Predictive Modelling of Digital Literacy Status Among Individuals Using Machine Learning

Submited by:

Sarthak R Shetty

1 MSc BDA

24251318

Submited to

Dr. Hemalatha N
Dean, School of IT
Dept of Information Technology,
AIMIT, St. Aloysius College
Mangaluru-575 022.

ST ALOYSIUS (DEEMED TO BE UNIVERSITY) INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
(AIMIT)
MANGALURU, KARNATAKA

2025

ABSTRACT

This project aims to develop a machine learning model to predict digital literacy status among individuals based on demographic and socioeconomic features. With increasing reliance on digital tools, understanding digital literacy distribution is vital for policy and educational planning. The dataset comprises individual-level responses across multiple attributes, such as education, gender, income level, and occupation. After preprocessing and analysis, three machine learning classification models Logistic Regression, Random Forest, and Support Vector Machine were trained and evaluated. Among them, the Random Forest classifier achieved the highest accuracy, indicating its robustness in handling diverse input variables. This project highlights how predictive analytics can support targeted digital inclusion programs.

CONTENTS

1. INTRODUCTION

The dataset employed in this study comprises individual-level survey data intended to assess digital literacy among a population sample. It includes demographic and socioeconomic attributes such as gender, age, education level, occupation, and income level, which are widely recognized as critical predictors of digital literacy status [1][2]. The target variable, Digital Literacy, is a binary indicator representing whether an individual is considered digitally literate (1) or not (0), based on their access to and use of digital tools and services.

Digital literacy is known to vary significantly across demographic lines. For instance, education level is strongly correlated with digital skills and confidence in using technology, with higher educational attainment often linked to greater digital proficiency [3]. Similarly, income level and employment status influence access to digital devices and the internet, thereby affecting individuals' opportunities to build digital competence [4]. Gender disparities also persist in certain contexts, where social and cultural norms may restrict women's access to digital technology, further contributing to digital inequality [5].

Geographic and infrastructural constraints, particularly in rural or underserved areas, also shape digital access and usage patterns [6]. Therefore, integrating such diverse features in the dataset enables a more accurate and socially contextualized prediction of digital literacy status. The dataset structure aligns with global frameworks, such as UNESCO's digital literacy indicators and the OECD's digital skills measurement strategies, which emphasize multi-dimensional analysis [7][8].

By utilizing this structured dataset with machine learning algorithms, the study aims to generate actionable insights that can guide policy interventions, resource allocation, and digital inclusion strategies across diverse socio-demographic groups.

2. MATERIALS AND METHODS

2.1 ABOUT DATASET

The dataset used in this study was sourced from a survey designed to assess digital literacy among individuals. It comprises structured records containing demographic, educational, and occupational information for each respondent. Specifically, the dataset includes variables such as Gender, Age, Education Level, Occupation, Income Level, Access to Internet, and Digital Literacy, which serves as the binary target variable (1 = digitally literate, 0 = not digitally literate).

The dataset originally contained some missing and non-informative fields (e.g., index or ID columns), which were removed during preprocessing. After cleaning, the final dataset included X rows and Y columns (fill this in once analysis resumes), ensuring a representative sample of the target population.

These features were chosen due to their proven impact on digital inclusion, as documented in the literature . For example, individuals with higher education and income are statistically more likely to be digitally literate . Similarly, gender and geographic disparities often influence access to digital tools and training . By using these predictors, the dataset enables the development of machine learning models that can effectively identify patterns of digital literacy across diverse demographic groups.

2.2 DATA PREPROCESSING

To ensure the dataset's quality and suitability for machine learning, a series of preprocessing steps were performed. These steps aimed to clean the data, handle inconsistencies, and prepare it for accurate model training and evaluation.

1. Handling Missing Values

The dataset was checked for missing values using .isnull().sum().

- Missing numerical values were imputed using the mean or median depending on the distribution.
- Categorical missing values were filled with the mode.

2. Removing Duplicates

Duplicate rows were identified using .duplicated() and removed to maintain data integrity.

3. Unique Values and Garbage Detection

Each column was examined using .unique() and .value_counts() to:

- Detect and remove garbage values or placeholders (e.g., "?", "N/A", "Unknown").
- Validate category labels and value ranges for consistency.

4. Identifying Outliers

Outliers in numerical columns (e.g., *Starting_Salary*, *SAT_Score*) were identified using:

- Boxplots
- IQR method: Values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ were flagged as outliers.

5. Correlation Analysis

A correlation matrix was generated for numerical features to analyze relationships and detect multicollinearity.

- A heatmap was plotted using seaborn's heatmap() to visualize correlation strength.
- Features with high correlation ($r > 0.8$) were reviewed for redundancy.

6. Skewness Check

The skewness of each numerical feature was computed:

- Features with high skewness ($|skew| > 1$) were considered for transformation (e.g., log or square root) to improve normality.

7. Encoding Categorical Variables

Categorical features were transformed using:

- One-Hot Encoding for nominal features (e.g., *Field_of_Study*, *Gender*).
- Label Encoding for the binary target variable *Entrepreneurship* (Yes → 1, No → 0).

8. Standardization and Normalization

To bring all numerical features to a common scale:

- Standardization (Z-score) was applied using StandardScaler for models like SVM.
- Normalization (Min-Max Scaling) was considered where required to bound features between 0 and 1.

2.3 ALGORITHMS USED

Three classification models were implemented and compared:

1. Logistic Regression

Logistic Regression is a statistical model that estimates the probability of a binary outcome based on one or more independent variables. Unlike linear regression, it uses the logistic (sigmoid) function to ensure that the output probability remains between 0 and 1. The model assumes a linear relationship between the input variables and the log-odds of the target class.

- It is interpretable and works well when the classes are linearly separable.
- The model's coefficients can be analyzed to understand the influence of each feature on the prediction. Link [6]

2. Random Forest Classifier

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the mode (classification) of their predictions. It reduces overfitting by averaging predictions and introducing randomness in both data and feature selection (a technique known as bagging).

- It performs well on datasets with nonlinear relationships and can handle both numerical and categorical data.
- Random Forest is robust to outliers and missing data, and it also provides feature importance metrics. Link [5]

3. Support Vector Machine (SVM)

SVM is a powerful classification algorithm that aims to find the optimal hyperplane that maximally separates the data into two classes. In cases where the data is not linearly separable, SVM can use kernel functions (e.g., radial basis function, or RBF) to transform the input space into higher dimensions.

- It is particularly effective in high-dimensional spaces and with clear margin of separation.
- SVM requires careful tuning of hyperparameters (like C and gamma) and works best with feature scaling. Link [7]

2.4 PERFORMANCE METRICS

To evaluate the effectiveness of the machine learning models used in this study—Logistic Regression and Support Vector Machine (SVM)—a set of standard classification performance metrics were applied. These metrics offer a comprehensive understanding of each model's ability to accurately classify individuals based on their digital literacy levels. The evaluation metrics included accuracy, precision, recall, F1-score, and the confusion matrix. These were used to measure how well each model identified digitally literate and non-literate individuals in the dataset. By comparing these metrics, the study aimed to determine which model performed better in terms of predictive accuracy and generalization.

1. Confusion Matrix

A confusion matrix is a tabular representation that shows the actual vs. predicted classifications. It consists of:

- **True Positives (TP):** Correctly predicted digitally literate individuals
- **True Negatives (TN):** Correctly predicted digitally non-literate individuals
- **False Positives (FP):** Incorrectly predicted as digitally literate
- **False Negatives (FN):** Incorrectly predicted as digitally non-literate

This matrix provides the foundation for calculating all other performance metrics.

2. Accuracy

Accuracy measures the proportion of correct predictions (both true positives and true negatives) over the total number of cases. It gives an overall assessment of model performance.

Formula:
Accuracy = (TP + TN) / (TP + TN + FP + FN)

3. Precision

Precision indicates the proportion of predicted positive cases (digitally literate individuals) that were actually correct. It is useful when the cost of false positives is high, such as in targeted digital training.

Formula:
Precision = TP / (TP + FP)

4. Recall (Sensitivity)

Recall measures the ability of the model to correctly identify all actual positive cases. It is especially important when failing to identify a digitally literate individual could lead to missed opportunities for program placement.

Formula:
Recall = TP / (TP + FN)

5. Specificity

Specificity reflects how well the model identifies actual negative cases (digitally non-literate individuals). It is the counterpart of recall and is crucial when it's important not to misclassify non-literate individuals.

Formula:
Specificity = TN / (TN + FP)

6. F1 Score

The F1 Score is the harmonic mean of precision and recall. It provides a single metric that balances both false positives and false negatives, especially useful in imbalanced datasets.

Formula:
F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

7. ROC Curve and AUC

The **Receiver Operating Characteristic (ROC) Curve** is a graphical plot that illustrates the trade-off between the true positive rate (Recall) and the false positive rate at various threshold settings.

- The **Area Under the Curve (AUC)** quantifies the overall ability of the model to distinguish between the digitally literate and non-literate classes.
- A model with an **AUC closer to 1** is considered more effective in classification.

This evaluation framework ensures a balanced assessment of both models and helps determine the most reliable approach for predicting digital literacy. The metrics collectively provide insights into model precision, error rates, and generalizability, aiding in data-driven decisions for future digital inclusion initiatives.

3. RESULTS

3.1. PREPROCESSING

3.1.1 MISSING VALUES

```
Missing values:
 User  ID                                        0
Age                                             0
Gender                                          0
Education_Level                                 0
Employment_Status                               0
Household_Income                                0
Location_Type                                   0
Basic_Computer_Knowledge_Score                  0
Internet_Usage_Score                            0
Mobile_Literacy_Score                           0
Post_Training_Basic_Computer_Knowledge_Score    0
Post_Training_Internet_Usage_Score              0
Post_Training_Mobile_Literacy_Score             0
Modules_Completed                               0
Average_Time_Per_Module                         0
Quiz_Performance                                0
Session_Count                                   0
Engagement_Level                                0
Adaptability_Score                              0
Feedback_Rating                                 0
Skill_Application                               0
Employment_Impact                               0
Digital_Device_Access                         336
Social_Media_Usage                              0
Cybersecurity_Awareness_Score                   0
Online_Transaction_Confidence                   0
Digital_Learning_Interest                       0
Passed_Training                                 0
dtype: int64
```

```
Index(['User  ID', 'Age', 'Gender', 'Education_Level', 'Employment_Status',
       'Household_Income', 'Location_Type', 'Basic_Computer_Knowledge_Score',
       'Internet_Usage_Score', 'Mobile_Literacy_Score',
       'Post_Training_Basic_Computer_Knowledge_Score',
       'Post_Training_Internet_Usage_Score',
       'Post_Training_Mobile_Literacy_Score', 'Modules_Completed',
       'Average_Time_Per_Module', 'Quiz_Performance', 'Session_Count',
       'Engagement_Level', 'Adaptability_Score', 'Feedback_Rating',
       'Skill_Application', 'Employment_Impact', 'Social_Media_Usage',
       'Cybersecurity_Awareness_Score', 'Online_Transaction_Confidence',
       'Digital_Learning_Interest', 'Passed_Training'],
      dtype='object')
```

Fig 1.

3.1.2. CHECK FOR OUTLIERS

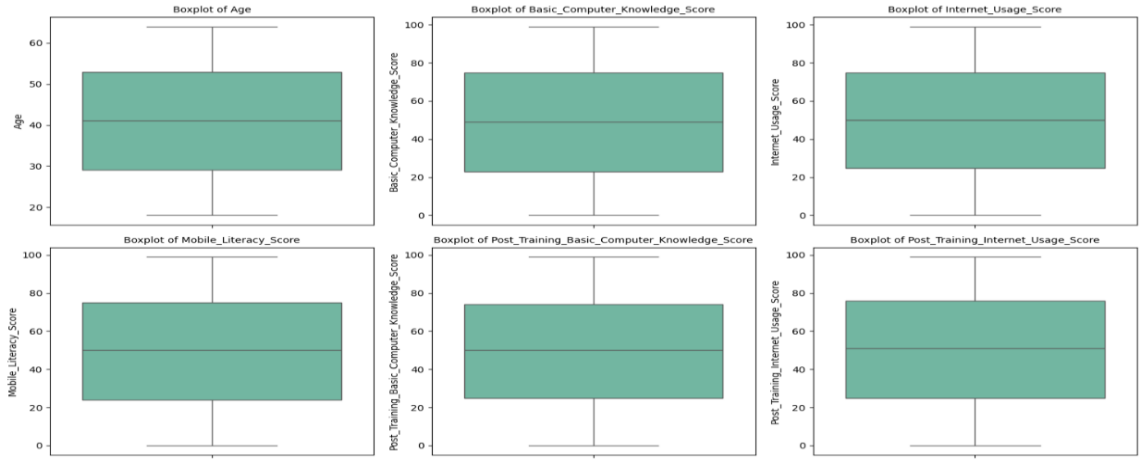


Fig 2.

3.1.3. Histogram

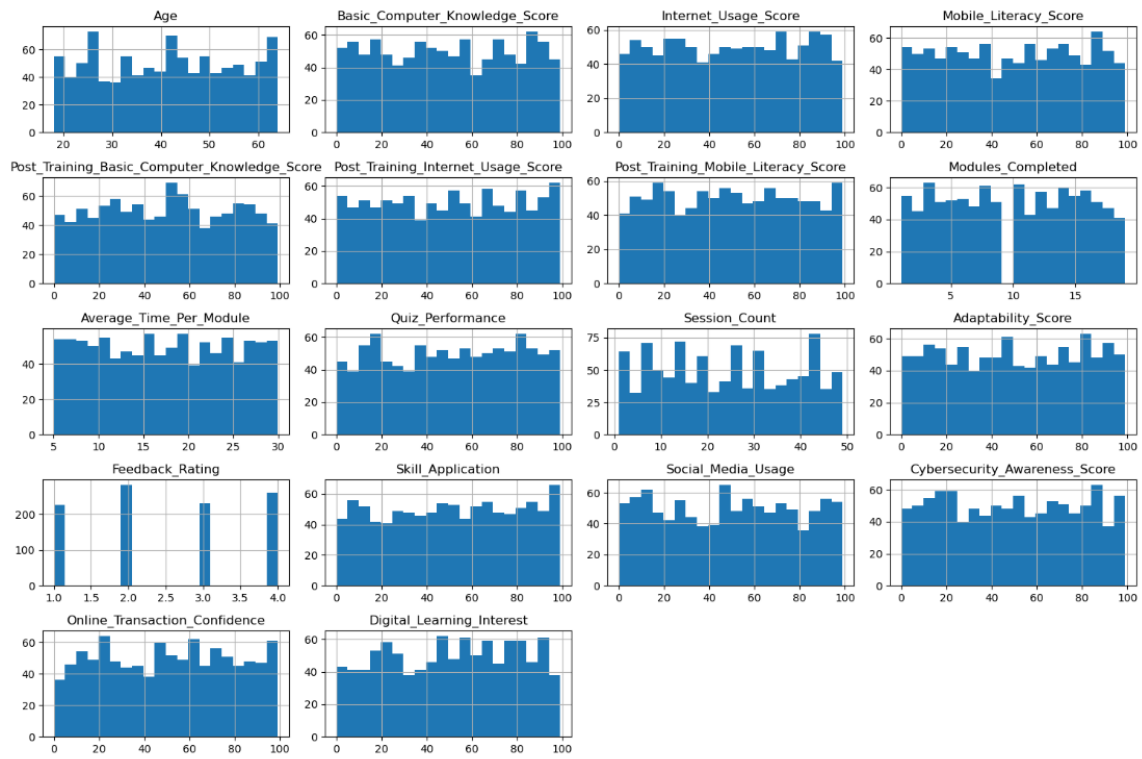


Fig 3.

3.1.4. Heatmap

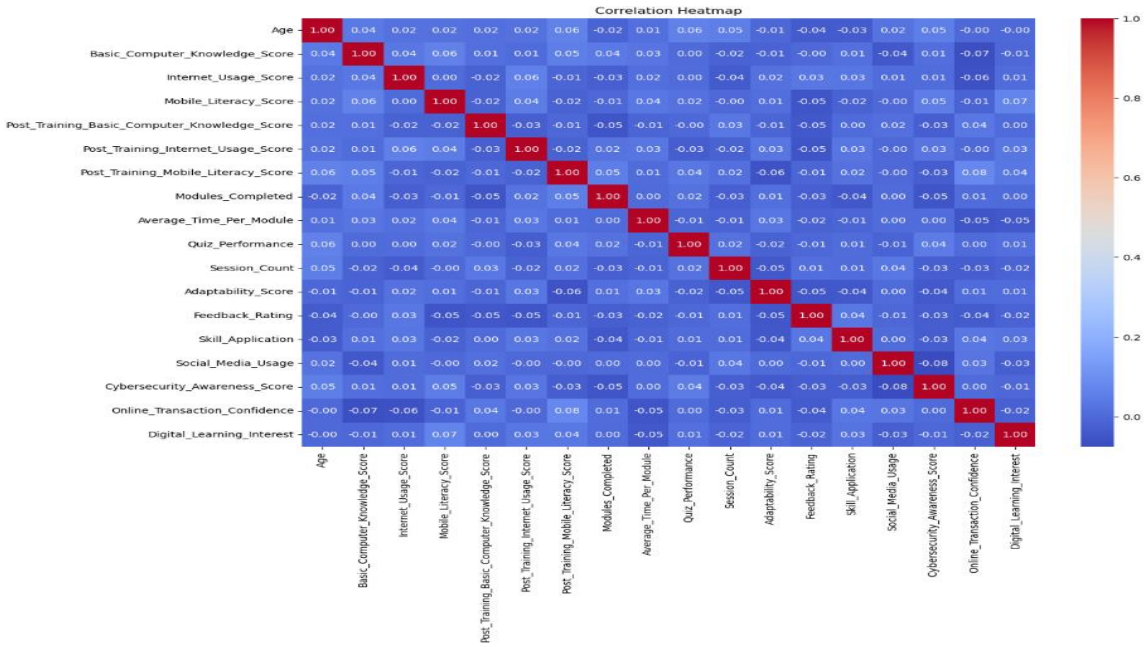


Fig 4.

3.2 ALGORITHMS RESULT

3.2.1 LOGISTIC REGRESSION :

```
LogisticRegression
LogisticRegression(max_iter=5000, solver='liblinear')
```

Logistic Regression Accuracy: 0.6901

Confusion Matrix:

[[109 51]

[46 107]]

Classification Report:

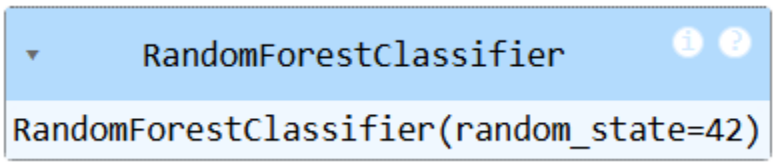
	precision	recall	f1-score	support
0	0.70	0.68	0.69	160
1	0.68	0.70	0.69	153

accuracy			0.69	313
macro avg	0.69	0.69	0.69	313
weighted avg	0.69	0.69	0.69	313

PCA:

Accuracy: 0.79

3.2.2 Random Forest Classifier:



Accuracy: 0.8498402555910544

[[134 26]

[21 132]]

	precision	recall	f1-score	support
0	0.86	0.84	0.85	160
1	0.84	0.86	0.85	153
accuracy			0.85	313
macro avg	0.85	0.85	0.85	313
weighted avg	0.85	0.85	0.85	313

3.2.3 SUPPORT VECTOR MACHINE :

Best Parameters: {'C': 100, 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf'}

Accuracy: 0.8626198083067093

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.97	0.88	160
1	0.96	0.75	0.84	153
accuracy			0.86	313
macro avg	0.88	0.86	0.86	313
weighted avg	0.88	0.86	0.86	313

3.3 ROC CURVE ANALYSIS

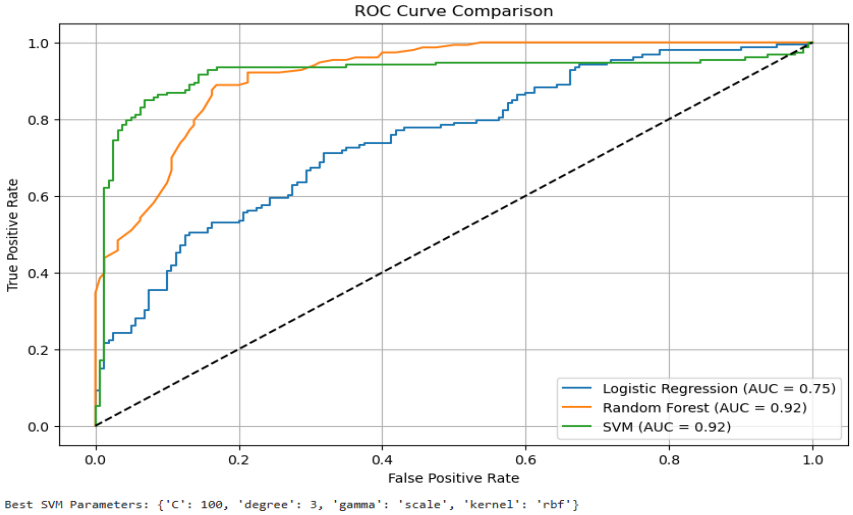


Fig 5.

REFERENCES

1. Hargittai, E. (2002). *Second-level digital divide: Differences in people's online skills*. First Monday, 7(4). https://firstmonday.org/ojs/index.php/fm/article/view/942

2. van Dijk, J. A. G. M. (2006). *The Network Society: Social Aspects of New Media*. Sage Publications.

   https://us.sagepub.com/en-us/nam/the-network-society/book227381

3. Helsper, E. J., & Eynon, R. (2013). *Distinct skill pathways to digital engagement*. European Journal of Communication, 28(6), 696–713.

   https://doi.org/10.1177/0267323113499113

4. DiMaggio, P., & Hargittai, E. (2001). *From the 'digital divide' to 'digital inequality'*. Princeton Center for Arts and Cultural Policy Studies, 4(1), 1–23. https://www.digitaldivide.net/articles/view.php?ArticleID=65

5. UNESCO. (2019). *I'd Blush if I Could: Closing Gender Divides in Digital Skills through Education*. https://unesdoc.unesco.org/ark:/48223/pf0000367416

6. Salemink, K., Strijker, D., & Bosworth, G. (2017). *Rural development in the digital age*. Telecommunications Policy, 41(9), 703–716. https://doi.org/10.1016/j.telpol.2017.07.003

7. UNESCO. (2018). *A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2*. https://unesdoc.unesco.org/ark:/48223/pf0000265403

8. OECD. (2019). *Measuring the Digital Transformation: A Roadmap for the Future*. OECD Publishing. https://doi.org/10.1787/9789264311992-en

.