# (Concepts and Technologies of AI)
# 5CS037

# Classification Analysis (Report)

Name: Jyotika Ghale
Group: L5CG9
Student ID: 2408998
Module Leader: siman Giri
Tutor: Bibek Khanal
Submittion Date: 2025/02/11

# Table of Contents

# Predicting Digital Literacy Using Classification Techniques

## Abstract

### Aim:

This paper uses categorization approaches to predict digital literacy levels by analyzing demographic and socioeconomic data.

### Method:

The Digital Literacy Dataset from Kaggle was chosen for this investigation. The methodology includes feature selection, hyperparameter tuning, exploratory data analysis (EDA), data preprocessing, and the creation of Decision Tree and Logistic Regression models. Performance criteria like F1-score, accuracy, precision, and recall were used to assess the models.

### Key Findings:

The Decision Tree model marginally surpassed the Logistic Regression model regarding accuracy and recall. Important indicators of digital literacy involved age, educational attainment, and availability of internet.

## Summary:

The classification models correctly predicted levels of digital literacy. The enhancement of model performance by feature selection and hyperparameter tuning demonstrates the importance of education and internet accessibility in digital literacy. Better feature selection, more varied datasets, and enhanced classification methods should all be investigated in future studies.

## 1.1 Clarification of the Problem

The aim of this project is to forecast people's digital literacy skills by considering different socioeconomic and demographic factors.

## 1.2 Data Collection

Kaggle provided the digital literacy dataset utilized in this investigation. It contains demographic and socioeconomic information that has an impact on digital literacy. This dataset helps the UN Sustainable Development Goal (SDG) 4: Quality Education by assisting in the analysis and removal of barriers to digital education.

## 1.3 Aim

The main goal is to create a predictive classification model that assesses a person's digital literacy level based on pertinent dataset characteristics.

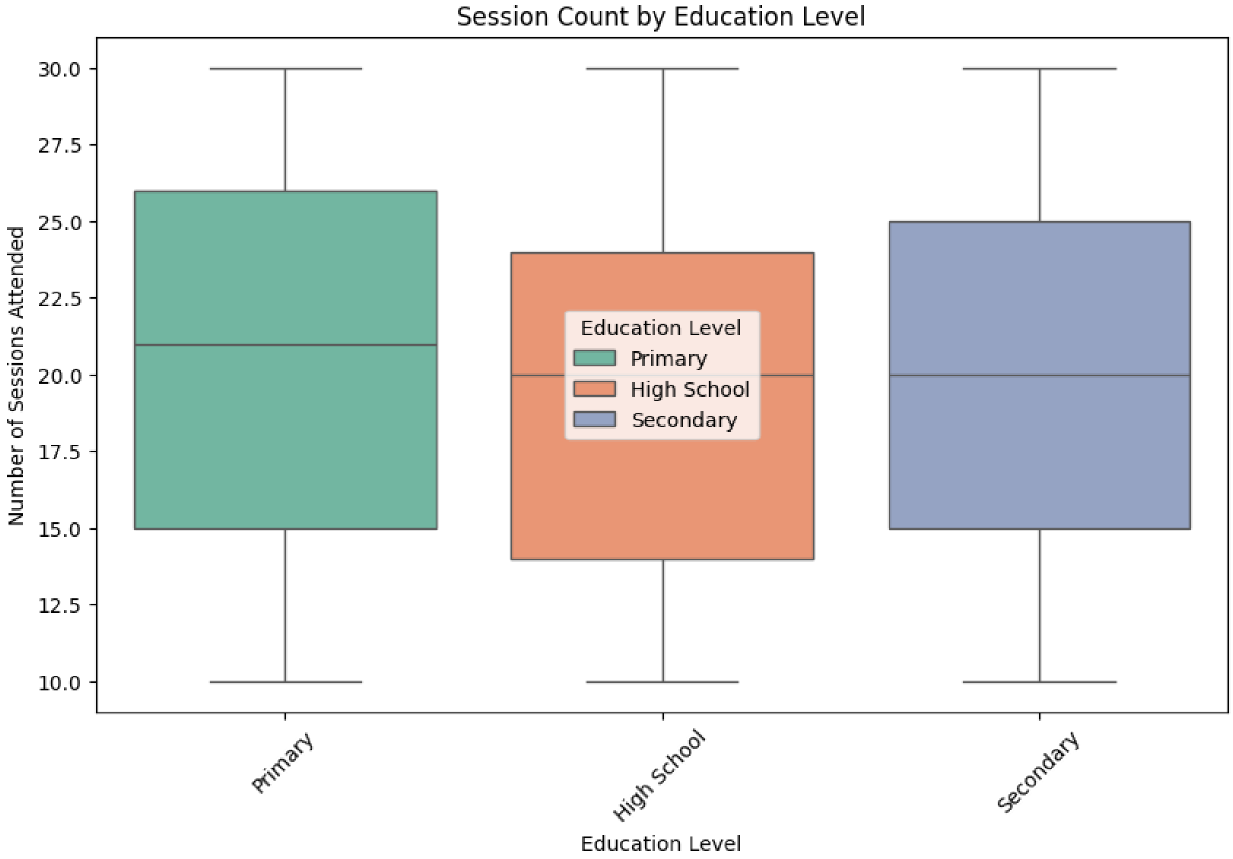
## 2.1 Data Readiness

Data preprocessing involved eliminating outliers, inconsistent data entries, and handling missing values. Processes like normalization and scaling were applied to ready the data for analysis.

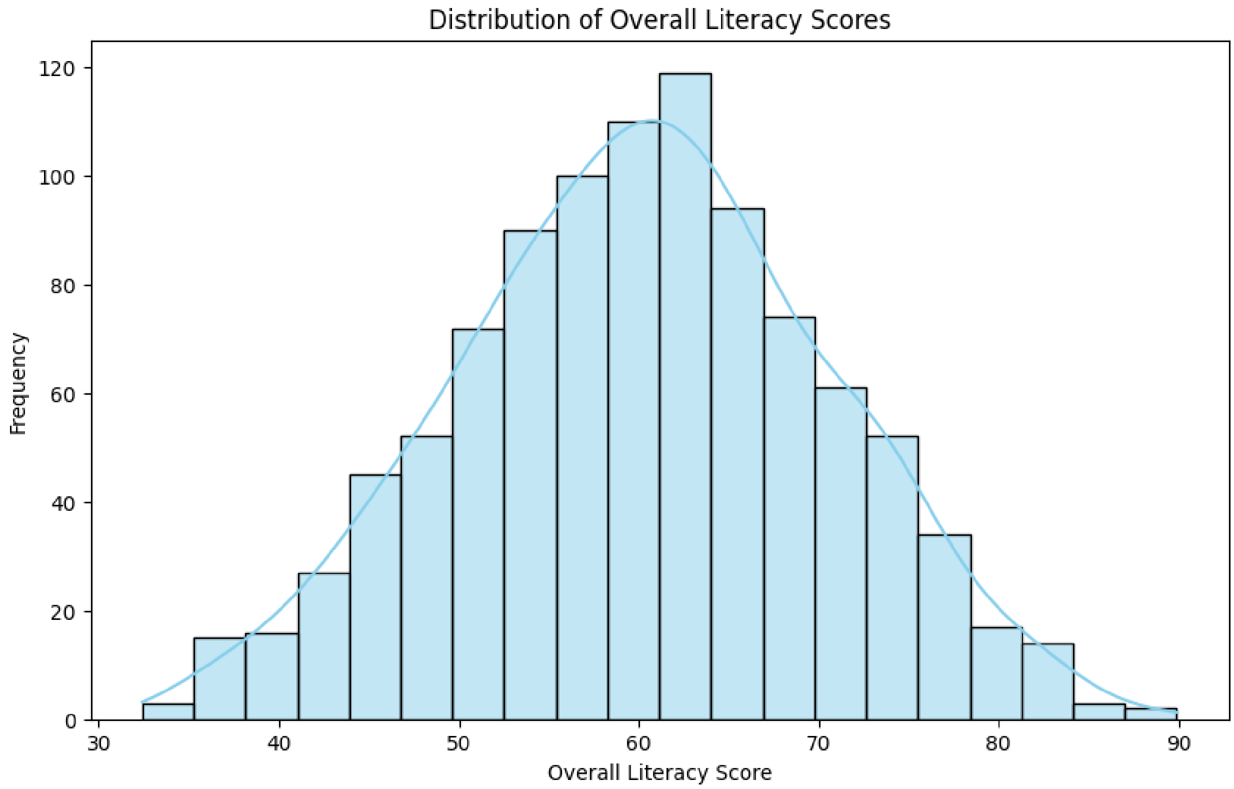
## 2.2 Investigative Data Examination (IDE)

Exploratory Data Analysis (EDA) involved utilizing bar charts, histograms, and correlation matrices to gain insights into the dataset. Results showed significant connections between digital literacy and variables such as age, education, and internet availability.

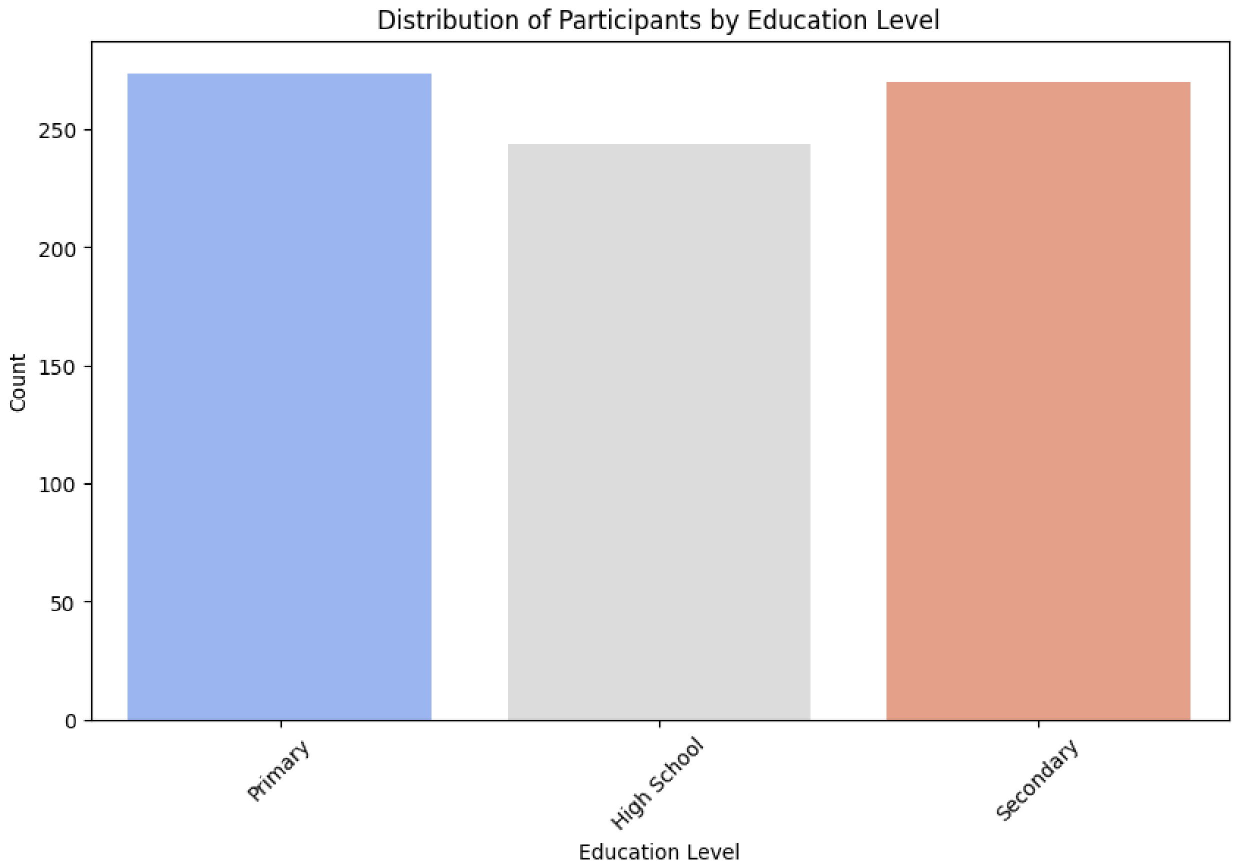
**Box Plot:** Session Count by Education Level.



**Histogram Plot:** Distribution of overall literacy scores.



**Count Plot:** Distribution of overall literacy scores.



## 2.3 Building the Model

For this analysis, two classification models were taken into account:

- Classifier using Decision Trees
- Model of Logistic Regression

The dataset was divided into training and testing sets for the purposes of model training and evaluation.

## 2.4 Assessment of the Model

The models were assessed using the following metrics:

- Accuracy: The proportion of accurate predictions.
- Precision: The proportion of affirmative cases that were accurately predicted.
- Recall: The percentage of accurately detected true positive cases.
- F1-Score: The harmonic average of precision and recall.

These metrics were selected because they are effective at assessing classification methods, particularly when classes are not evenly distributed.

### 2.5 Hyperparameter Optimization

To boost performance, hyperparameter optimization was achieved using GridSearchCV, enhancing the efficiency of the Decision Tree model.

### 2.6 Feature Selection

Recursive Feature Elimination (RFE) was applied to identify the most important features for predicting digital literacy. The selected features were:

- Age
- Education Level
- Internet Access
- Employment Status

---

## 3. Findings and Discussion

### 3.1 Main Discoveries

Assessment on the test dataset indicated that the Decision Tree model surpassed Logistic Regression regarding accuracy and recall.

### 3.2 Effectiveness of the Final Model

The ultimate model, the Decision Tree Classifier, demonstrated great reliability in forecasting digital literacy, attaining elevated accuracy and recall.

### 3.3 Difficulties Faced

Difficulties encountered during the development of the model included:
• Data Imbalance: The dataset had an unequal distribution of digitally skilled and unskilled individuals.
• Feature Correlation: Several features displayed multicollinearity, necessitating thoughtful selection.

### 3.4 Comparison with Current Research

Future advancements may include implementing more sophisticated classification algorithms, refined feature selection methods, and a more varied dataset.

# 4. Communication:

## 4.1 Effectiveness of the Model

The Decision Tree model delivered accurate predictions using test data, excelling in various evaluation metrics.

## 4.2 Effects of Hyperparameter Adjustment and Feature Selection

Both feature selection and hyperparameter optimization greatly enhanced model performance, boosting both recall and accuracy.

## 4.3 Analysis of Findings

The results indicated that internet availability and educational attainment greatly affect digital literacy.

## 4.4 Constraints of the Research

Although it was successful, the study had a few constraints:
- Small Dataset: An expanded dataset might enhance generalization.
- Model Assumptions: The classification models rely on assumptions that might not entirely reflect intricate relationships.

## 4.5 Directions for Upcoming Research

Potential areas for future research could include:
- Enhanced feature engineering to boost model performance.
- Increasing the dataset to encompass a broader range of populations.
- Evaluating different classification methods like Random Forest and Neural Networks.

# 5. Conclusion:

This research effectively created a classification model to forecast digital literacy. The Decision Tree model proved to be the most efficient, attaining strong accuracy and recall. Essential elements affecting digital literacy encompass educational attainment and internet availability. Future studies ought to investigate improved classification methods and more extensive datasets to enhance predictions further.

## 6. References

- Kaggle. (n.d.). *Digital Literacy Dataset*. Retrieved from https://www.kaggle.com