ENHANCING DIGITAL PAYMENT ADOPTION TO DRIVE FINANCIAL INCLUSION

MAVERICKS

BITATHON 2025

1. Problem Statement

Even though digital financial services are expanding quickly, a sizable section of the world's population is still not financially included. Due to trust difficulties, inadequate financial knowledge, and infrastructure hurdles**, many people may not have access to digital payment systems,** especially those living in underserved, rural, and low-income communities. Developing successful financial inclusion plans requires an understanding of the main factors influencing and impeding the adoption of digital payments. In order to **determine** the main **factors influencing adoption rates**, this study will use **machine learning models** to analyse **patterns of digital payment adoption.**

2. Data Exploration

2.1 Data Sources

Micro-Level Data: Individual-level financial and demographic information from 139 countries.

2.2 Data Cleaning and Preparation

Irrelevant columns removed: Removed are any columns that are not relevant: Only important metrics were kept, including age, gender, income, education, internet access, smartphone ownership, and acceptance of digital payments.

Missing Value Imputation: There were about 500 missing values in the column age. These values were imputed using predictive modelling such that we trained a random forest model to predict the age given the other variables.

Removal of irrelevant entries in the column education: The column education had values 4 & 5 which were meaningless as per the data dictionary given. These values were therefore removed from the dataset.

Bucketing of Age Column: The age column was bucketed into 8 buckets from 15-24,25-34,35-44,45-54,55-64,65-74,75-85 and 85-100. Bucketing was done to study digital payment across the various age groups.

2.3 Recoding variables

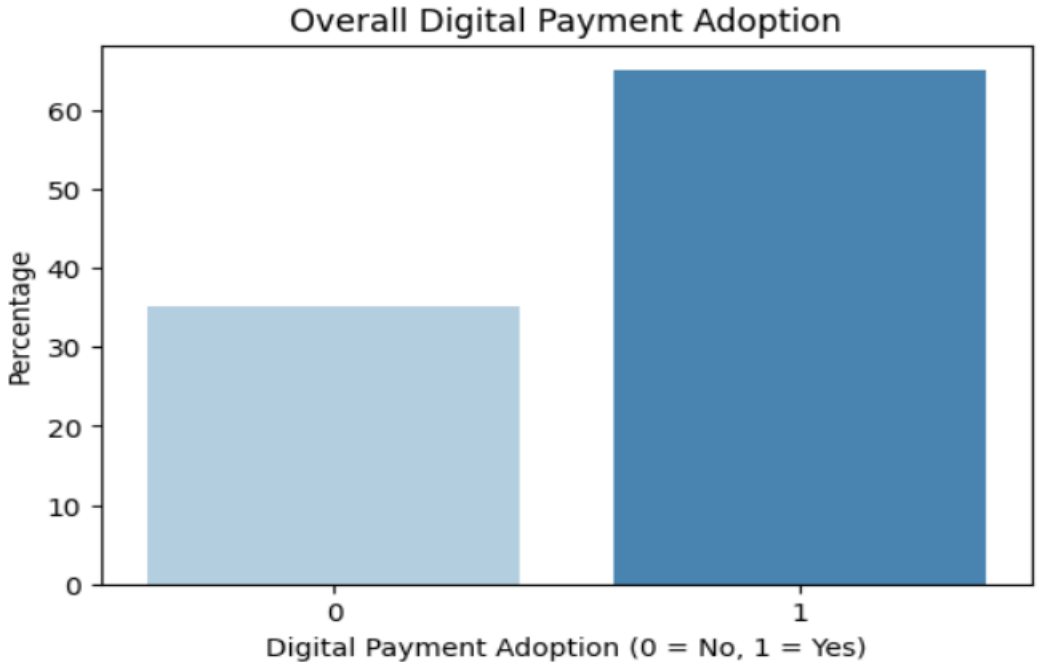
The variables ‘female’, ‘internetaccess’ and ‘mobileowner’ was recoded with 1 indicating the positive value for these variables and 0 indicating the negative value for these variables. This binary representation was done to improve the interpretability.

2.4 Data Standarization

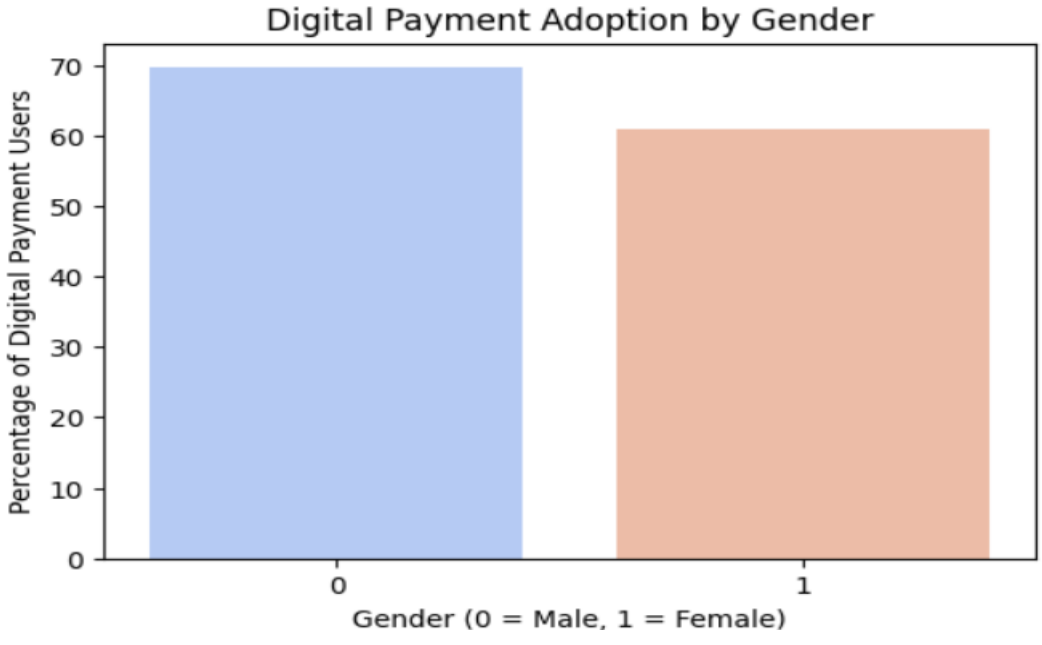
StandardScaler was used to standardise numerical features in order to improve model performance prior to the model building process.

3. Initial Descriptive Analysis

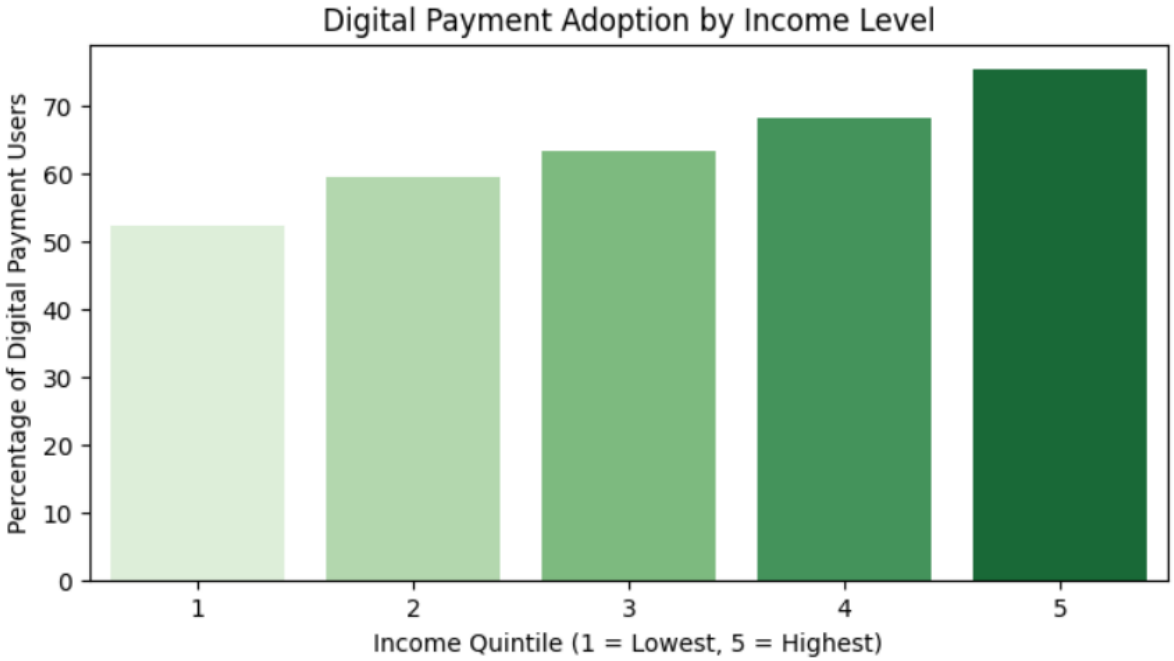
- The variable ‘anydigpayment’ was used to calculate the digital payment rate. Of all the respondents in the dataset, 76% of individuals in the dataset reported using digital payments indicating a 76% digital payment adoption rate.



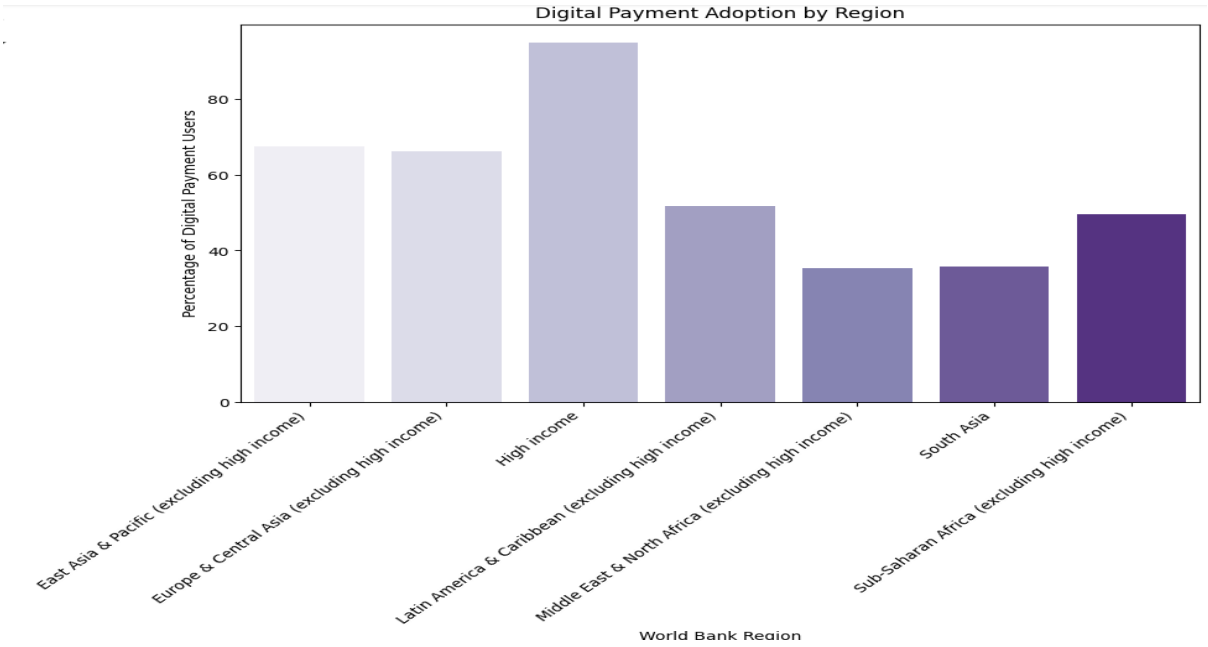
- **Adoption By Gender:** Males showed slightly higher adoption rates compared to females.



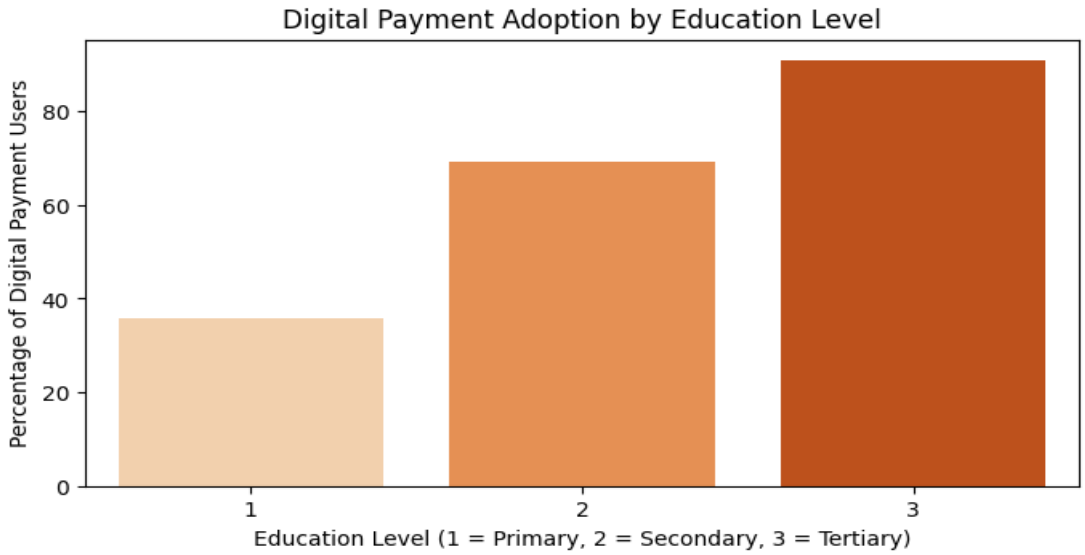
- **Adoption by Income Level**: Higher-income individuals had significantly higher adoption rates. It varied from about 53% in the 1st quintile to 72 % in the 5th quintile. While low adoption levels were expected in the lower income groups, among those in the highest quintile of income the adoption level is still only around 73%, which indicates reluctance to adopt digital payment methods even among those earning the most.



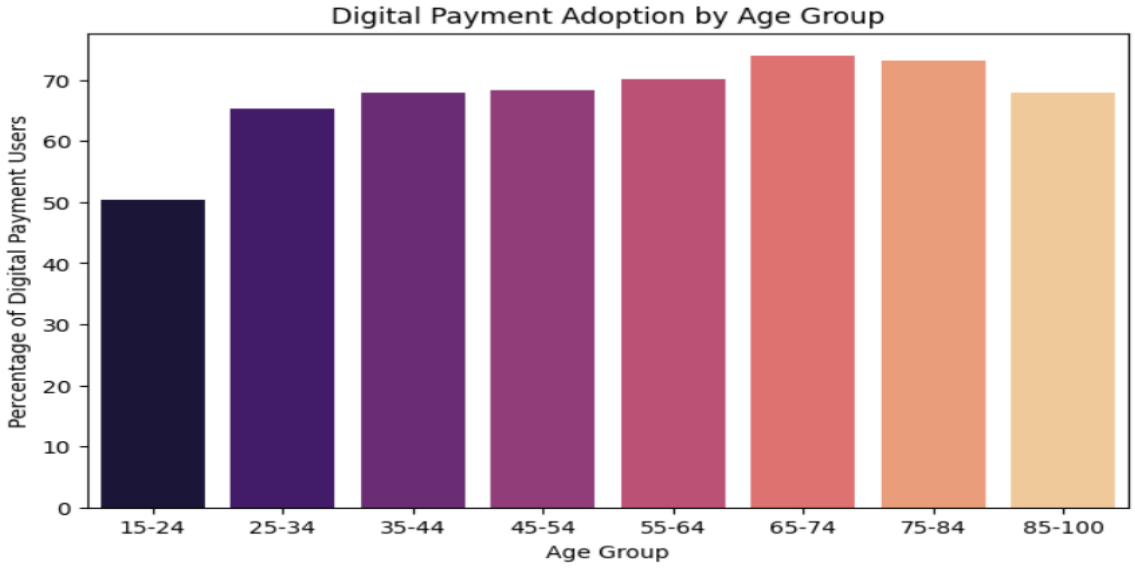
- **Adoption by Region**: Among the low- and mid-income levels, East Asia& pacific and Europe & central Asia had the highest adoption, while South Asia lagged. The high income was studied as a separate category in each of the regions.



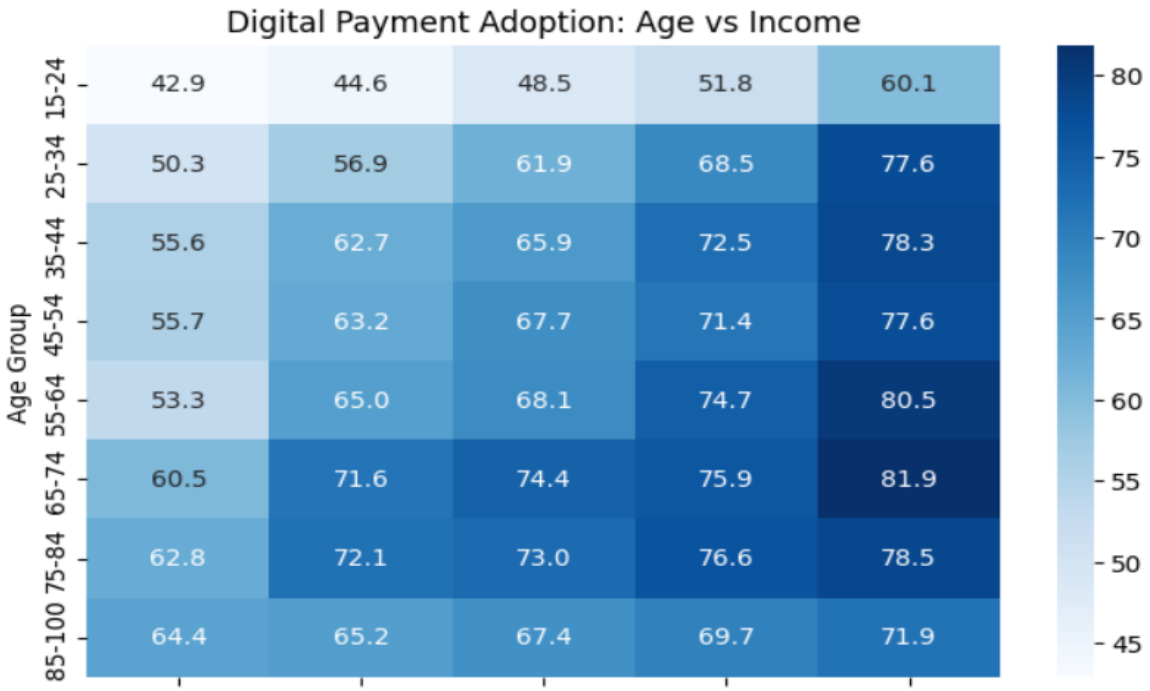
- **Adoption By Education**
  There was a stark difference in adoption rates basis education with the digital adoption rates. Among those with only a primary education the adoption rate was significantly low at around 37% while among those who had a tertiary level of education the adoption rate was significantly higher at around 90%. This shows that education level is a significant predicator of the digital payment adoption rate.



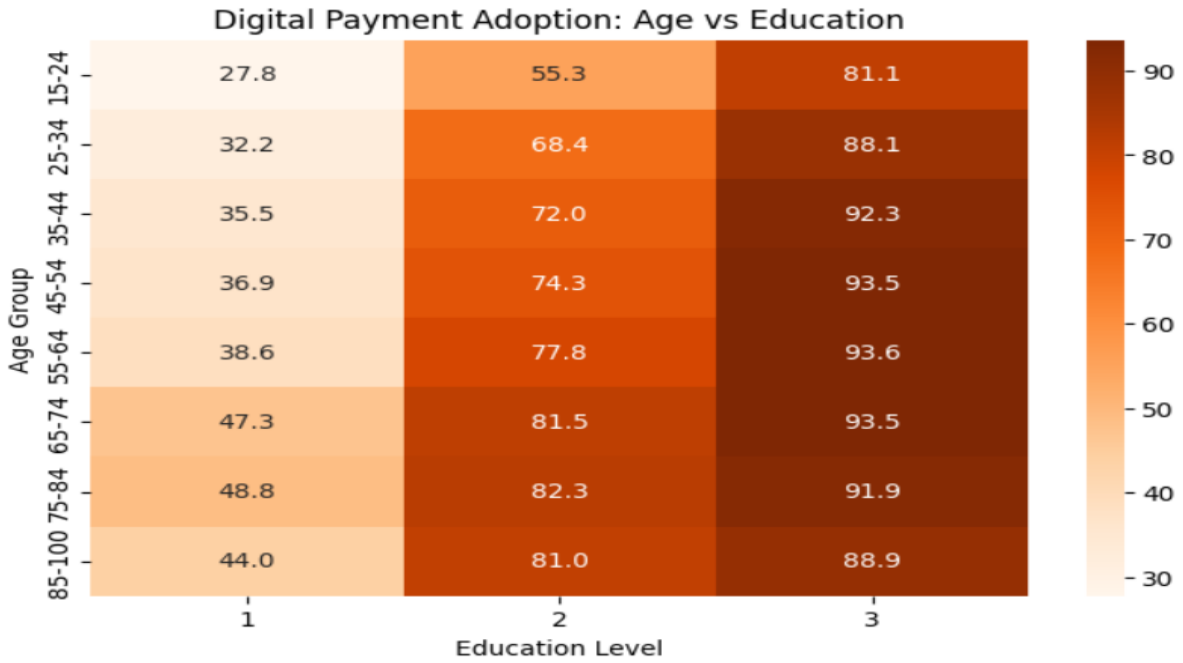
- **Adoption By Age**
  Adoption of digital payments rises from the 15–24 age group (about 50%) to the 65–74 and 75–84 age groups (over 70%), according to the bar chart that depicts this trend. There is a modest reduction in the 85–100 age range, which could be caused by technological hurdles. This implies that younger users are not the only ones driving the growth of digital payments, underscoring the need for financial technology solutions that are age inclusive.



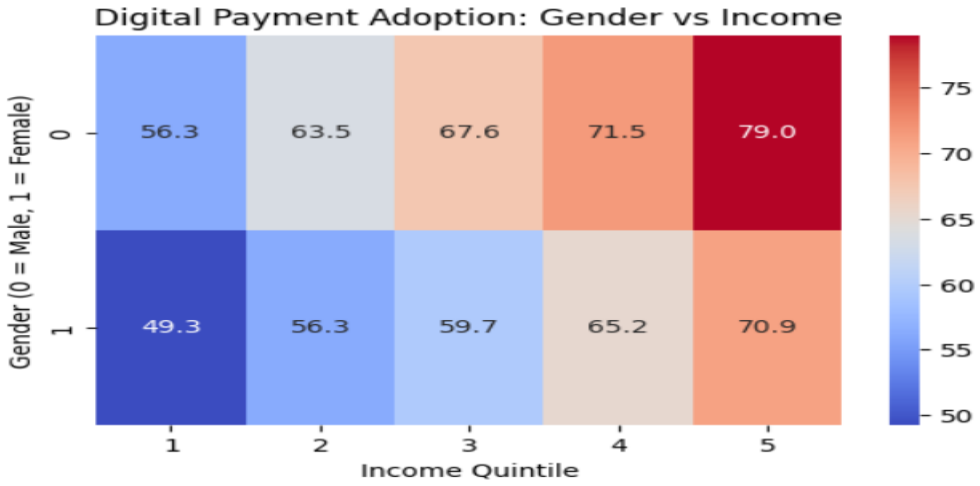
- **Combined Effect of Age and Income on digital payments adoption.**
  The use of digital payments across age and economic brackets is depicted in the heatmap below. Across all age categories, adoption rates often rise with income, suggesting a close relationship between financial capability and the use of digital payments. While middle-aged and older persons (ages 55–84) have the highest adoption rates, especially in higher income brackets, younger age groups (15–24) have lower adoption rates, especially at lower income levels. Despite greater income levels, adoption is somewhat lower for the oldest age group (85–100), either as a result of technological hurdles or a preference for traditional banking. This realisation emphasises the necessity of focused financial inclusion initiatives, particularly for the younger and older generations.



- **Influence of Education and Age on Digital Payment Adoption**
  This heatmap illustrates how usage of digital payments varies by age group and education level. Higher levels of education are clearly associated with higher adoption rates across all age groups. Adoption is substantially lower among those with the least amount of education, especially in the younger and older age groups. On the other hand, the highest adoption rates—which frequently surpass 90%—are generally seen among persons with the highest levels of education. In younger groups, where adoption almost triples between the lowest and greatest education levels, the influence of education is particularly noticeable. These findings emphasise how important digital education and financial literacy are to increasing the uptake of digital payments, particularly among younger and less educated populations.



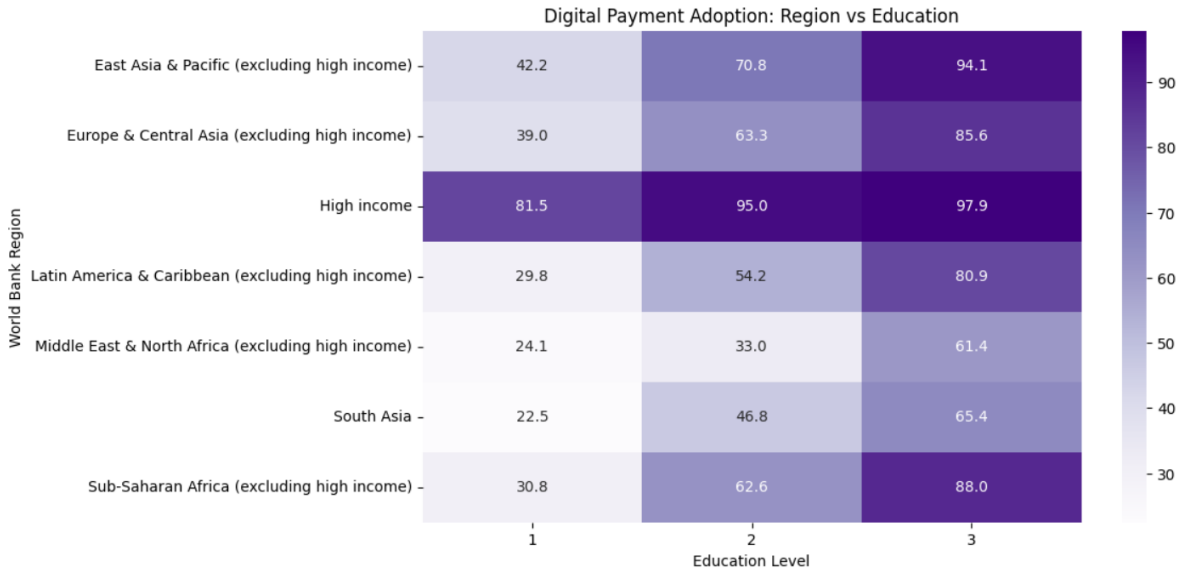
- **Impact of Gender and Income on Digital Payment Adoption**
  The association between male and female adoption of digital payments and income levels is depicted in this heatmap. Both genders' adoption of digital payments rises with income, suggesting a robust relationship between financial capability and the use of digital transactions. At every income quintile, however, adoption rates are consistently higher for males than for females. At lower income levels, where men accept digital payments at a significantly faster rate than women, the gender difference is particularly noticeable. The gender gap narrows but does not disappear as money increases. These results imply that although income is a significant determinant in the adoption of digital payments, adoption rates may also be influenced by gender-specific factors like financial literacy, access to digital banking, and social norms.



- **Impact of Education and Regional Disparities on Digital Payment Adoption**

The heatmap below highlights the relationship between education levels and digital payment adoption across different World Bank regions. A clear trend emerges as a higher education levels strongly correlate with increased digital payment adoption in all regions.

  - **High-Income Countries** show the highest adoption rates, with even individuals at the lowest education level displaying significant digital payment usage.
  - **Developing Regions** (e.g., South Asia, Sub-Saharan Africa, and the Middle East & North Africa) exhibit lower adoption rates, particularly among individuals with minimal education. However, adoption increases sharply with higher education.
  - **Regional Disparities** exist, with Europe & Central Asia and East Asia & Pacific showing relatively higher adoption rates compared to Latin America, South Asia, and Sub-Saharan Africa at similar education levels.
  - **Policy Implications**: Bridging the digital payment gap in lower-income regions may require targeted interventions in financial literacy, digital infrastructure, and access to banking services, particularly for those with lower education levels.

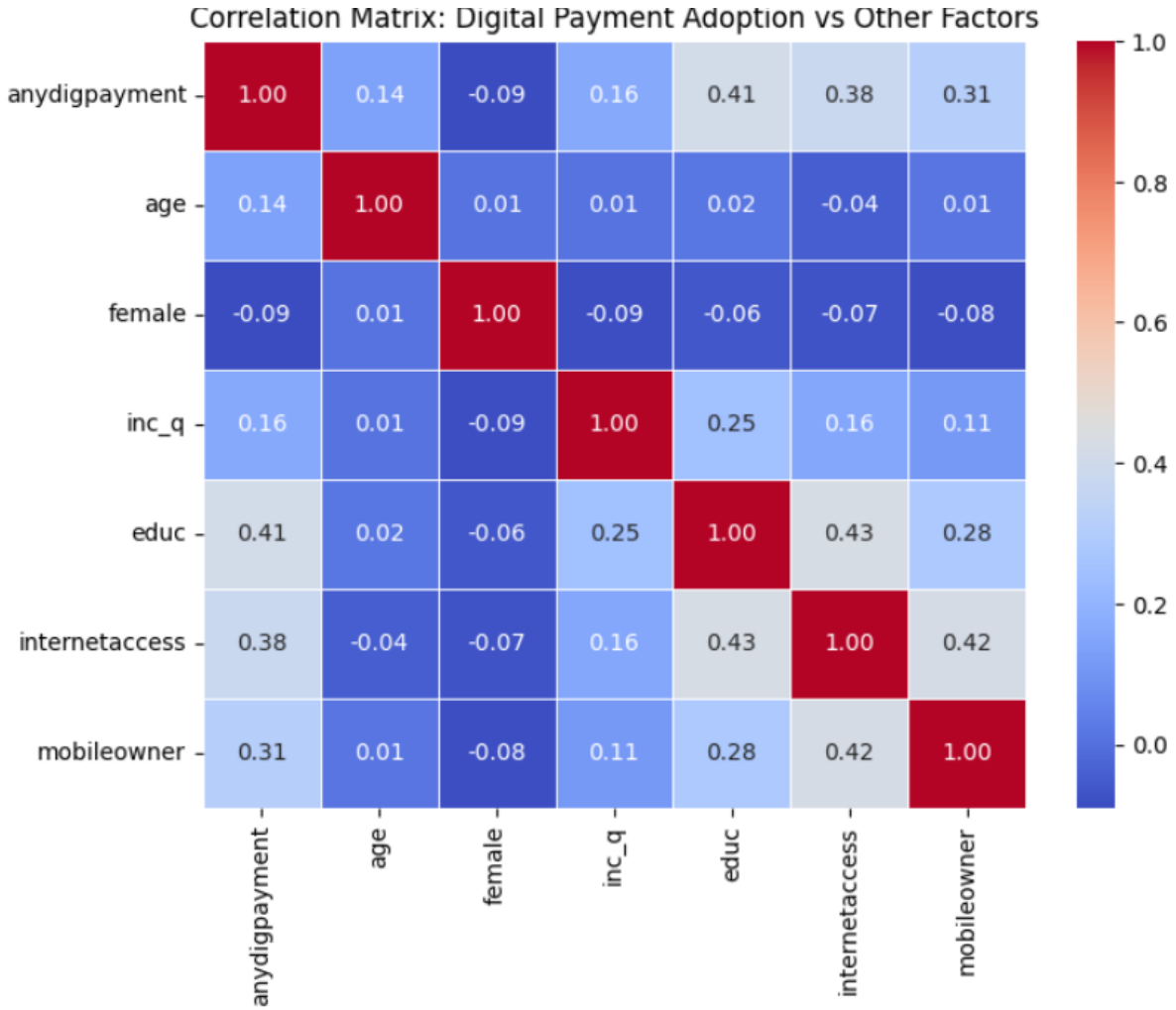


- **Correlation Matrix**

This correlation matrix reveals the relationship between digital payment adoption and various socioeconomic factors. Key insights include:

  - **Education (0.41)** and **internet access (0.38)** show the highest positive correlation with digital payment adoption, suggesting that higher education levels and internet connectivity significantly enhance digital payment usage.
  - A moderate positive correlation (**0.31**) between mobile ownership and digital payment adoption highlights the role of mobile devices in enabling financial transactions.
  - While higher income levels generally promote digital payment use, the impact is less pronounced compared to education and internet access.

- Gender (**-0.09**) and age (**0.14**) exhibit weak correlations, indicating that digital payment adoption is relatively uniform across these demographics.



4. Predictive Modelling

4.1 Model Building

After the initial EDA and finding the effect of various variables on the adoption of digital payment methods, we went ahead and built a predictive model that predicts if a person would adopt digital payment method given the following predictor variables

a) Age
b) Gender
c) Income Quantile
d) Education level
e) How the respondent makes a utility payment
f) Access To Internet
g) Mobile Phone Ownership

The target variable was anydigpayment. We built 4 classification models using

a) Decision Tree
b) Logistic Regression
c) Random Forest
d) XGBoost Classifier

The four different models were tried in order to evaluate which model would perform the best. A train test split of 80-20 was used.

4.2 Model Evaluation

The four models were evaluated using the metrics of accuracy, precision, recall and f1 score. The evaluation scores from the various models and their respective confusion matrix is given below

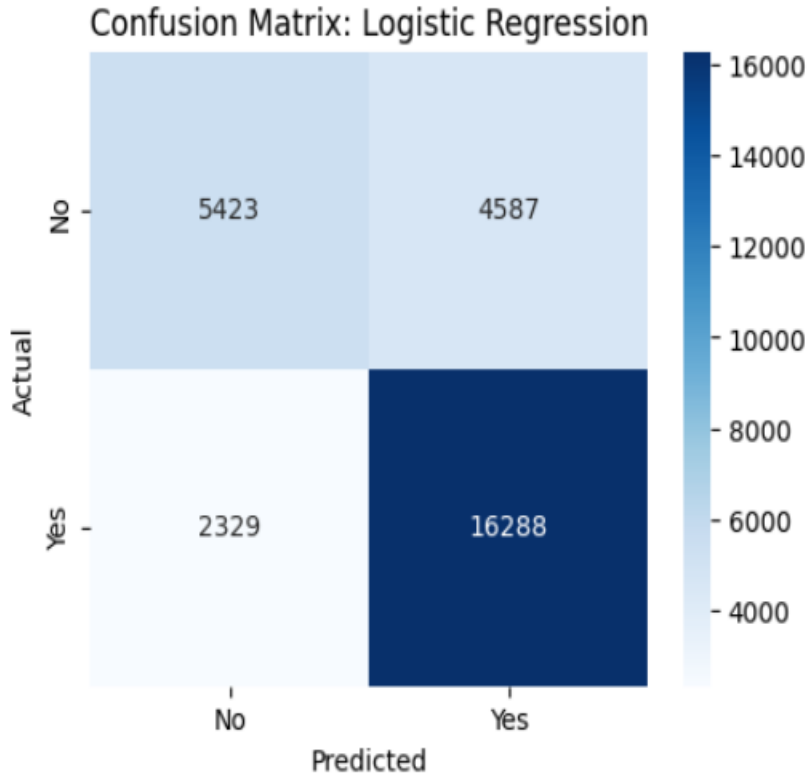
a) Logistic Regression

```
Logistic Regression Accuracy: 0.7584

Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.54      0.61     10010
           1       0.78      0.87      0.82     18617

    accuracy                           0.76     28627
   macro avg       0.74      0.71      0.72     28627
weighted avg       0.75      0.76      0.75     28627
```



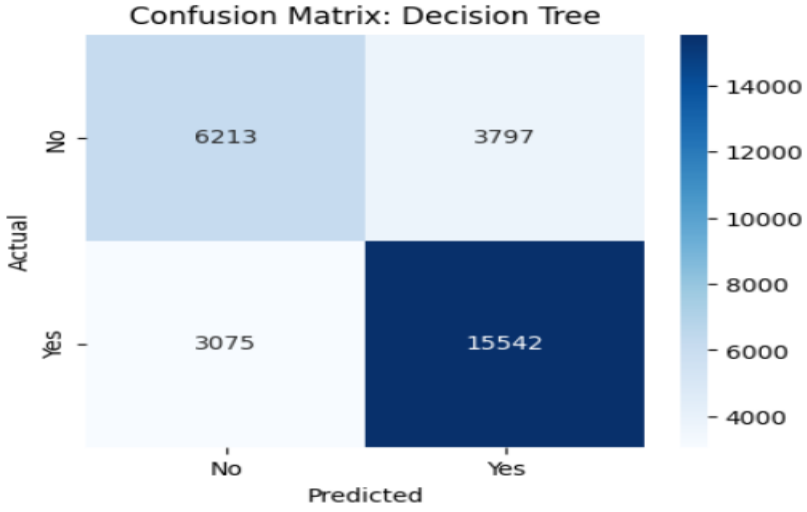
b) Decision Tree

```
Decision Tree Accuracy: 0.7599

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.62      0.64     10010
           1       0.80      0.83      0.82     18617

    accuracy                           0.76     28627
   macro avg       0.74      0.73      0.73     28627
weighted avg       0.76      0.76      0.76     28627
```

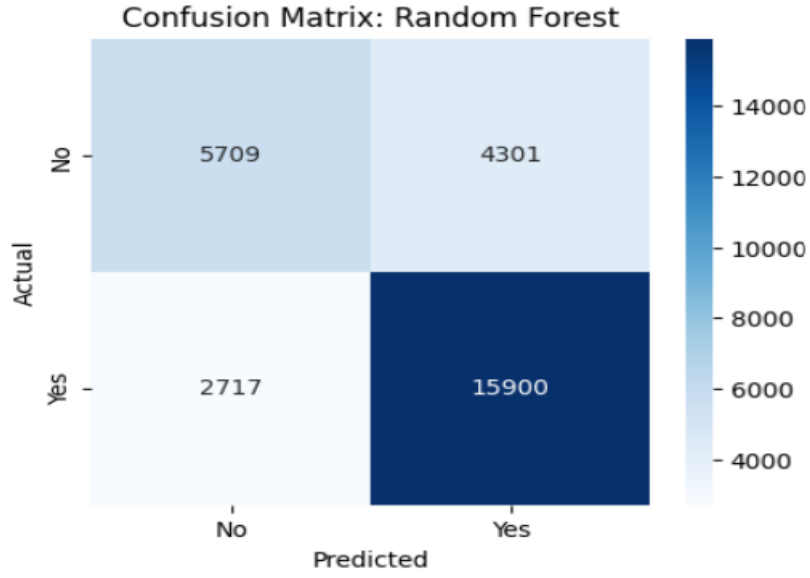


c) Random Forests

```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.57      0.62     10010
           1       0.79      0.85      0.82     18617

    accuracy                           0.75     28627
   macro avg       0.73      0.71      0.72     28627
weighted avg       0.75      0.75      0.75     28627
```



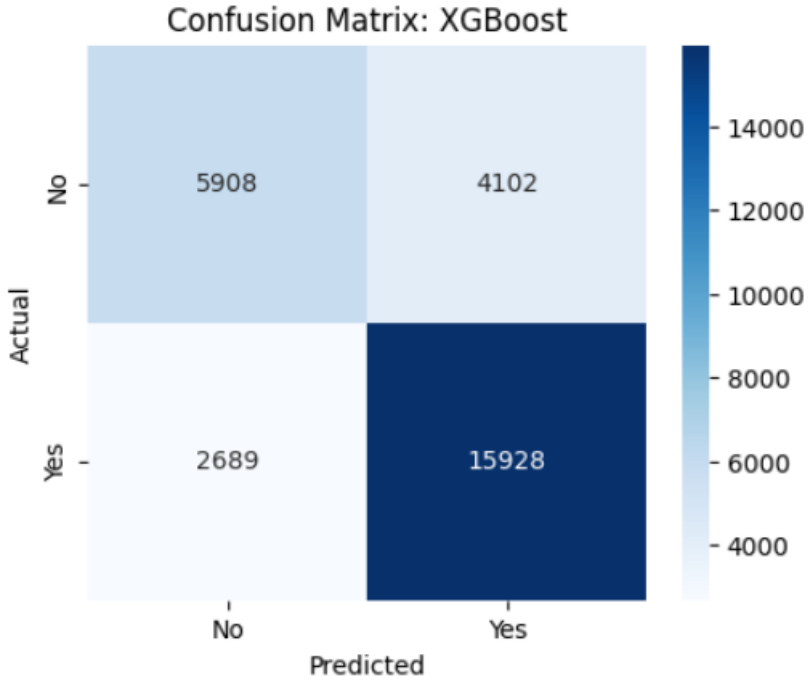
d) XGBoost

```
XGBoost Accuracy: 0.7628

XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.59      0.64     10010
           1       0.80      0.86      0.82     18617

    accuracy                           0.76     28627
   macro avg       0.74      0.72      0.73     28627
weighted avg       0.76      0.76      0.76     28627
```



Since the dataset was imbalanced one, accuracy was not a good enough measure to evaluate how good the model is. Since in this case we needed to minimize the false negatives for class 0 (that is participants predicted as will adopt a digital payment method but actually does not), we decided to go with recall for the negative class as the key metric for model evaluation.

For the following reasons**, Decision Tree** was chosen as the **best model** even though XGBoost had a little higher accuracy (76.28%):

Interpretability: Policymakers and financial institutions can easily comprehend adoption considerations thanks to Decision Trees' explicit decision rules.

Performance and Simplicity Balance: It provides high accuracy (75.99%) without resorting to the overfitting and unnecessary complexity that Random Forest is known for.

Better Handling of Non-Linearity: Decision Trees, as opposed to Logistic Regression, are able to identify intricate linkages in the adoption trends of digital payments.

Reduced Computational Cost: Decision Trees are feasible for extensive financial inclusion research since they use less computing power than XGBoost.

Highest recall For Class 0: Decision tree had the highest recall for class 0(i.e those not adopting digital payment methods). This model therefore had the least wrong predictions for people who didn't adopt a digital payment method but were predicted to do so while compared to the other models.

5. Results and Implications

Improve Financial Literacy and Digital Education

- Strengthen digital education initiatives to increase adoption, especially among lower-income and less-educated populations.
- Develop targeted financial literacy programs for younger and older age groups to bridge knowledge gaps.
- Collaborate with educational institutions to integrate digital finance modules into curriculums.

Expand Digital Infrastructure & Internet Access

- Invest in internet connectivity, particularly in underserved rural and low-income regions.
- Encourage affordable smartphone access and digital banking solutions.
- Develop mobile-friendly financial services to improve accessibility.

Address Income-Related Barriers

- Provide incentives and subsidies for digital financial tools for lower-income groups.
- Encourage fintech innovations that cater to people with irregular incomes.
- Design financial products with minimal transaction fees to encourage adoption.

Strengthen Trust in Digital Payments

- Address gender-related adoption gaps by promoting financial education among women.
- Implement robust consumer protection policies to safeguard users from fraud.
- Increase awareness and transparency about data security and digital transactions.

Implement Data-Driven Financial Inclusion Policies

- Utilize machine learning insights from the study to create targeted policy interventions.
- Develop region-specific policies based on adoption trends and barriers identified in the study.
- Encourage regulatory frameworks that promote digital payment adoption while ensuring security and accessibility.

6. Appendix-Codes

Loading the Data set

```
import pandas as pd

country_agg_data = pd.read_excel("CountryLevel_DatabankWide.xlsx",
sheet_name=0)
country_databank = pd.read_csv("micro_world_139countries.csv", encoding='ISO-
8859-1')
```

Retaning only the relevant columns

```
selected_columns = [ 'economy', 'regionwb', 'age', 'female', 'inc_q', 'educ',
   'anydigpayment', 'pay_utilities',   'internetaccess', 'mobileowner', 'year']
country_databank = country_databank[selected_columns].copy()
```

Data Cleaning & Preparation

```
country_databank['female'] = country_databank['female'].replace({2: 0, 1: 1})
country_databank['internetaccess'] = country_databank['internetaccess'].replace({2:
0})
country_databank['mobileowner'] = country_databank['mobileowner'].replace({2:
0})  # (1 = Yes, 0 = No)

# Missing value treatment
country_databank = country_databank.dropna(subset=['anydigpayment'])
country_databank['pay_utilities'].fillna(0, inplace=True)
country_databank = country_databank[~country_databank['educ'].isin([4, 5])]
```

```
print(country_databank.head())

# Save cleaned data
country_databank.to_csv("Cleaned_Digital_Payment_Data.csv", index=False)

import matplotlib.pyplot as plt
import seaborn as sns
# Summary statistics
summary_stats = country_databank.describe()
summary_stats
```

Digital Payment Adoption Rate

```
digital_payment_rate =
country_databank['anydigpayment'].value_counts(normalize=True) * 100
print("Digital Payment Adoption Rate:\n", digital_payment_rate)

gender_payment = country_databank.groupby('female')['anydigpayment'].mean() *
100
print("\nDigital Payment Adoption by Gender:\n", gender_payment)

income_payment = country_databank.groupby('inc_q')['anydigpayment'].mean() * 100
print("\nDigital Payment Adoption by Income Level:\n", income_payment)

education_payment = country_databank.groupby('educ')['anydigpayment'].mean() *
100
print("\nDigital Payment Adoption by Education Level:\n", education_payment)

region_payment = country_databank.groupby('regionwb')['anydigpayment'].mean() *
100
print("\nDigital Payment Adoption by Region:\n", region_payment)

plt.figure(figsize=(6,4))
sns.barplot(x=digital_payment_rate.index, y=digital_payment_rate.values,
palette="Blues")
plt.xlabel("Digital Payment Adoption (0 = No, 1 = Yes)")
plt.ylabel("Percentage")
plt.title("Overall Digital Payment Adoption")
plt.show()
```

Digital Payment Rate by Gender

```
plt.figure(figsize=(6,4))
sns.barplot(x=gender_payment.index, y=gender_payment.values, palette="coolwarm")
plt.xlabel("Gender (0 = Male, 1 = Female)")
```

```
plt.ylabel("Percentage of Digital Payment Users")
plt.title("Digital Payment Adoption by Gender")
plt.show()
```

Digital Payment Rate by Income Level

```
plt.figure(figsize=(8,4))
sns.barplot(x=income_payment.index, y=income_payment.values, palette="Greens")
plt.xlabel("Income Quintile (1 = Lowest, 5 = Highest)")
plt.ylabel("Percentage of Digital Payment Users")
plt.title("Digital Payment Adoption by Income Level")
plt.show()
```

Digital Payment Rate by Education Level

```
plt.figure(figsize=(8,4))
sns.barplot(x=education_payment.index, y=education_payment.values,
palette="Oranges")
plt.xlabel("Education Level (1 = Primary, 2 = Secondary, 3 = Tertiary)")
plt.ylabel("Percentage of Digital Payment Users")
plt.title("Digital Payment Adoption by Education Level")
plt.show()
```

Digital Payment Rate by Region

```
plt.figure(figsize=(10,6))
sns.barplot(x=region_payment.index, y=region_payment.values, palette="Purples")
plt.xticks(rotation=45, ha="right")
plt.xlabel("World Bank Region")
plt.ylabel("Percentage of Digital Payment Users")
plt.title("Digital Payment Adoption by Region")
plt.show()
```

Pivot table for variables

```
pivot_table = pd.pivot_table(
    country_databank,
    values='anydigpayment',
    index=['regionwb', 'inc_q', 'educ', 'female'],
    aggfunc='mean'
) * 100

pivot_table = pivot_table.reset_index()
pivot_table.rename(columns={'anydigpayment': 'Digital Payment Adoption (%)'},
inplace=True)

from IPython.display import display
```

```
display(pivot_table)
```

Random Forest Regressor for imputing missing age

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.impute import SimpleImputer

features = ['female', 'inc_q', 'educ', 'internetaccess', 'mobileowner', 'anydigpayment']

age_known = country_databank.dropna(subset=['age'])

age_missing = country_databank[country_databank['age'].isna()]

X = age_known[features]
y = age_known['age']

imputer = SimpleImputer(strategy='most_frequent')
X = imputer.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

if not age_missing.empty:
   X_missing = imputer.transform(age_missing[features])
   predicted_ages = model.predict(X_missing)

   country_databank.loc[country_databank['age'].isna(), 'age'] = predicted_ages

print("Missing age values after imputation:", country_databank['age'].isna().sum())
```

Bucketing age for easier visualization

```
age_bins = [15, 25, 35, 45, 55, 65, 75, 85, 100]
age_labels = ["15-24", "25-34", "35-44", "45-54", "55-64", "65-74", "75-84", "85-
100"]

country_databank['age_group'] = pd.cut(country_databank['age'], bins=age_bins,
labels=age_labels, right=False)

age_adoption = country_databank.groupby('age_group')['anydigpayment'].mean() *
100
```

```
print("Digital Payment Adoption Rates by Age Group:\n", age_adoption)

plt.figure(figsize=(8,5))
sns.barplot(x=age_adoption.index, y=age_adoption.values, palette="magma")
plt.xlabel("Age Group")
plt.ylabel("Percentage of Digital Payment Users")
plt.title("Digital Payment Adoption by Age Group")
plt.show()
```

Pivots/Graphs from EDA

```
# Age Group vs Income Level
pivot_age_income = pd.pivot_table(
   country_databank, values='anydigpayment', index='age_group', columns='inc_q',
aggfunc='mean'
) * 100
pivot_age_income

# Heatmap for Age Group vs Income Level
plt.figure(figsize=(8, 5))
sns.heatmap(pivot_age_income, cmap="Blues", annot=True, fmt=".1f")
plt.xlabel("Income Quintile")
plt.ylabel("Age Group")
plt.title("Digital Payment Adoption: Age vs Income")
plt.show()

# Age Group vs Education Level
pivot_age_education = pd.pivot_table(
   country_databank, values='anydigpayment', index='age_group', columns='educ',
aggfunc='mean'
) * 100
pivot_age_education

# Heatmap for Age Group vs Education Level
plt.figure(figsize=(8, 5))
sns.heatmap(pivot_age_education, cmap="Oranges", annot=True, fmt=".1f")
plt.xlabel("Education Level")
plt.ylabel("Age Group")
plt.title("Digital Payment Adoption: Age vs Education")
plt.show()

# Gender vs Income Level
pivot_gender_income = pd.pivot_table(
```

```
    country_databank, values='anydigpayment', index='female', columns='inc_q',
aggfunc='mean'
) * 100
pivot_gender_income

# Heatmap for Gender vs Income Level
plt.figure(figsize=(6, 4))
sns.heatmap(pivot_gender_income, cmap="coolwarm", annot=True, fmt=".1f")
plt.xlabel("Income Quintile")
plt.ylabel("Gender (0 = Male, 1 = Female)")
plt.title("Digital Payment Adoption: Gender vs Income")
plt.show()


# Region vs Education Level
pivot_region_education = pd.pivot_table(
    country_databank, values='anydigpayment', index='regionwb', columns='educ',
aggfunc='mean'
) * 100
pivot_region_education


# Heatmap for Region vs Education Level
plt.figure(figsize=(10, 6))
sns.heatmap(pivot_region_education, cmap="Purples", annot=True, fmt=".1f")
plt.xlabel("Education Level")
plt.ylabel("World Bank Region")
plt.title("Digital Payment Adoption: Region vs Education")
plt.xticks(rotation=0)
plt.show()

# Region vs Age Group
pivot_region_age = pd.pivot_table(
    country_databank, values='anydigpayment', index='regionwb',
columns='age_group', aggfunc='mean'
) * 100
pivot_region_age


# Heatmap for Region vs Age Group
plt.figure(figsize=(10, 5))
sns.heatmap(pivot_region_age, cmap="Blues", annot=True, fmt=".1f")
plt.xlabel("Age Group")
```

```
plt.ylabel("World Bank Region")
plt.title("Digital Payment Adoption: Region vs Age")
plt.xticks(rotation=0)
plt.show()


# Education Level vs Gender
pivot_education_gender = pd.pivot_table(
   country_databank, values='anydigpayment', index='educ', columns='female',
aggfunc='mean'
) * 100

# Heatmap for Education Level vs Gender
plt.figure(figsize=(6, 4))
sns.heatmap(pivot_education_gender, cmap="coolwarm", annot=True, fmt=".1f")
plt.xlabel("Gender (0 = Male, 1 = Female)")
plt.ylabel("Education Level")
plt.title("Digital Payment Adoption: Education vs Gender")
plt.yticks(rotation=0)
plt.show()


correlation_columns = ['anydigpayment', 'age', 'female', 'inc_q', 'educ', 'internetaccess',
'mobileowner']

correlation_matrix = country_databank[correlation_columns].corr()

print("\nCorrelation Matrix:\n", correlation_matrix)

plt.figure(figsize=(8,6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5)
plt.title("Correlation Matrix: Digital Payment Adoption vs Other Factors")
plt.show()
```

Predictive Modelling

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Import Models
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier

features = ['age', 'female', 'inc_q', 'educ', 'internetaccess', 'mobileowner']
target = 'anydigpayment'

country_databank = country_databank.dropna(subset=features + [target])

X = country_databank[features]
y = country_databank[target]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42)

def train_evaluate_model(model, model_name):
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    print(f"\n{model_name} Accuracy: {accuracy:.4f}")

    print(f"\n{model_name} Classification Report:\n", classification_report(y_test,
y_pred))

    conf_matrix = confusion_matrix(y_test, y_pred)
    plt.figure(figsize=(5,4))
    sns.heatmap(conf_matrix, annot=True, cmap="Blues", fmt='d', xticklabels=["No",
"Yes"], yticklabels=["No", "Yes"])
    plt.xlabel("Predicted")
    plt.ylabel("Actual")
    plt.title(f'Confusion Matrix: {model_name}')
    plt.show()

# Model 1: Logistic Regression
log_model = LogisticRegression()
train_evaluate_model(log_model, "Logistic Regression")
```

```
# Model 2: Decision Tree
dt_model = DecisionTreeClassifier(max_depth=5, random_state=42)
train_evaluate_model(dt_model, "Decision Tree"


# Model 3: Random Forest
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
train_evaluate_model(rf_model, "Random Forest")

# Model 4: XGBoost
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss',
random_state=42)
train_evaluate_model(xgb_model, "XGBoost")
```