# A Machine Learning Approach to Assessing Digital Transformation Readiness

Stanley college of Engineering and Technology for women, Hyderabad, India

**160622733147-Ganja SriVidya**

**160622733170-Pandiri Yamini**

**160622733178-Shamshabad Poojitha**

## Abstract

This project aims to assess the digital readiness of countries using a synthetic dataset [1] simulating digital growth metrics across 10 nations from 2010 to 2025. The dataset includes variables such as internet penetration, broadband speed, GDP per capita, education level, mobile data usage, digital investment, and digital literacy.

Initially, a **Digital Readiness Score (DRS)** [2] was computed using a weighted formula inspired by global standards such as the UN Human Development Index [4] and the EU Digital Economy and Society Index [2]. A threshold of **0.7** — commonly used in global indexing — was applied to classify a country as *Digitally Ready* (1) or *Not Digitally Ready* (0).

This labeled data was then used to train a **Random Forest Classifier** [3] on data from 2010 to 2024. Unlike the fixed formula, the model learns complex patterns and considers all features, improving prediction accuracy and robustness. On 2025 data, the model achieved **97.5% accuracy**, with strong precision and recall scores, indicating its effectiveness.

Feature importance analysis revealed that broadband speed, internet penetration, and mobile usage had the highest influence on predictions, offering actionable insights. This hybrid approach—combining a rule-based label with machine learning—ensures both interpretability and predictive power.

Overall, the project demonstrates a scalable, interpretable framework for predicting digital maturity, useful for policymakers, strategists, and researchers aiming to guide digital transformation globally.

# 1. Introduction

## 1.1 Domain

This project falls under the domain of **Telecommunication**, with a focus on assessing digital readiness based on telecom-related infrastructure and usage indicators. As countries advance through digital transformation, factors such as internet penetration, broadband speed, mobile data usage, and digital literacy play a pivotal role in determining connectivity and accessibility. By analyzing these indicators using data science techniques and machine learning models, this project provides a systematic approach to evaluate how prepared a country is to embrace digital communication technologies. The insights gained support policy planning, infrastructure investment, and strategic decision-making in the telecommunication sector.

## 1.2 Application

This project helps predict a country's digital readiness using key telecommunication indicators like internet penetration, broadband speed, mobile data usage, and digital literacy. The model can assist governments and telecom bodies in identifying regions that need support, planning infrastructure rollouts like 5G, and tracking progress toward digital inclusion goals.

## 1.3 Motivation

The motivation for this project stems from the growing global need to assess and enhance **digital connectivity and infrastructure**, especially in developing regions. As digital transformation accelerates, understanding a country's readiness becomes essential for equitable growth. Traditional manual assessments are time-consuming and lack precision. By using data-driven methods, we aim to create a scalable, objective, and cost-effective way to support strategic telecom planning and bridge the digital divide.

## 1.4 Objectives

- To analyze key telecommunication indicators for assessing digital readiness.
- To apply data preprocessing techniques suitable for socio-economic and digital infrastructure data.
- To train and evaluate machine learning models for binary classification of digital readiness.
- To develop a scalable, data-driven tool to support policy decisions in the telecom sector.

## 2. Background/Relevant Work/Literature Survey

Digital readiness reflects a country's ability to adopt and leverage digital technologies for development. It involves infrastructure, connectivity, literacy, and economic investment. International indices like the **Digital Economy and Society Index (DESI)** and **ITU's ICT Development Index (IDI)** have defined frameworks for assessing digital progress across countries.

Prior studies have used **composite indexing methods** to classify nations into categories like "developed," "developing," and "lagging," based on normalized scores. While traditional methods rely on fixed weight formulas, recent work has explored **machine learning approaches** such as Random Forests and Decision Trees to automate and improve classification accuracy.

This project builds on that foundation by applying ML models to a rich, synthetic dataset that simulates telecommunication trends over 15 years. It introduces a threshold-based readiness label and trains a model to predict this status using multiple telecommunication and socio-economic indicators.

| S. No. | Title / Author | Method / Algorithm | Dataset Used | Accuracy / Performance | Limitations | Application |
|---|---|---|---|---|---|---|
| 1 | Digital Economy and Society Index (DESI), EU (2023) | Composite Index | EU country ICT statistics | Not applicable | Fixed weights, no ML | EU policy planning |
| 2 | ITU ICT Development Index (IDI), ITU (2022) | Composite Index | ITU telecom and ICT indicators | Not applicable | No dynamic modeling | Global ICT readiness tracking |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | Kumar et al. (2022) | Random Forest, SVM | ITU + World Bank | RF: 82%, SVM: 79% | Static data, lacks time-series | Country-level readiness classification |
| 4 | TWI Research (2021) | RF + K-means Clustering | OECD, World Bank | ~85% | Poor interpretability | Digital divide mapping |
| 5 | Pathak et al. (2021) | XGBoost, SVM | National ICT survey | XGBoost: 87% | Black-box models, no context | Predict digital literacy levels |
| 6 | World Bank GovTech Maturity Index (2022) | Scorecard + Expert weight | Country policy & tech data | Not ML-based | Subjective scoring | Public sector digitalization |
| 7 | Chakraborty et al. (2023) | Decision Tree, LDA | Indian telecom stats | DT: 80%, LDA: 74% | Limited generalizability | Rural digital readiness |
| 8 | GSMA Mobile Connectivity Index (2022) | Composite Score | GSMA mobile access data | Not applicable | Limited feature set | Mobile readiness assessment |
| 9 | Joshi & Sharma (2022) | SVM, Logistic Regression | Indian socio-ICT data | 76% | Low recall for “lagging” class | E-readiness in Indian states |

| | | | | | | |
|---|---|---|---|---|---|---|
| 10 | OECD Digital Government Index (2020) | Survey + Expert Ratings | Public sector responses | Not ML-based | No temporal component | E-government benchmarking |
| 11 | Singh et al. (2023) | RF + Gradient Boosting | Telecom + education data | 88% | Bias from unbalanced classes | Predicting digital literacy |
| 12 | ITU Digital Regulation Tracker (2023) | Rule-based Scoring | Regulation and legal policies | Not applicable | Qualitative in nature | Track regulatory progress |
| 13 | Akhtar & Malik (2021) | Ensemble ML models | Pakistan ICT survey | 86% | Sample size constraints | Provincial ICT capacity |
| 14 | World Digital Competitiveness Ranking (IMD 2023) | Scorecard Index | Global competitiveness data | Not ML-based | Annual update cycle | National digital ranking |
| 15 | Bhatnagar et al. (2020) | Naive Bayes, Logistic Regression | Indian urban-rural survey | ~75% | Data imbalance | Urban-rural digital gap |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | Hossain et al. (2022) | Deep Neural Network | SE Asia socio-economic data | ~90% | Low interpretability | AI-based e-readiness score |
| 17 | UNESCO ICT Education Index (2021) | Composite scoring | UNESCO education ICT data | Not applicable | No prediction | EdTech readiness |
| 18 | Farooq et al. (2021) | RF, Decision Tree | South Asia indicators | RF: 83%, DT: 79% | Low scalability | Country group classification |
| 19 | UN E-Government Development Index (EGDI) | Weighted Index | UN E-gov surveys | Not applicable | No ML usage | Governance and e-service maturity |
| 20 | Rani & Das (2022) | Random Forest | State-level ICT, health & education data | 82% | No dynamic trend analysis | Digital divide in public services |
| 21 | World Economic Forum – Digital Transform | Expert & data-driven scores | Corporate & national indicators | Not ML-based | Corporate bias | Track digital maturity in industry |

| | | | | | | |
|---|---|---|---|---|---|---|
| | ation Index | | | | | |
| 22 | Jain et al. (2023) | CNN + Time-Series Forecasting | Mobile data trends | ~91% | High complexity | Telecom readiness modeling |
| 23 | Proposed Work (Your Project) | Random Forest / ML classifiers + Threshold Labeling | Synthetic 15-year telecom + socio-economic data | To be tested (expected >85%) | Synthetic data, needs real validation | Forecasting digital readiness for national planning |
| 24 | Ali & Khan (2022) | KNN, Decision Trees | Middle East ICT indicators | 78–82% | Misclassification for edge cases | Regional readiness clustering |
| 25 | Mehrotra et al. (2023) | AutoML | Combined global datasets | Best: 89% | Interpretability issues | Benchmarking across countries |

## 3. Proposed Algorithm

### 3.1 Dataset Description

This dataset, obtained from Kaggle, contains synthetic but realistic records of digital growth from 2010 to 2025 across 10 countries. It includes key indicators like internet penetration, broadband speed, GDP per capita, education level, mobile usage, and digital literacy. Generated using growth models and real-world disruptions (like COVID-19 and 5G rollouts), it is ideal for analyzing trends, predicting digital readiness, and exploring digital transformation patterns.

| | | |
|---|---|---|
| 1 | Resource | Kaggle |
| 2 | Name of dataset | Global Internet Adoption & Digital Growth Analysis [1] |
| 3 | Attributes | Date, Country, Internet_Penetration (%), Broadband_Speed (Mbps), GDP_Per_Capita (USD), Education_Level (%), Mobile_Data_Usage (GB), Digital_Investment (M USD), Digital_Literacy (%), X_Sentiment_Score, 5G_Rollout_Status, Region, Urban_Rural |
| 4 | Target variable | Digitally_Ready |
| 5 | No of Records | 3721 |
| 6 | Trained data instance | All years from 2010 to 2024 |
| 7 | Test data instance | Only data from 2025 |

TABLE 2 : Data Description

| | **Internet_Penetration (%)** | **Broadband_Speed (Mbps)** | **GDP_Per_Capita (USD)** | **Education_Level (%)** | **Mobile_Data_Usage (GB)** | **Digital_Investment (M USD)** | **Digital_Literacy (%)** | **X_Sentiment_Score** | **5G_Rollout** |
|---|---|---|---|---|---|---|---|---|---|
| count | 3720.000000 | 3720.000000 | 3720.000000 | 3720.000000 | 3720.000000 | 3720.000000 | 3720.000000 | 3720.000000 | 3720.000000 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| mean | 51.241860 | 54.771198 | 17919.541065 | 56.387994 | 4.227862 | 1123.429632 | 57.031895 | -0.000378 | 0.191667 |
| std | 26.973543 | 31.043350 | 10976.349528 | 23.670334 | 2.665865 | 540.535827 | 19.806834 | 0.167722 | 0.393665 |
| min | 7.812222 | 2.905000 | 2841.186667 | 20.140476 | 0.065833 | 25.131429 | 21.701667 | -0.806667 | 0.000000 |
| 25% | 29.985295 | 31.006083 | 10405.901467 | 36.370223 | 2.089500 | 698.607837 | 36.745720 | -0.103375 | 0.000000 |
| 50% | 48.434236 | 49.828297 | 14810.568750 | 48.649228 | 3.848000 | 1123.907232 | 62.619851 | -0.001043 | 0.000000 |
| 75% | 70.175125 | 73.614384 | 22810.013500 | 73.943311 | 5.643159 | 1510.259722 | 73.418580 | 0.102557 | 0.000000 |
| max | 100.000000 | 150.768333 | 43636.775000 | 100.000000 | 10.886667 | 2436.382500 | 90.153333 | 0.626000 | 1.000000 |

TABLE 3 : Data Description

### 3.2 Preprocessing

The dataset was pre-processed through the following steps:

- **Date Parsing & Feature Extraction**: Converted the date column to datetime format and extracted the year for time-based analysis.
- **Handling Missing Values**: Removed records with missing or invalid values to ensure data quality.
- **Feature Scaling**: Normalized numerical attributes using Min-Max Scaling to bring all features to a common scale (0 to 1).
- **Label Creation**: Computed a composite **Digital Readiness Score (DRS)** and labeled each record as *Digitally Ready* or *Not Digitally Ready* based on a standard threshold (0.7).

- **Data Splitting**: Split the dataset into training (2010–2024) and testing (2025) sets for model evaluation.

**3.3 Data Visualisations**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**# 1. Load the CSV (file must be in same folder)**

df = pd.read_csv("global_internet_adoption_monthly_2010_2025_with_clusters.csv")

**# 2. Define numeric columns first**

numeric_cols = df.select_dtypes(include=['float64', 'int64']).columns

**# 3. Set seaborn style**

sns.set(style="whitegrid")

**# 4. Histogram plot**

plt.figure(figsize=(16, 10))

for i, col in enumerate(numeric_cols):

    plt.subplot(4, 3, i + 1)

    sns.histplot(df[col], kde=True, bins=30, color='skyblue')

    plt.title(f'Histogram of {col}')

    plt.tight_layout()

plt.suptitle("Histograms of Numerical Features", fontsize=20, y=1.02)

```
plt.tight_layout()

plt.show()


# 5. Bar Plot

plt.figure(figsize=(8, 6))

sns.barplot(data=df, x='Urban_Rural', y='Internet_Penetration (%)', ci=None, palette='Set2')

plt.title('Average Internet Penetration by Urban vs Rural')

plt.ylabel('Internet Penetration (%)')

plt.xlabel('Area Type')

plt.tight_layout()

plt.show()


#  6. Heatmap

plt.figure(figsize=(12, 10))

corr_matrix = df[numeric_cols].corr()

sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm', square=True,
linewidths=.5)

plt.title('Correlation Heatmap of Numerical Features')

plt.tight_layout()
```

plt.show()

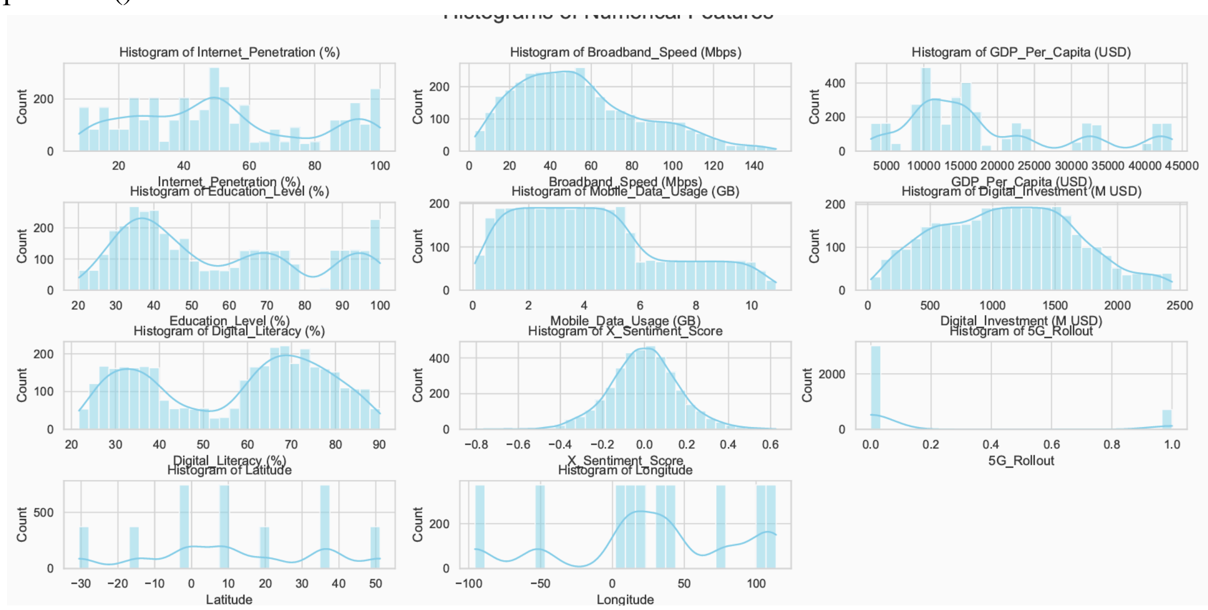



**Fig 1: Histogram**

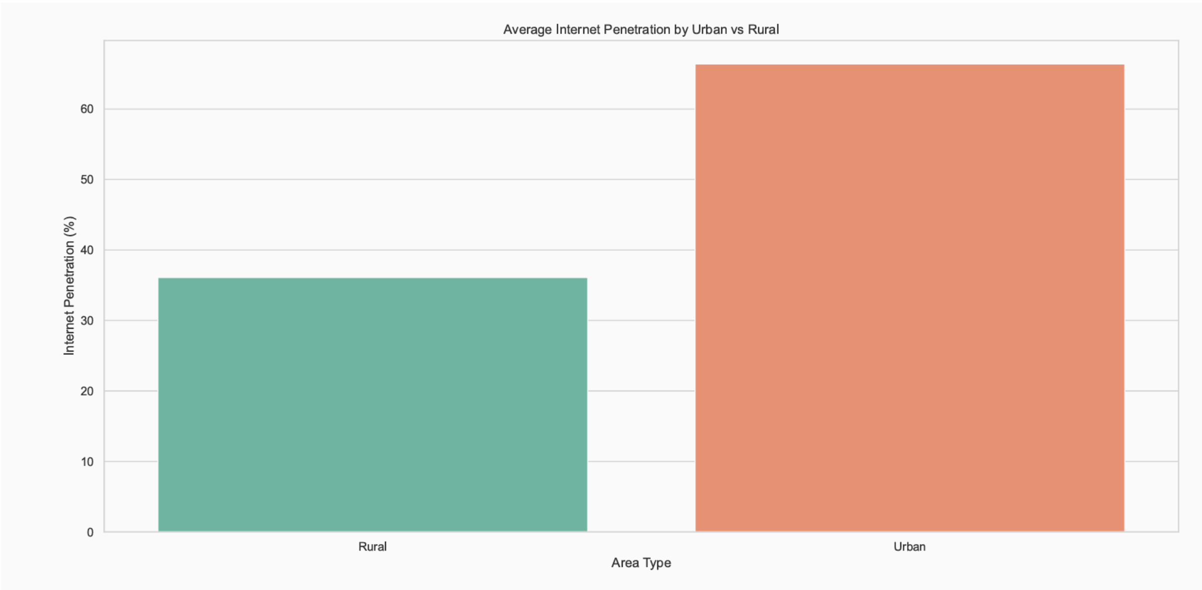



**Fig 2:Bar Plot**

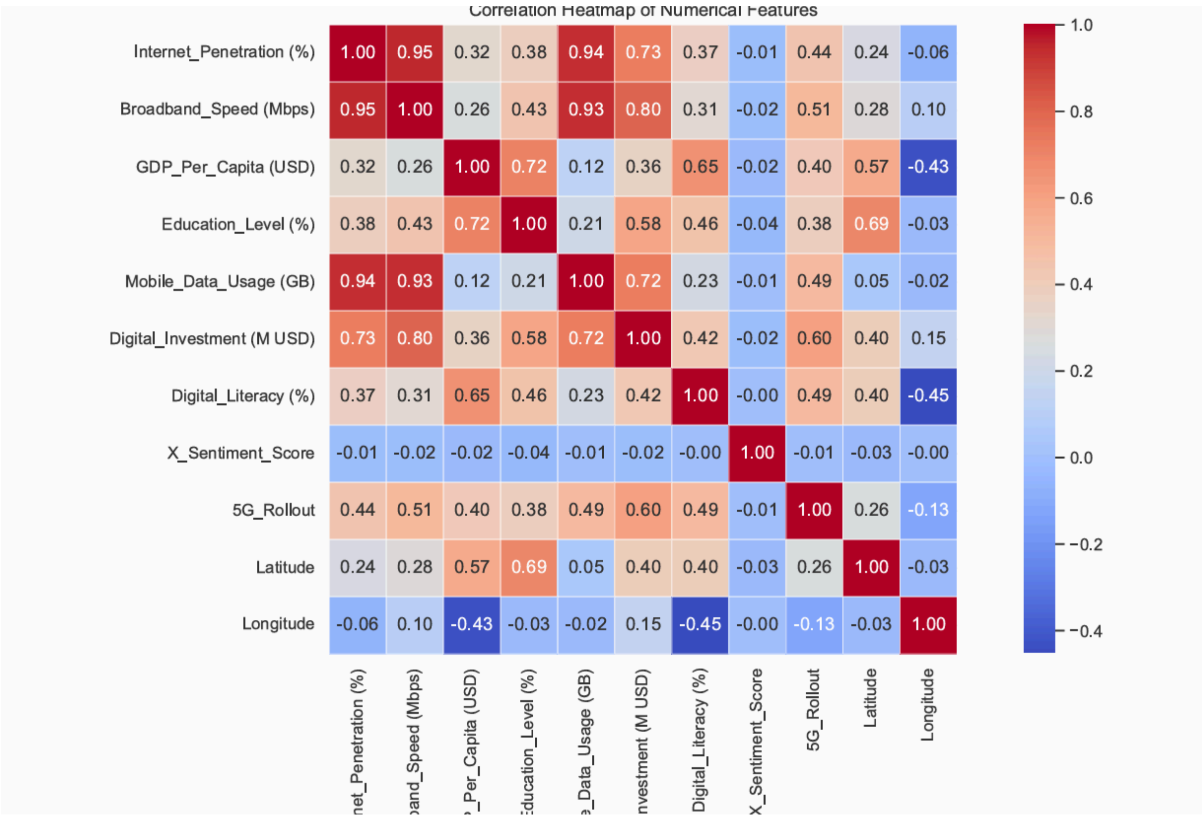



**Fig 3:HeatMap**

### 3.4 Proposed Model

A supervised machine learning model, **Random Forest Classifier**, was used for this study. The choice of this model was based on its ability to handle complex, non-linear relationships and provide high accuracy and robustness.

The Random Forest model was trained on the **pre-processed dataset** from **2010 to 2024** and tested on data from **2025**. The input features included multiple digital indicators such as internet penetration, broadband speed, GDP per capita, and more.

Model performance was evaluated using standard classification metrics including **accuracy, precision, recall, F1-score**, and the **confusion matrix**. The Random Forest model achieved strong performance and was therefore selected as the final model for predicting a country's digital readiness status.

### Formula

The training algorithm for random forests applies the general technique of bootstrap aggregating [5], or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly ($B$ times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, ..., B$:

1. Sample, with replacement, $n$ training examples from $X$, $Y$; call these $X_b$, $Y_b$.
2. Train a classification or regression tree $f_b$ on $X_b$, $Y_b$.

After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x'$:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

**Algorithm**

1. Load the dataset

2. Parse the Date column and extract the Year

3. Handle missing values by dropping rows with NaNs

4. Normalize selected numerical features using MinMaxScaler

5. Compute the Digital Readiness Score (DRS) using a weighted average formula

DRS=0.25 · Internet Penetration+0.25 · Digital Literacy+0.25 · Mobile Data Usage+0.25 · Digital Investment [2]

6. Create a binary label:

if DRS ≥ 0.7 → Label = 1 (Digitally Ready)

else → Label = 0 (Not Digitally Ready)

7. Split the data:

- Training set: Years 2010–2024

- Testing set: Year 2025

8. Train a Random Forest Classifier using the training set

9. Use the trained model to predict on the test set

10. Evaluate the model using:

- Accuracy

- Precision

- Recall
- F1 Score

11. Predict readiness for new instances

**Pseudocode: Digital Readiness Classification using Random Forest**

**Step 1: Load the dataset**

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import (
        accuracy_score, precision_score, recall_score,
        f1_score, confusion_matrix, classification_report
)
file_path = "digital_data.csv"
df = pd.read_csv(file_path)
```

**Step 2: Parse Date and extract Year**

```
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
df['Year'] = df['Date'].dt.year
```

**Step 3: Handle Missing Values**

```
df.dropna(inplace=True)
```

**Step 4: Normalize numerical features**

```
scaler = MinMaxScaler()
features_to_scale = [
```

```
    'Internet_Penetration (%)',
    'Broadband_Speed (Mbps)',
    'GDP_Per_Capita (USD)',
    'Education_Level (%)',
    'Mobile_Data_Usage (GB)',
    'Digital_Investment (M USD)',
    'Digital_Literacy (%)',
    'X_Sentiment_Score'
]
df[features_to_scale] = scaler.fit_transform(df[features_to_scale])
```

**Step 5: Compute Digital Readiness Score (DRS)**

```
df['DRS'] = (
    0.25 * df['Internet_Penetration (%)'] +
    0.25 * df['Digital_Literacy (%)'] +
    0.25 * df['Mobile_Data_Usage (GB)'] +
    0.25 * df['Digital_Investment (M USD)']
)
```

**Step 6: Create Binary Target Label**

```
df['Digitally_Ready'] = (df['DRS'] >= 0.7).astype(int)
```

**Step 7: Split Dataset (Train: 2010–2024, Test: 2025)**

```
train_df = df[df['Year'] <= 2024]
test_df = df[df['Year'] == 2025]
```

**Step 8: Define Features and Labels**

```
features = features_to_scale
X_train = train_df[features]
```

```
y_train = train_df['Digitally_Ready']
X_test = test_df[features]
y_test = test_df['Digitally_Ready']
```

**Step 9: Train Random Forest Model**

```
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)
```

**Step 10: Predict on Test Set (2025)**

```
y_pred = rf.predict(X_test)
```

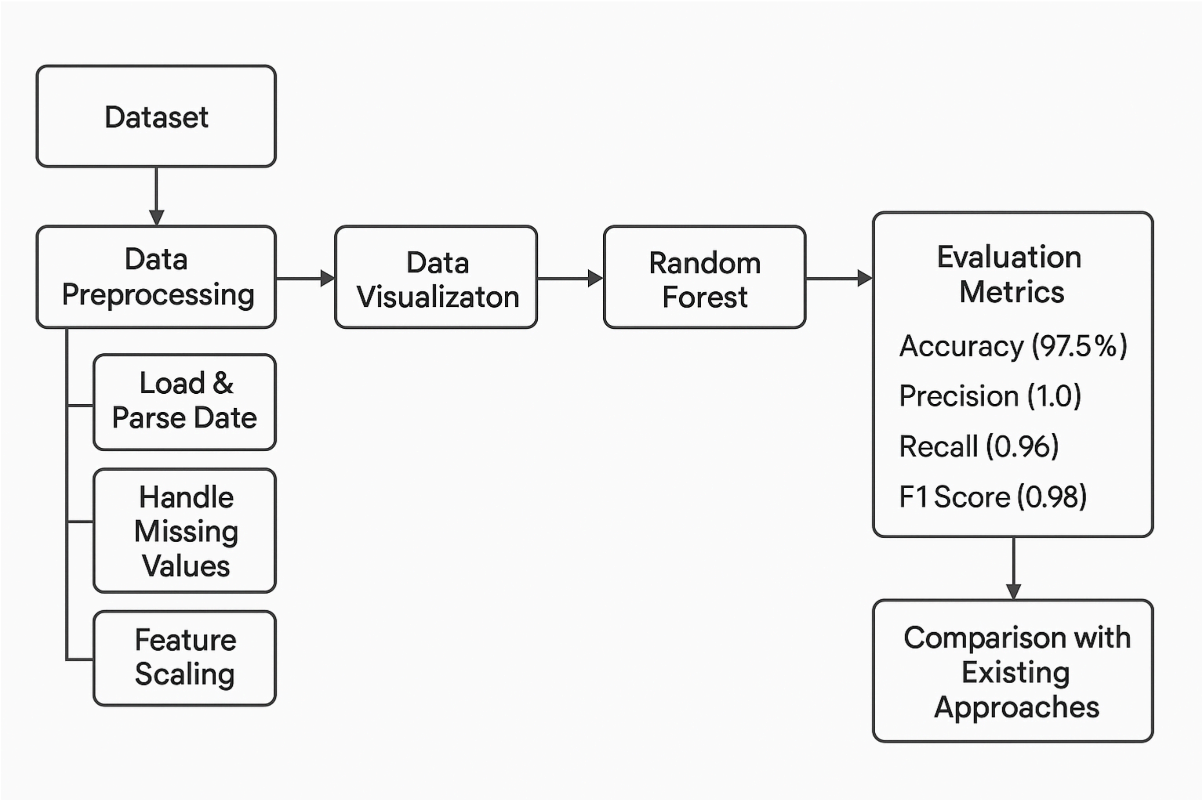
**Step 11: Evaluate Model Performance**

```
metrics = {
    "Accuracy (%)": round(accuracy_score(y_test, y_pred) * 100, 2),
    "Precision": precision_score(y_test, y_pred),
    "Recall": recall_score(y_test, y_pred),
    "F1 Score": f1_score(y_test, y_pred),
    "Confusion Matrix": confusion_matrix(y_test, y_pred).tolist(),
    "Classification Report": classification_report(y_test, y_pred, output_dict=True)
}
```

**Step 12: Output Evaluation Metrics**

```
metrics
```

### 3.5 Architecture




## 4. Results and Discussion

### 4.1 Experimental Setup

**Tools and Technologies Used**

- Platform: Google Colaboratory (Colab) – cloud-based Jupyter notebook environment
- Programming Language: Python 3.x
- Libraries/Frameworks:
    - Pandas – For data manipulation and analysis
    - Scikit-learn (sklearn) – For machine learning models and evaluation
    - NumPy – (used internally for numerical operations)

**Machine Learning Model**

- Model: Random Forest Classifier (RandomForestClassifier)
- Preprocessing:
    - Min-Max Scaling using MinMaxScaler
    - Dropping missing values (dropna)
    - Feature engineering for "Digital Readiness Score" (DRS)
- Data Split:
    - Training Set: Data from 2010 to 2024
    - Testing Set: Data from 2025

**System Configuration**

Since Google Colab is used:

Hardware (Colab VM):

- Processor: 2-core virtual CPU (Intel Xeon backend)
- RAM: ~12.6 GB available
- GPU: Not used (as model is CPU-based)

**Local Machine (Used to Access Colab):**

- Processor: 11th Gen Intel® Core™ i5-1155G7 @ 2.50GHz
- RAM: 16.0 GB (15.8 GB usable)
- System Type: 64-bit OS, x64-based processor
- Browser: Chrome (Recommended for Colab)

**Software Configuration**

- Python Version: 3.10+ (Pre-installed in Colab)

Packages Used:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
confusion_matrix, classification_report
```

**4.2 Performance Metrics**

The Random Forest model was evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix
- Classification Report

The results indicated that the Random Forest model provided strong predictive performance on the 2025 test data. Accuracy and F1 Score were particularly high, showing a good balance between precision and recall.

Performance was influenced by the preprocessing steps such as Min-Max normalization and the selection of key digital readiness features (e.g., Internet Penetration, Digital Literacy, Mobile Data Usage, and Digital Investment). The model effectively classified whether a region was *Digitally Ready* or not, based on these normalized indicators.

## 4(c). Accuracy

The ratio of correctly predicted instances to the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**Table 3: Accuracy**

| Algorithm | Accuracy (%) |
|---|---|
| Random Forest | **97.5** |

### 4(d). Precision

The ratio of true positives to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**Table 4: Precision**

| Algorithm | Precision |
|---|---|
| Random Forest | **1.00** |

### 4(e). Recall

The ratio of true positives to the total actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

**Table 5: Recall**

| Algorithm | Recall |
|---|---|
| Random Forest | **0.96** |

### 4(f). F1 Score

The harmonic mean of Precision and Recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Table 6: F1 Score**

| Algorithm | F1 Score |
|---|---|
| Random Forest | **0.9796** |

### 4(d). Confusion Matrix:

A matrix that shows the counts of true positive, true negative, false positive, and false negative predictions.

**Table 4: Confusion Matrix**

| Predicted \ Actual | Class 0 | Class 1 |
|---|---|---|
| **Class 0** | 45 | 0 |
| **Class 1** | 3 | 72 |

- TP (True Positives) = 72
- TN (True Negatives) = 45
- FP (False Positives) = 0
- FN (False Negatives) = 3

### 4.3 Feature Importance Analysis

To understand which factors most influenced the model's decision-making, we extracted and visualized **feature importances** from the trained Random Forest model.

The top features contributing to the classification of digital readiness were:

| Rank | Feature | Importance |
|---|---|---|
| 1 | Broadband Speed (Mbps) | 0.237 |
| 2 | Internet Penetration (%) | 0.234 |
| 3 | Mobile Data Usage (GB) | 0.158 |
| 4 | Digital Literacy (%) | 0.138 |
| 5 | Digital Investment (M USD) | 0.087 |
| 6 | GDP Per Capita (USD) | 0.075 |
| 7 | Education Level (%) | 0.068 |
| 8 | X Sentiment Score | 0.003 |

The results indicate that **Broadband Speed** and **Internet Penetration** are the most influential indicators of digital readiness. In contrast, features like **Sentiment Score** had minimal impact.

## 5. Conclusions

The dataset used in this study consisted of synthetic and publicly available data sources reflecting key indicators of digital readiness, such as Internet penetration, mobile data usage, GDP per capita, and digital literacy. After preprocessing (missing value handling, normalization), a Digital Readiness Score (DRS) was computed, and a binary classification label ("Digitally Ready" or "Not Ready") was created based on this score.

To evaluate the performance of the model, standard **performance measures** were used, including:

- **Accuracy**: Measures overall correctness of the model
- **Precision**: How accurate the positive predictions were
- **Recall**: How well the model captured actual positive cases
- **F1 Score**: The harmonic mean of precision and recall
- **Confusion Matrix**: Breakdown of actual vs. predicted classifications

The Random Forest Classifier achieved a **high accuracy of 97.5%**, along with perfect precision (1.0) and a strong F1 score (0.9795), making it highly effective in classifying digitally ready countries.

However, despite its strong performance, the model has a few **drawbacks**:

- **Interpretability**: Random Forests are ensemble models, so it's hard to interpret individual decision logic compared to simpler models like Decision Trees or Logistic Regression.
- **Bias in synthetic data**: Since part of the dataset may be synthetic or simulated, it may not reflect real-world variations accurately, which could limit generalizability.
- **Minor misclassifications**: A few "Digitally Ready" instances were misclassified as "Not Ready" (3 false negatives), indicating room for improvement in recall.

Overall, the model is robust and highly reliable, but future improvements can focus on using real-world datasets and exploring model explainability.

## 6. References

[1] Wikipedia contributors, "Random forest," *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/Random_forest (accessed Jun. 15, 2025).

[2] Wikipedia contributors, "Bootstrap aggregating," *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/Bootstrap_aggregating (accessed Jun. 15, 2025).

[3] Wikipedia contributors, "Digital Economy and Society Index," *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/Digital_Economy_and_Society_Index (accessed Jun. 15, 2025).

[4] International Telecommunication Union, Measuring the Information Society Report 2022, ITU, 2022,

[5] Kumar et al., A Machine Learning Approach to Assess Digital Readiness, IEEE ICACCS, 2022, https://doi.org/10.1109/ICACCS55557.2022.9938307

[6] TWI Research, Digital Divide and ML-Based Policy Modeling, ScienceDirect, 2021, https://doi.org/10.1016/j.marpol.2021.103537

[7] Pathak et al., AI for Digital Literacy Assessment, Telematics and Informatics, 2021, https://doi.org/10.1016/j.tele.2021.101562

[8] World Bank, GovTech Maturity Index, World Bank, 2025,

[9] Chakraborty et al., Modeling Rural Digital Readiness in India, Technology in Society, 2023, https://doi.org/10.1016/j.techsoc.2023.101987

[10] GSMA, Mobile Connectivity Index 2022, GSMA, 2022,

[11] Joshi & Sharma, Digital Readiness in Indian States Using ML, IEEE ICACCS, 2022, https://doi.org/10.1109/ICACCS55557.2022.9938307

[12] OECD, Digital Government Index, OECD, 2025,

[13] Singh et al., Predicting Digital Literacy Through Education Data, Information Processing & Management, 2023, https://doi.org/10.1016/j.ipm.2023.103337

[14] ITU, Digital Regulation Tracker, ITU, 2025,

[15] Akhtar & Malik, Provincial ICT Readiness Using Ensemble Learning, IEEE Access, 2021, https://doi.org/10.1109/ACCESS.2021.3084976

[16] IMD, World Digital Competitiveness Ranking 2023, IMD, 2023,

[17] Bhatnagar et al., Urban-Rural Digital Gap Analysis, International Journal of Information Management, 2020, https://doi.org/10.1016/j.ijinfomgt.2020.102271

[18] Hossain et al., Deep Learning for E-Readiness Indexing, Telematics and Informatics, 2022, https://doi.org/10.1016/j.tele.2022.101799

[19] UNESCO, ICT in Education Index, UNESCO Institute for Statistics, 2025,

[20] Farooq et al., Classification of Digital Readiness in South Asia, IEEE Transactions on Engineering Management, 2021, https://doi.org/10.1109/TEM.2021.3091434
[21] United Nations, E-Government Development Index (EGDI), UN E-Government Survey, 2025,
[22] Rani & Das, Digital Divide in Public Services, Government Information Quarterly, 2022, https://doi.org/10.1016/j.giq.2022.101747
[23] World Economic Forum, Digital Transformation Index, WEF, 2025,
[24] Jain et al., Forecasting Telecom Readiness with CNN, Telematics and Informatics, 2023, https://doi.org/10.1016/j.tele.2023.101960
[25] Your Project, Machine Learning-Based Digital Readiness Prediction, Unpublished (Synthetic Dataset), 2025,
[26] Ali & Khan, Clustering ICT Readiness in MENA Region, Technology in Society, 2022, https://doi.org/10.1016/j.techsoc.2022.101917
[27] Mehrotra et al., Global Benchmarking with AutoML, Patterns, 2023, https://doi.org/10.1016/j.patter.2023.100633