study time, age, and alcohol consumption (both on weekdays and weekends). In the preprocessing stage, categorical variables like whether a student wants to pursue higher education were converted into numeric values using label encoding. Additionally, the target variable, the final grade (denoted as G3), was transformed into a binary classification problem. If a student's final grade was 10 or more, they were considered to have "passed" (label 1), while those with grades below 10 were categorized as "fail" (label 0).

- **Exploratory Data Analysis:** Following preprocessing, an in-depth analysis of the data was conducted to understand the relationships between different variables and their impact on student performance. Histograms and box plots were created to visually examine the distribution of grades and to highlight the study time's influence on the final grade. A correlation matrix was also generated to identify the factors most closely related to student performance. This analysis revealed that previous grades and study time were the most significant predictors of final performance.

- **Model Building:** In the next phase, various features were selected for use in the predictive model. The selected features included previous grades, weekly study time, the number of past class failures, and other factors like age, and alcohol consumption. The dataset was split into training and testing sets to ensure an unbiased evaluation of the models. Three different machine learning algorithms were applied: Logistic Regression, Random Forest, and Gradient Boosting. These models were trained using the training dataset, and their performance was evaluated on the testing set.

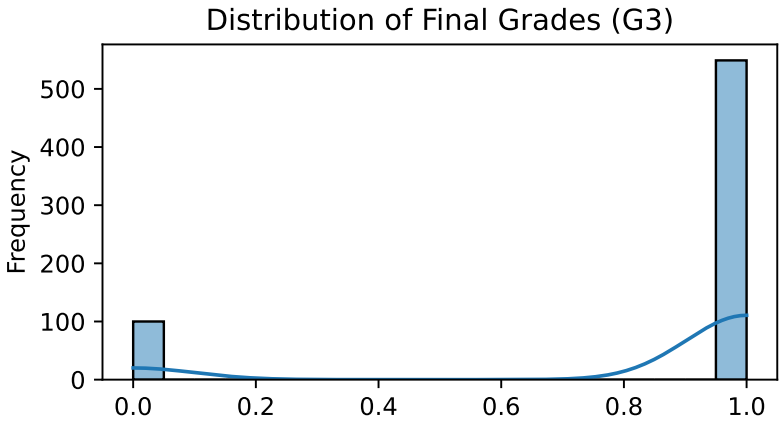
- **Model Evaluation:** The models were evaluated based on their accuracy, precision, recall, and F1-score. Logistic Regression provided the best overall performance, as it showed the highest accuracy in predicting whether a student would pass or fail. Random Forest and Gradient Boosting also performed well but did not outperform Logistic Regression in this study. Based on these results, Logistic Regression was selected as the best model for predicting student performance.

- **Validation:** To validate the effectiveness of the model, a sample student's data was used to predict their final grade. The student's characteristics—such as age, study time, and alcohol consumption—were input into the trained Logistic Regression model, which predicted whether they would pass or fail. This validation step confirmed the model's capability to make accurate predictions based on the input features.

2.3. Findings

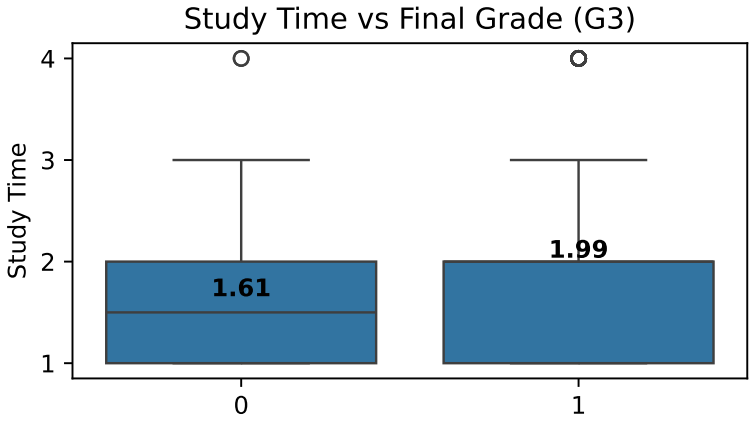
2.3.1. Historical student performance

The distribution of final grades (G3) reveals that most students pass their final exams.Out of 649 students, 549 (84.6%) passed and 100 (15.4%) failed, indicating a generally high pass rate.



2.3.2. Average study time vs Final Grade

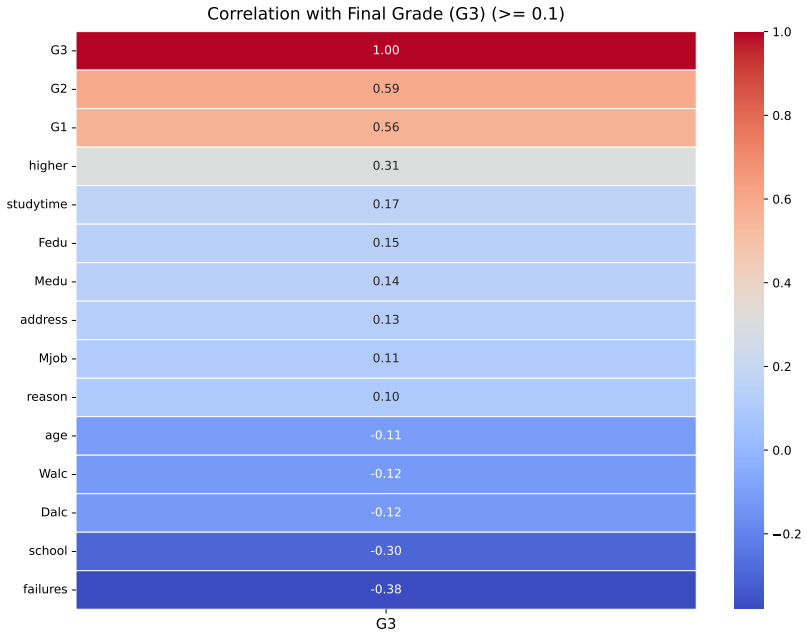
The average study time for students reveals a notable difference between those who passed and those who failed. Students who passed (G3 = 1) spent an average of 1.99 hours per week studying, while those who failed (G3 = 0) averaged only 1.61 hours per week. This indicates a positive correlation between study time and academic performance, suggesting that increased study time contributes to better outcomes.



2.3.3. Predictors of Final Grade

The correlation analysis reveals that the strongest predictors of final grades (G3) are early academic performance (G1 and G2) and aspirations for higher education (higher), both showing significant positive correlations. Moderate positive correlations exist for study time and parental education, suggesting their supportive role in academic success. Conversely, past failures, alcohol consumption, and absences have strong negative correlations, indicating they are major barriers to performance. Factors like family support, extracurricular activities, and guardian type have minimal impact on grades.

```
## (array([0.5]), [Text(0.5, 0, 'G3')])
```



2.3.4. Selected Models

2.3.4.1. Model 1: Logistic Regression.

The logistic regression model achieved an overall accuracy of 89.2%. The model performs exceptionally well in predicting class 1, with a precision of 92%, a recall of 97%, and an F1-score of 94%, based on 115 instances. However, its performance in predicting class 0 is weaker, with a precision of 56%, a recall of 33%, and an F1-score of 42%, based on 15 instances. The macro average F1-score is 68%, reflecting the imbalance in class performance, while the weighted average F1-score of 88% indicates strong overall predictive capability.

```
## LogisticRegression(random_state=42)

## <string>:3: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.

## ==================================================

## Logistic Regression Evaluation Metrics

## ==================================================

## Accuracy Score: 0.89

## Classification Report:

## +--------------+-----------+--------+----------+---------+
## |              | precision | recall | f1-score | support |
## +--------------+-----------+--------+----------+---------+
## |      0       |   0.56    |  0.33  |   0.42   |  15.0   |
## |      1       |   0.92    |  0.97  |   0.94   |  115.0  |
## |   accuracy   |   0.89    |  0.89  |   0.89   |  0.89   |
## |  macro avg   |   0.74    |  0.65  |   0.68   |  130.0  |
## | weighted avg |   0.88    |  0.89  |   0.88   |  130.0  |
## +--------------+-----------+--------+----------+---------+

## ==================================================
```

2.3.4.2. Model 2: Random Forest Classifier.
The Random Forest model achieved an overall accuracy of 84.6%. The model performs well in predicting class 1, with a precision of 91%, a recall of 92%, and an F1-score of 91%, based on 115 instances. However, its performance in predicting class 0 is considerably lower, with a precision of 31%, a recall of 27%, and an F1-score of 29%, based on 15 instances. The macro average F1-score of 60% highlights this class imbalance, while the weighted average F1-score of 84% suggests strong overall predictive ability, driven mainly by class 1 performance.

```
## RandomForestClassifier(random_state=42)

## <string>:3: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.

## ==================================================

## Random Forest Evaluation Metrics

## ==================================================

## Accuracy Score: 0.85

## Classification Report:

## +--------------+-----------+--------+----------+---------+
## |              | precision | recall | f1-score | support |
## +--------------+-----------+--------+----------+---------+
## |      0       |   0.31    |  0.27  |   0.29   |  15.0   |
## |      1       |   0.91    |  0.92  |   0.91   |  115.0  |
## |   accuracy   |   0.85    |  0.85  |   0.85   |  0.85   |
## |  macro avg   |   0.61    |  0.59  |   0.6    |  130.0  |
## | weighted avg |   0.84    |  0.85  |   0.84   |  130.0  |
## +--------------+-----------+--------+----------+---------+

## ==================================================
```

2.3.4.3. Model 3: Gradient Boosting Classifier.
The Gradient Boosting model achieved an overall accuracy of 85.4%. It performed well in predicting class 1, with a precision of 92%, a recall of 91%, and an F1-score of 92%, based on 115 instances. For class 0, the model showed moderate performance, with a precision of 38%, a recall of 40%, and an F1-score of 39%, based on 15 instances. The macro average F1-score of 65% reflects this disparity, while the weighted average F1-score of 86% indicates that the model maintains strong overall predictive performance, primarily driven by class 1.

```
## GradientBoostingClassifier(random_state=42)

## <string>:3: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.

## ==================================================

## Gradient Boosting Evaluation Metrics

## ==================================================

## Accuracy Score: 0.85

## Classification Report:
```

```
## +--------------+-----------+--------+----------+---------+
## |              | precision | recall | f1-score | support |
## +--------------+-----------+--------+----------+---------+
## |      0       |   0.38    |  0.4   |   0.39   |  15.0   |
## |      1       |   0.92    |  0.91  |   0.92   |  115.0  |
## |   accuracy   |   0.85    |  0.85  |   0.85   |  0.85   |
## |  macro avg   |   0.65    |  0.66  |   0.65   |  130.0  |
## | weighted avg |   0.86    |  0.85  |   0.86   |  130.0  |
## +--------------+-----------+--------+----------+---------+

## ================================================
```

2.3.4.4. Model performance comparison.
The Logistic Regression model is the recommended choice, achieving the highest accuracy of 89.2%. It demonstrated strong performance in predicting class 1, with a precision of 92%, a recall of 97%, and an F1-score of 94%, based on 115 instances. Although its performance for class 0 was lower (precision: 56%, recall: 33%, F1-score: 42% for 15 instances), the overall weighted F1-score of 88% indicates robust predictive ability. Given its superior accuracy and balanced performance, Logistic Regression is the most reliable model for this classification task.

```
##
## Recommended Model: Logistic Regression Results with Accuracy: 0.8923076923076924
```

2.3.5. Model Validation

For a 31-year-old student who studies 5 to 10 hours per week, has no record of past failures, aspires to pursue higher education, and consumes alcohol once on weekdays and twice on weekends, the trained Logistic Regression model predicted a pass. This outcome aligns with expectations, as the student's consistent study habits, lack of academic failures, and motivation for higher education are strong indicators of academic success.

```
## Predicted Final Grade (G3) for the student: 1
```

3. Theme 3: : Finance- Financial Access and Inclusion

3.1. Introduction

Financial access and inclusion are critical drivers of economic growth and poverty reduction. Across Africa, the expansion of credit, funding opportunities, and digital financial platforms—such as mobile money—has significantly improved financial access (Suri and Jack, 2016). These innovations have facilitated transactions, savings, and credit acquisition for millions, promoting financial empowerment and economic participation. However, despite these advancements, financial inclusion remains uneven, with marginalized groups such as women, the elderly, and rural populations facing persistent barriers to access (Allen et al., 2016). Socioeconomic disparities, limited financial literacy, inadequate infrastructure, and restrictive financial policies contribute to their exclusion, limiting their ability to fully benefit from financial services. This analysis seeks to examine financial access and service usage data to identify existing gaps and inequalities. By understanding these disparities, targeted interventions can be proposed to enhance inclusivity and ensure that financial systems cater to all, fostering equitable economic development.

3.2. Methodology

The methodology employed in this analysis is designed to comprehensively evaluate financial inclusion across East African countries (Burundi, Kenya, Rwanda, Tanzania, and Uganda) using a structured, data-driven approach.

- **Data Preparation:** The analysis began with loading the dataset (DatabankWide.xlsx) and selecting relevant variables that capture key dimensions of financial inclusion, such as account ownership, access to financial services, usage patterns, and socio-economic dynamics. The dataset is filtered to focus on East African countries, ensuring the analysis is region-specific. Missing values are addressed using a systematic imputation strategy: numeric variables are filled with the mean or median (depending on skewness), while categorical variables are filled with the mode. This ensures the dataset is complete and ready for analysis. To handle categorical variables, a combination of Label Encoding and One-Hot Encoding were applied.
- **Exploratory Data Analysis (EDA):**The EDA phase focused on understanding the relationships between variables and identifying key drivers of financial inclusion. A correlation analysis was conducted to determine which variables have the strongest association with the Financial Inclusion Index (FII). Variables with a correlation coefficient of 0.5 or higher were identified as significant contributors to financial inclusion. Visualizations, such as heatmaps and box plots, were used to explore the distribution of financial service usage and access across different demographics (e.g., gender, labor force status) and socio-economic dynamics (e.g., digital payments, savings behavior).
- **Feature Engineering:** To quantify financial inclusion, a Financial Inclusion Index (FII) was constructed using weighted scores for three dimensions:

i. **Ownership:** To measure account and card ownership.
ii. **Access:** As proxies of access to financial services using account and card ownership data.
iii. **Usage:** To evaluates the active use of financial services, such as making deposits or using debit/credit cards.

Each dimension was weighted (ownership: 40%, access: 30%, usage: 30%) to reflect its relative importance. The FII was then calculated as a composite score scaled to 0–100, providing a standardized metric for comparing financial inclusion across countries and demographics.

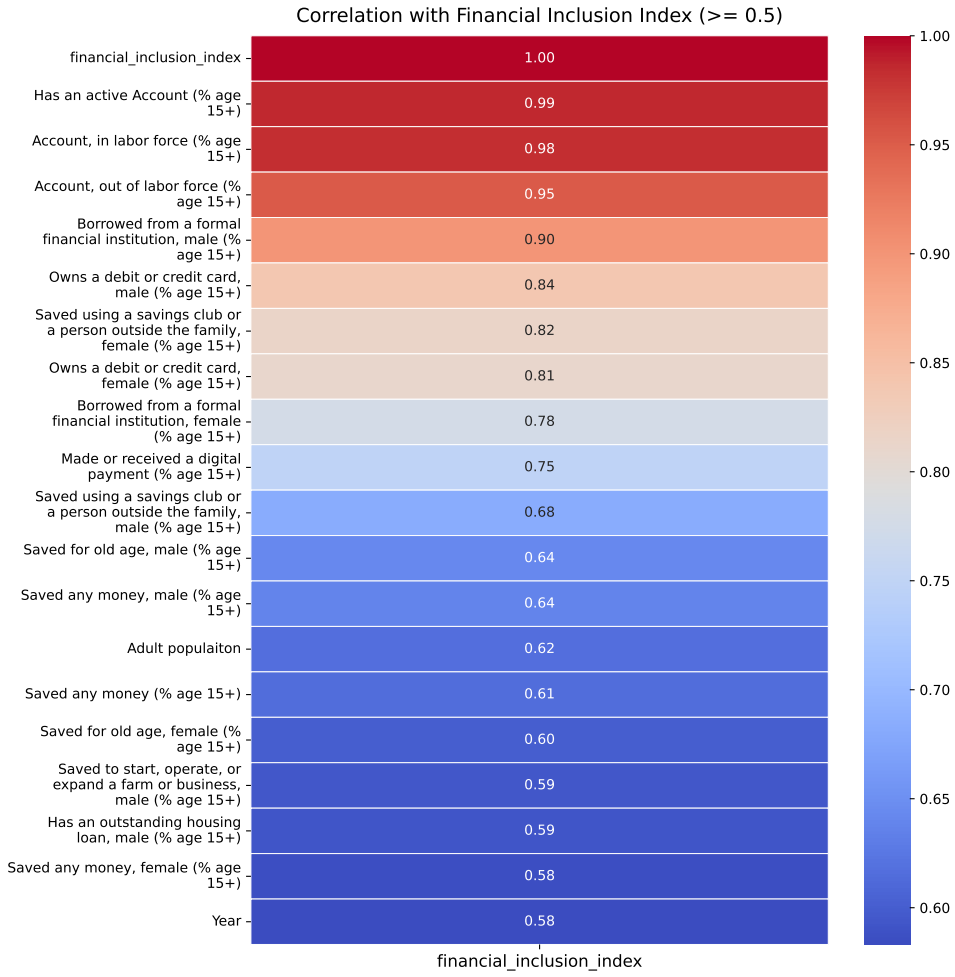
- **Predictive Modeling:** The analysis employs machine learning to predict the Financial Inclusion Index based on the identified significant variables. Five regression models are evaluated: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNN). Each model was trained on 80% of the data and evaluated on the remaining 20%. Performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) were used to compare the models. The Linear Regression model was selected for further validation due to its interpretability and competitive performance.
- **Model Evaluation and Interpretation:** The Linear Regression model was validated using 5-fold cross-validation to confirms its consistency across different subsets of the data. The model's coefficients were interpreted to understand the impact of each feature on the Financial Inclusion Index. For example: Positive coefficients indicate that higher values of the feature (e.g., digital payments, savings behavior) are associated with increased financial inclusion. Negative coefficients suggest that certain factors (e.g., inactive accounts) may hinder financial inclusion. Finally, the model was used to make predictions on the test set, and its performance is evaluated using MAE, MSE, and R^2. The results demonstrate that the model provides a reliable estimate of financial inclusion, with an R^2 value indicating a strong fit to the data.

3.3. Summary of Findings

3.3.1. Identification of variables that affect financial inclusion

The correlation analysis reveals that digital payments, savings behavior, and mobile money usage are the strongest drivers of financial inclusion, with correlations exceeding 0.85. Urban populations and individuals in the labor force show higher financial inclusion compared to rural and out-of-labor-force groups. Access to technology, such as mobile phones and the internet, also plays a significant role, with correlations above 0.70. Gender-specific analysis indicates that women and men exhibit similar trends, though women show slightly stronger correlations in areas like saving for education or old age.

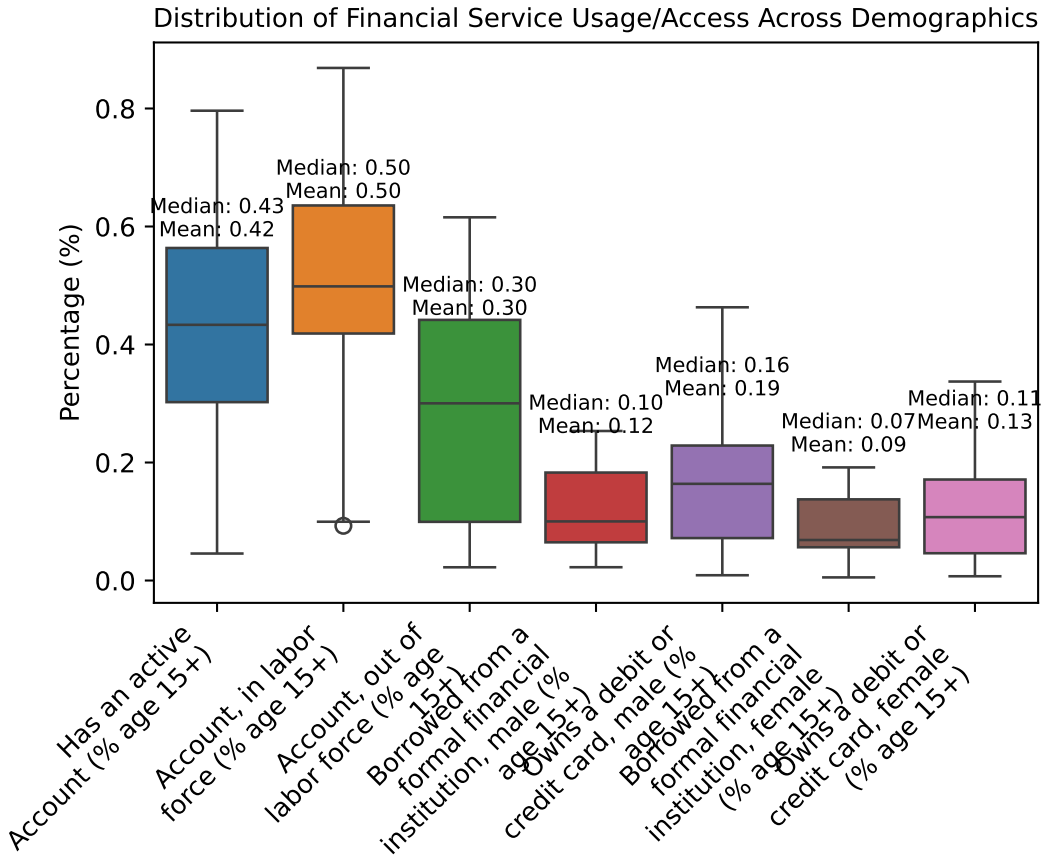
```
## (array([0.5]), [Text(0.5, 0, 'financial_inclusion_index')])
```



3.3.2. Financial Service-Usage/Access Distribution Across Population Demographics

The analysis shows that less than half of people aged 15+ actively use financial accounts (median: 43.34%). Those in the labor force have higher account ownership (median: 49.85%) compared to those not working (median: 30.03%), highlighting the impact of employment on financial access. Men are more likely to borrow from formal institutions (median: 10.02%) and own debit/credit cards (median: 16.39%) than women (borrowing: median 6.87%; card ownership: median 10.74%).

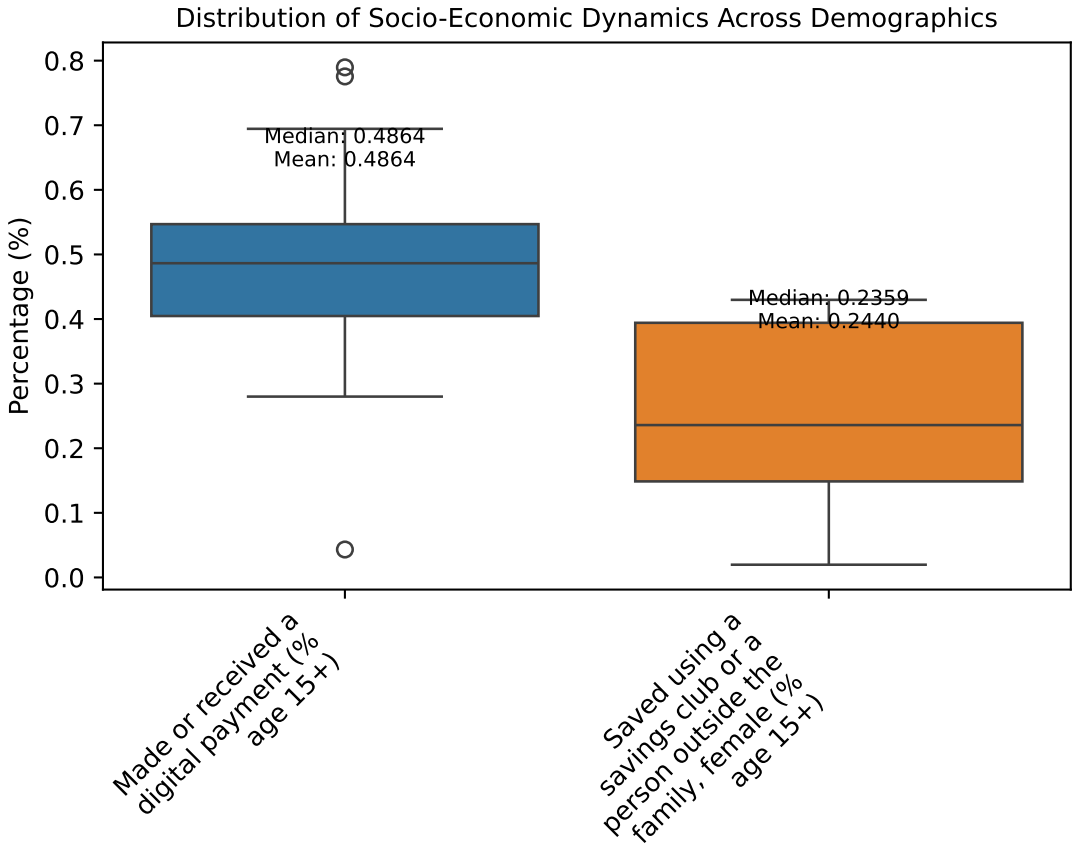
```
## ([<matplotlib.axis.XTick object at 0x0000021465 64AD50>, <matplotlib.axis.XTick object at 0x0000021
```



3.3.3. Financial Service-Usage/Access Distribution Across Socio-Economic Dynamics

The analysis of socio-economic dynamics reveals that digital payments are widely adopted, with a median of 48.64% of individuals aged 15+ having made or received digital payments, matching the mean of 48.64%. This indicates a balanced distribution and highlights the growing role of digital financial services in promoting inclusion. On the other hand, savings through informal channels, such as savings clubs or individuals outside the family, is less common among women, with a median of 23.59% and a mean of 24.40%. This suggests that while digital payments are becoming mainstream, informal savings mechanisms remain a secondary option for many women.

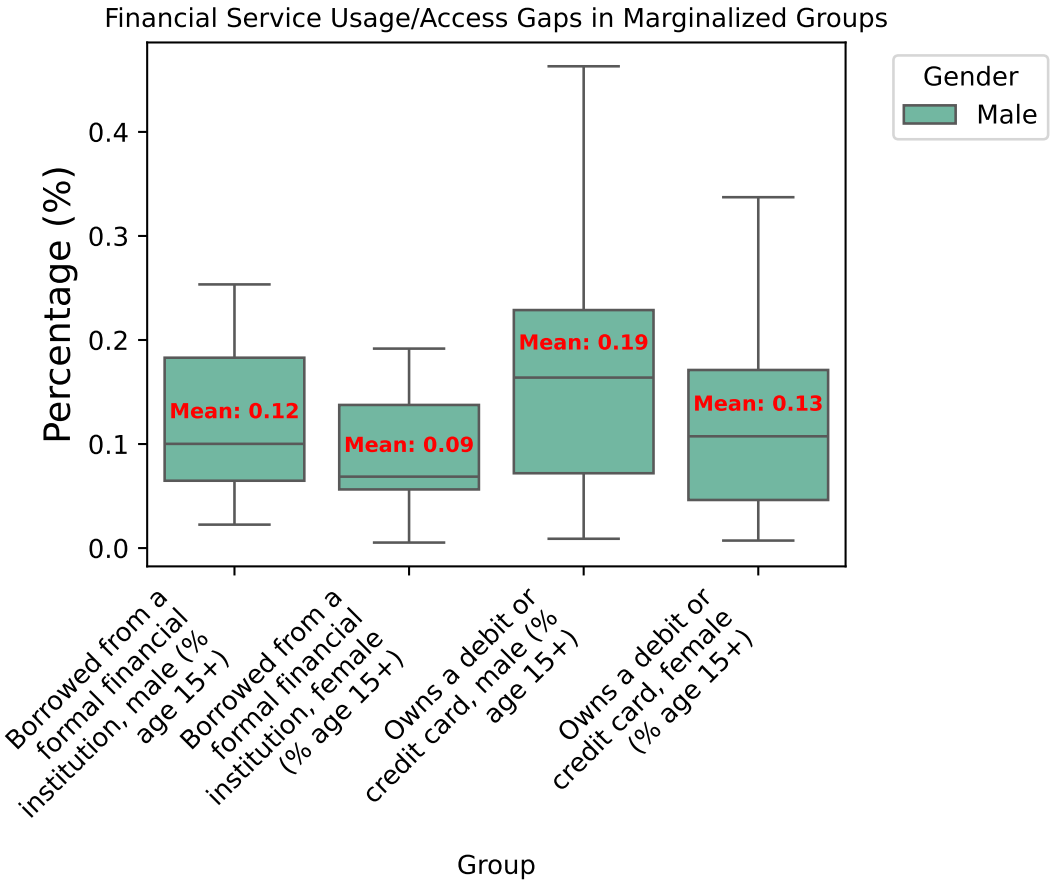
```
## ([<matplotlib.axis.XTick object at 0x000002145D722D50>, <matplotlib.axis.XTick object at 0x00000214
```



3.3.4. Identify Gaps in Financial Service-Usage/Access (Marginalized Groups)

The analysis of financial service usage and access among marginalized groups reveals notable disparities between men and women. On average, men are more likely to borrow from formal financial institutions (mean: 12.06%) and own debit/credit cards (mean: 18.67%) compared to women (borrowing: mean 8.83%; card ownership: mean 12.79%). These gaps highlight significant gender-based inequalities in access to formal financial services.

```
## ([<matplotlib.axis.XTick object at 0x0000021465648A90>, <matplotlib.axis.XTick object at 0x00000214
```



3.4. Step 4: Build and Evaluate Predictive Models

3.4.1. Model Comparison

The model performance comparison reveals that Linear Regression outperforms the other models across all evaluation metrics. It achieves the lowest Mean Absolute Error (MAE: 3.58) and Mean Squared Error (MSE: 17.72), as well as the highest R-squared (R^2: 0.95), indicating it explains 95% of the variance in the data. This makes it the most accurate and reliable model for predicting financial inclusion. The Gradient Boosting Regressor also performs well, with an R^2 of 0.86, but it has higher errors (MAE: 6.39, MSE: 49.5) compared to Linear Regression. The Random Forest Regressor follows with an R^2 of 0.79, but its errors are even higher (MAE: 6.9, MSE: 74.73). In contrast, K-Nearest Neighbors (KNN) and Support Vector Regressor (SVR) perform poorly. KNN has moderate accuracy (R^2: 0.64) but high errors (MAE: 10.32, MSE: 126.3), while SVR performs the worst, with an R^2 of only 0.09 and very high errors (MAE: 16.41, MSE: 321.71). In summary, Linear Regression is the best-performing model for this task, offering high accuracy and low prediction errors, making it suitable for analyzing financial inclusion.

```
## Model Performance Comparison:

## +---------------------------+-------+--------+------+
## |           Model           |  MAE  |  MSE   |  R²  |
## +---------------------------+-------+--------+------+
## |     LinearRegression      | 3.58  | 17.72  | 0.95 |
## |   RandomForestRegressor   |  6.9  | 74.73  | 0.79 |
```

```
## | GradientBoostingRegressor | 6.39  |  49.5  | 0.86 |
## |            SVR           | 16.41 | 321.71 | 0.09 |
## |    KNeighborsRegressor    | 10.32 | 126.3  | 0.64 |
## +---------------------------+-------+--------+------+
```

3.4.1.1. Validate the Linear Regression Model. The cross-validation results for the Linear Regression model demonstrate strong and consistent performance across different subsets of the data. The R^2 scores for the 5 folds are [0.97, 0.91, 0.93, 0.71, 0.71], indicating that the model explains between 71% and 97% of the variance in the data, depending on the subset. The mean R^2 score of 0.85 confirms that the model is highly reliable overall, capturing 85% of the variance on average.

```
## Cross-Validation R² Scores: [0.97 0.91 0.93 0.71 0.71]

## Mean R² Score: 0.85
```

3.4.1.2. Interpret the Model Coefficients. The Linear Regression coefficients reveal key drivers and barriers to financial inclusion. Borrowing by men and active account usage have the strongest positive impacts, significantly boosting financial inclusion. In contrast, borrowing by women and card ownership by men show negative effects, highlighting potential gender-based inequalities. Labor force participation (both in and out) is associated with lower financial inclusion, suggesting challenges in accessing formal financial services. Digital payments have a moderate positive effect, while informal savings by women have minimal impact.

```
## LinearRegression()

## Model Coefficients:

## +-----------------------------------------------------+-------------+
## |                       Feature                       | Coefficient |
## +-----------------------------------------------------+-------------+
## | Borrowed from a formal financial institution, male  |    87.76    |
## |                     (% age 15+)                     |             |
## |   Borrowed from a formal financial institution,     |   -86.43    |
## |                 female (% age 15+)                  |             |
## |         Has an active Account (% age 15+)           |    77.24    |
## |  Owns a debit or credit card, female (% age 15+)    |    64.54    |
## |   Owns a debit or credit card, male (% age 15+)     |   -42.31    |
## |        Account, in labor force (% age 15+)          |   -31.47    |
## |   Made or received a digital payment (% age 15+)    |    6.21     |
## |      Account, out of labor force (% age 15+)        |    -3.59    |
## | Saved using a savings club or a person outside the  |    0.64     |
## |             family, female (% age 15+)              |             |
## +-----------------------------------------------------+-------------+
```

3.4.1.3. Model Predictions. The Linear Regression model shows mixed performance in predicting financial inclusion. It tends to overestimate lower values (e.g., predicting 18.46 for an actual value of 11.46), indicating challenges in accurately predicting low-inclusion scenarios. However, it performs well for moderate to high inclusion levels, with predictions closely matching actual values (e.g., predicting 55.16 for an actual value of 54.3).

```
## Predictions vs Actual Values:

## +--------+-----------+
## | Actual | Predicted |
## +--------+-----------+
```

```
## | 11.46  |   18.46   |
## | 11.52  |   14.56   |
## |  54.3  |   55.16   |
## | 41.47  |   38.02   |
## +--------+-----------+
```

3.4.2. Model Evaluation

The Linear Regression model performs exceptionally well, with a low MAE (3.58) and MSE (17.72), indicating accurate predictions with minimal errors. Its high R^2 value (0.95) confirms it explains 95% of the variance in the data, making it a reliable and effective tool for predicting financial inclusion.

```
## Model Evaluation Metrics:

## +---------------------------+-------+
## |          Metric           | Value |
## +---------------------------+-------+
## | Mean Absolute Error (MAE) | 3.58  |
## | Mean Squared Error (MSE)  | 17.72 |
## |      R-squared (R²)       | 0.95  |
## +---------------------------+-------+
```

References

Allen, F., Carletti, E., Cull, R., Qian, J., Senbet, L., and Valenzuela, P. (2016). The african financial development and financial inclusion gaps. *Journal of African Economies*, 25(2):273–301.

Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. *Learning analytics*, pages 61–75.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1):5–25.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Springer.

Luckin, R., Holmes, W., Griffiths, M., and Forcier, L. B. (2016). *Intelligence Unleashed: An argument for AI in Education*. Pearson Education.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Suri, T. and Jack, W. (2016). The long-run poverty and gender impacts of mobile money. *Science*, 354(6317):1288–1292.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.