Project Report : Aqua-Analytics

This document outlines the **project submitted** for the **Tech Ideathon 2025**, detailing the **problem statement**, our **innovative solution**, the **implementation plan**, and a thorough **impact analysis**.

Section 1 : Title

Project Title: Aqua-Analytics: A Data-Driven Insight into India's Drinking Water Accessibility

Tagline: From Coverage Metrics to Implementation Integrity

Submitted By: Sayantan Naha [**Innovation ID - ID20259NXL4I**]

Section 2 : The Problem Statement

Access to safe drinking water is a cornerstone of public health and economic stability in India. While national initiatives like the "Har Ghar Jal" mission have made monumental strides in expanding tap water coverage, a new, more complex challenge has emerged: ensuring the long-term integrity and verifiable success of this massive undertaking. Official reports may show high coverage percentages, but these numbers can often mask deep-seated issues. There is a critical gap between progress that is *reported* by local bodies and progress that is officially *certified* as complete and functional. This "Certification Deficit" represents a significant risk, indicating potential problems in data quality, last-mile execution, or administrative hurdles. Without a system to proactively identify and understand the drivers of this gap, resources cannot be targeted effectively, and the true impact of the mission remains opaque, putting the long-term sustainability of the investment at risk.

Section 3: Innovative Solution

Our solution, **Aqua-Analytics, is an AI-powered platform** designed to shift the focus from simple coverage metrics to a deeper analysis of implementation integrity. It provides a multi-layered insight into India's water accessibility challenge.

1. **Diagnostic Analysis:** We first aggregate and analyze multiple public datasets to create a comprehensive, state-level view of the current situation. This includes not just overall coverage, but also our own uniquely engineered metrics like a **Village Inequality Index** and a **PWS Infrastructure Gap**.
2. **Predictive AI:** The core of our innovation is an AI model that predicts a state's risk of developing a high **"Certification Deficit."** By analyzing the relationship between foundational infrastructure, socio-economic factors, and on-the-ground inequality, our model uncovers the hidden drivers of implementation challenges, allowing for proactive, data-driven interventions.

Section 4: The Core Innovation - The "Certification Deficit"

The most powerful feature of our project is a custom-engineered metric we call the **"Certification Deficit."**

Definition: *The percentage of villages that have been reported as 100% covered by the "Har Ghar Jal" mission but have not yet been officially certified.*

This metric serves as a powerful **"Implementation Integrity Score."** A high Certification Deficit doesn't mean the work hasn't been done; it acts as a critical early-warning signal for a wide range of potential issues:

- Administrative bottlenecks
- Data quality problems
- Delays in third-party verification
- On-the-ground quality issues preventing final sign-off

By predicting this single metric, we can help leaders move from a reactive to a proactive governance model.

Section 5: Implementation & Technology Stack

This project was executed as an **MVP (Minimum Viable Product)** to demonstrate feasibility, using a modern, agile, and accessible technology stack.

Implementation Phases:

1. **Data Aggregation & Cleaning:** Sourced and cleaned 6 distinct datasets from data.gov.in using Google Sheets.
2. **Feature Engineering:** Created unique, insightful metrics (Village_Inequality_Index, PWS_Infrastructure_Gap, Certification_Deficit_Pct) using VLOOKUP and custom formulas.
3. **Predictive Modeling:** Used a coding approach with Python (pandas, scikit-learn) in Google Colab to train a Random Forest Regressor model, chosen for its high performance and interpretability.
4. **Dashboarding:** Developed a multi-page interactive dashboard using Google Looker Studio to visualize and communicate the findings.

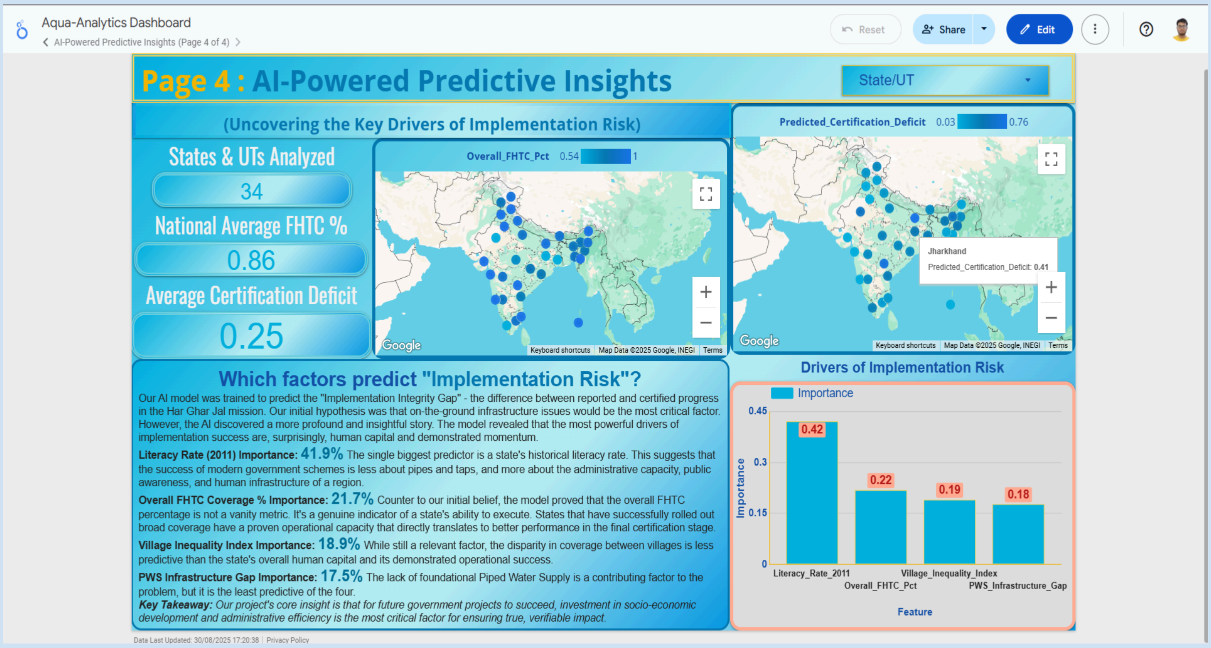
Technology Stack:

- **Data Preparation:** Google Sheets
- **AI/ML:** Python, Pandas, Scikit-learn (via Google Colab)
- **BI/Visualization:** Google Looker Studio

Section 6: Live Interactive Dashboard

Our final output is a 4-page interactive dashboard that guides the user through the **project's analytical story**, from the national overview to the final predictive insights.

Click Here to Explore the Live Aqua-Analytics Dashboard



Page 4 : AI-Powered Predictive Insights

(Uncovering the Key Drivers of Implementation Risk)

States & UTs Analyzed: 34

National Average FHTC %: 0.86

Average Certification Deficit: 0.25

Which factors predict "Implementation Risk"?

Our AI model was trained to predict the "Implementation Integrity Gap" - the difference between reported and certified progress in the Har Ghar Jal mission. Our initial hypothesis was that on-the-ground infrastructure issues would be the most critical factor. However, the AI discovered a more profound and insightful story. The model revealed that the most powerful drivers of implementation success are, surprisingly, human capital and demonstrated momentum.

Literacy Rate (2011) Importance: 41.9% The single biggest predictor is a state's historical literacy rate. This suggests that the success of modern government schemes is less about pipes and taps, and more about the administrative capacity, public awareness, and human infrastructure of a region.

Overall FHTC Coverage % Importance: 21.7% Counter to our initial belief, the model proved that the overall FHTC percentage is not a vanity metric. It's a genuine indicator of a state's ability to execute. States that have successfully rolled out broad coverage have a proven operational capacity that directly translates to better performance in the final certification stage.

Village Inequality Index Importance: 18.9% While still a relevant factor, the disparity in coverage between villages is less predictive than the state's overall human capital and its demonstrated operational success.

PWS Infrastructure Gap Importance: 17.5% The lack of foundational Piped Water Supply is a contributing factor to the problem, but it is the least predictive of the four.

Key Takeaway: *Our project's core insight is that for future government projects to succeed, investment in socio-economic development and administrative efficiency is the most critical factor for ensuring true, verifiable impact.*

(A screenshot of the main dashboard page is placed here)

Section 7: The Key Predictive Insight

Our AI model was trained to predict the "**Implementation Integrity Gap.**" While we hypothesized that infrastructure issues would be the most critical factor, the AI discovered a more profound story.

The model revealed that the most powerful drivers of implementation risk are **human capital** and **demonstrated operational momentum**.

1. Literacy Rate (41.9% Importance):
The single biggest predictor is a state's historical literacy rate. This suggests that the success of modern government schemes is less about pipes and taps, and more about the administrative capacity, public awareness, and human infrastructure of a region.
2. Overall FHTC Coverage % (21.7% Importance):
The model proved that the overall coverage percentage is a genuine indicator of a state's ability to execute. States that have successfully rolled out broad coverage have a proven operational capacity that directly translates to better performance in the final certification stage.

Section 8: Impact Analysis

Aqua-Analytics is designed to create tangible, real-world impact across multiple domains.

- **Social Impact:** By ensuring that reported progress translates to certified, functional connections, our tool helps guarantee that communities receive a consistent supply of safe drinking water, directly improving public health.
- **Economic Impact:** A proactive, predictive approach allows for the targeted allocation of funds to address administrative bottlenecks before they become costly delays, ensuring a higher return on investment for public funds.
- **Governance & Policy Impact:** The platform provides policymakers with an objective, data-driven early-warning system. It shifts governance from being reactive to proactive, ensuring that resources flow to the areas that need the most administrative and operational support.

Section 9: Future Scope & Scalability

This MVP is a powerful proof-of-concept. The next phases will focus on transforming it into a fully automated, production-ready platform.

- **Real-Time API Integration:** Transition from static datasets to live API feeds from government portals, using workflow automation tools like **n8n or Python** scripts to update the data daily.
- **Granular Analysis:** Expand the model to the **district** and even **village (Panchayat) level** as more granular data becomes available, providing hyper-local insights.
- **Enriched Data Sources:** Incorporate additional data streams, such as real-time water quality data, weather patterns, and public grievance reports, to make the **predictive model** even more **robust and accurate**.

Section 10: Thank You

Thank you for your time and consideration.

Sayantan Naha

Aqua-Analytics : An Insight into India's Water Accessibility

State/UT ▾

Predicting Implementation Integrity for India's Water Mission

This dashboard analyzes the current state of water access and uses AI to predict future challenges in the certification and implementation of the Har Ghar Jal mission, identifying states that require proactive support.

By Sayantan Naha

National Progress Report

Diagnostic Deep Dive

AI-powered Predictive Insights

Download Report

Page 2 : The National Progress Report

State/UT ▾

National Average FHTC %

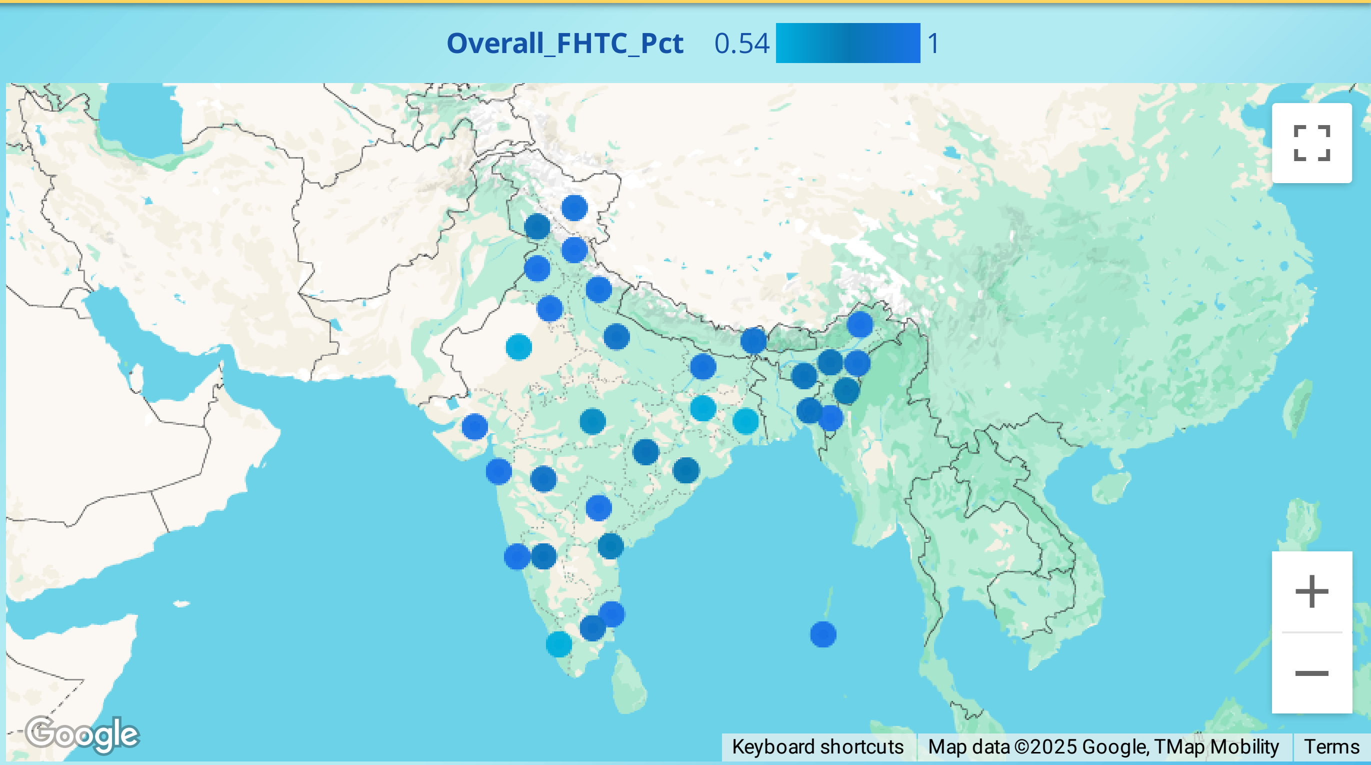
0.86

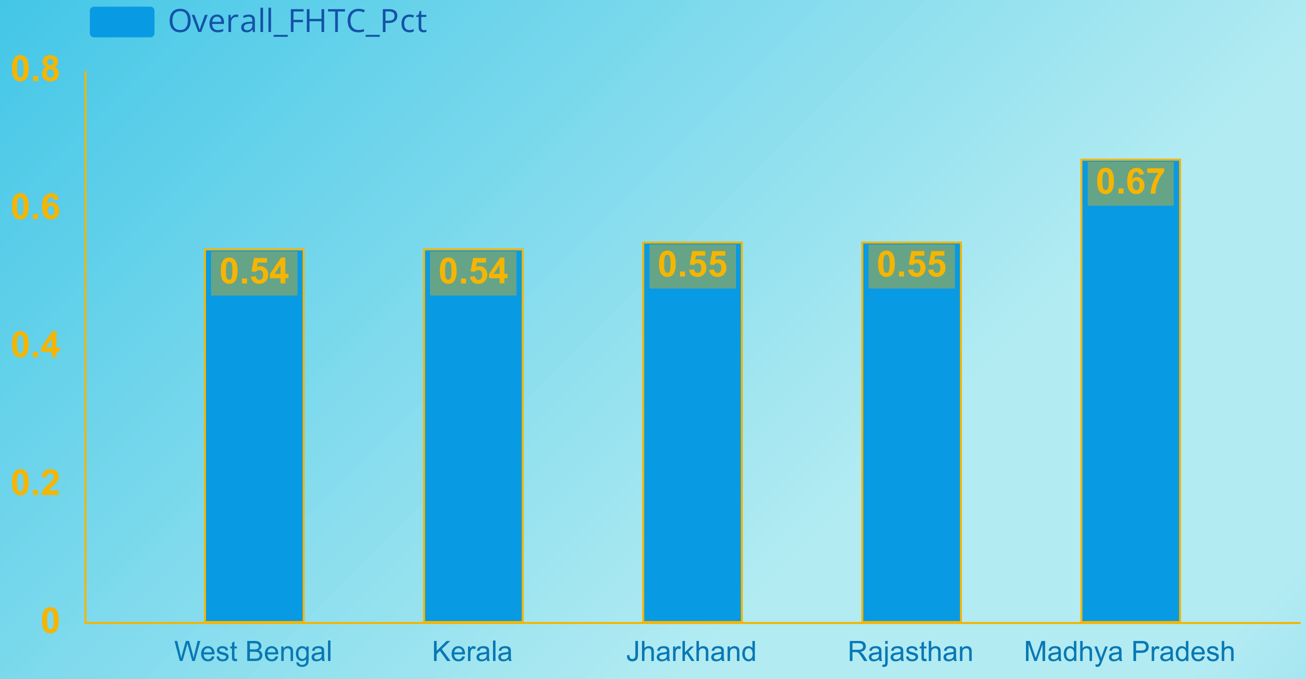
States & UTs Analyzed

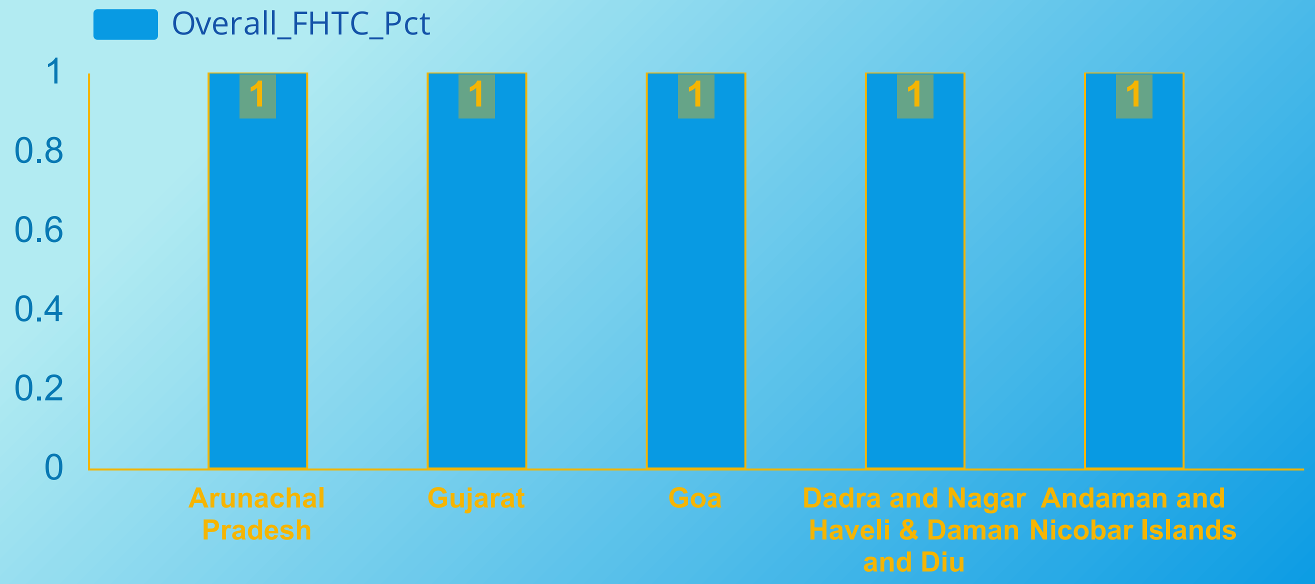
34

Average Certification Deficit

0.25







Page 3 : The Diagnostic Deep Dive

State/UT ▾

National Average FHTC %

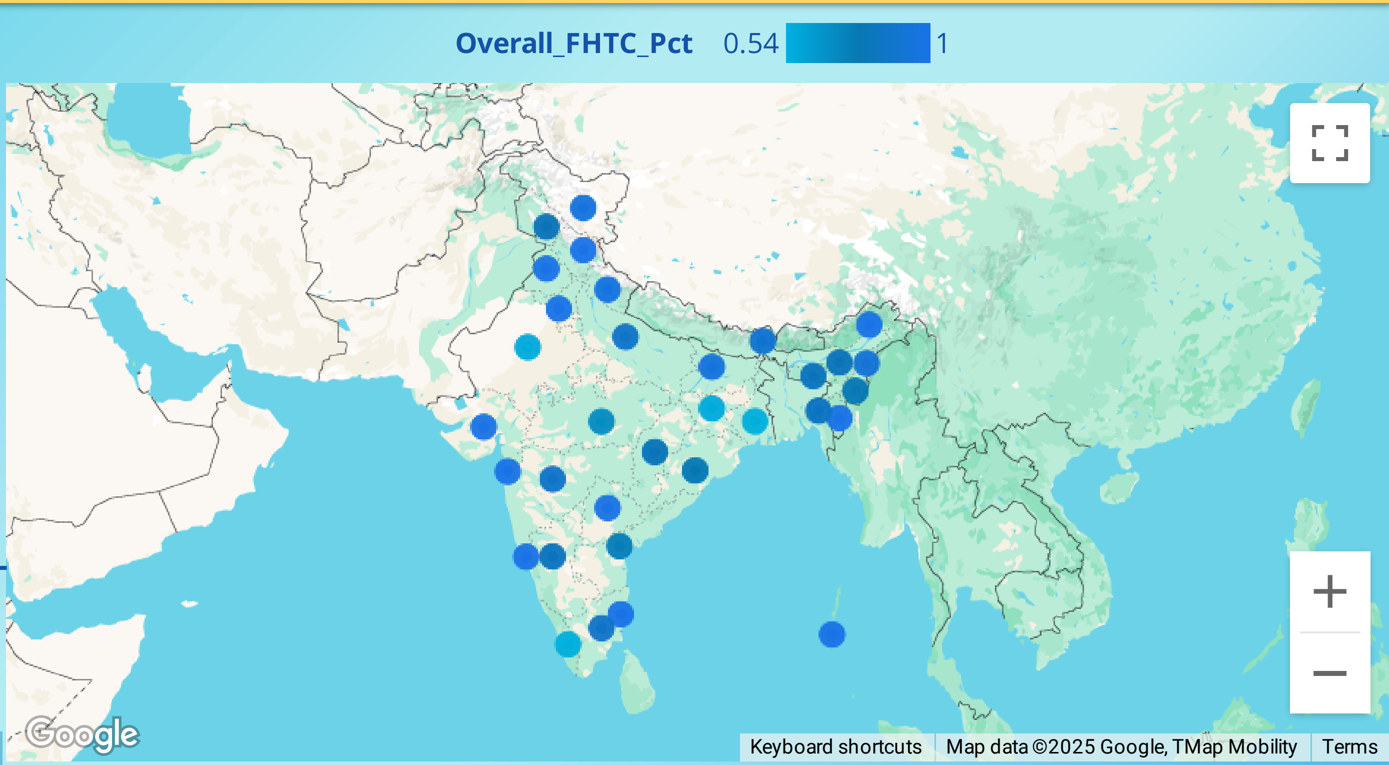
0.86

States & UTs Analyzed

34

Average Certification Deficit

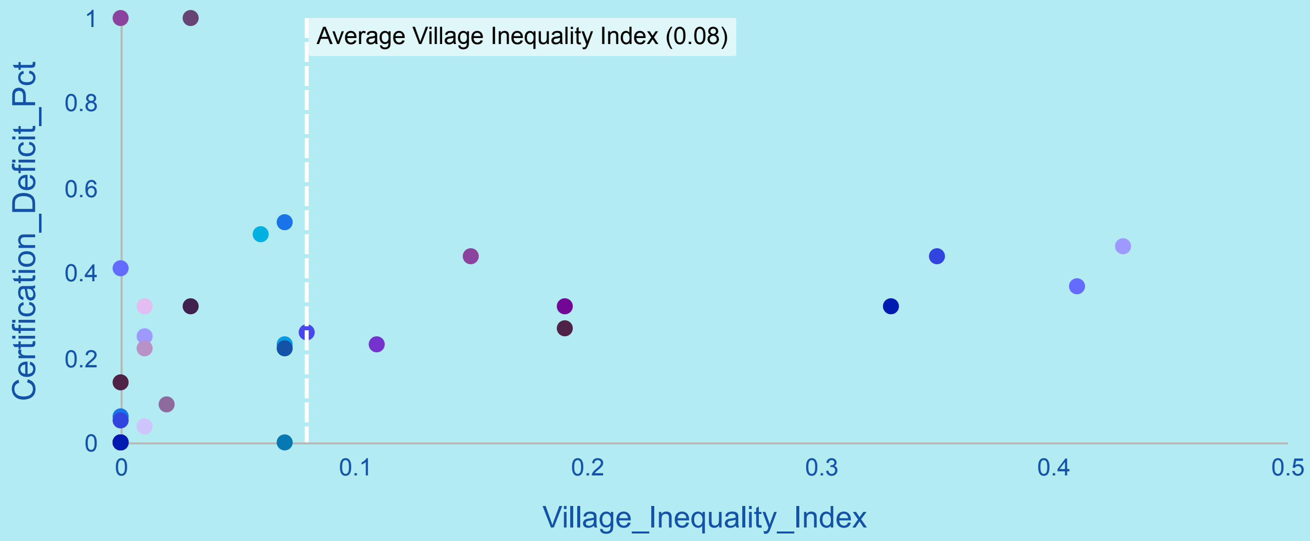
0.25



State-wise Risk Factor Analysis Table

State/UT ❶ ▾	Village_Inequality_Index	PWS_Infrastructure_Gap	Certification_Deficit_Pct ❷ ▾
West Bengal	0.41	0	0.37
Uttarakhand	0.01	0	0.22
Uttar Pradesh	0.08	0.01	0.26
Tripura	0.01	0	0.04
Telangana	0	0	1
Tamil Nadu	0.02	0	0.09
Sikkim	0.01	0	0.25
Rajasthan	0.43	0.03	0.46
Punjab	0	0	0
Puducherry	0	0	0

The Link Between Inequality and Certification Gaps



Page 4 : AI-Powered Predictive Insights

State/UT ▾

(Uncovering the Key Drivers of Implementation Risk)

States & UTs Analyzed

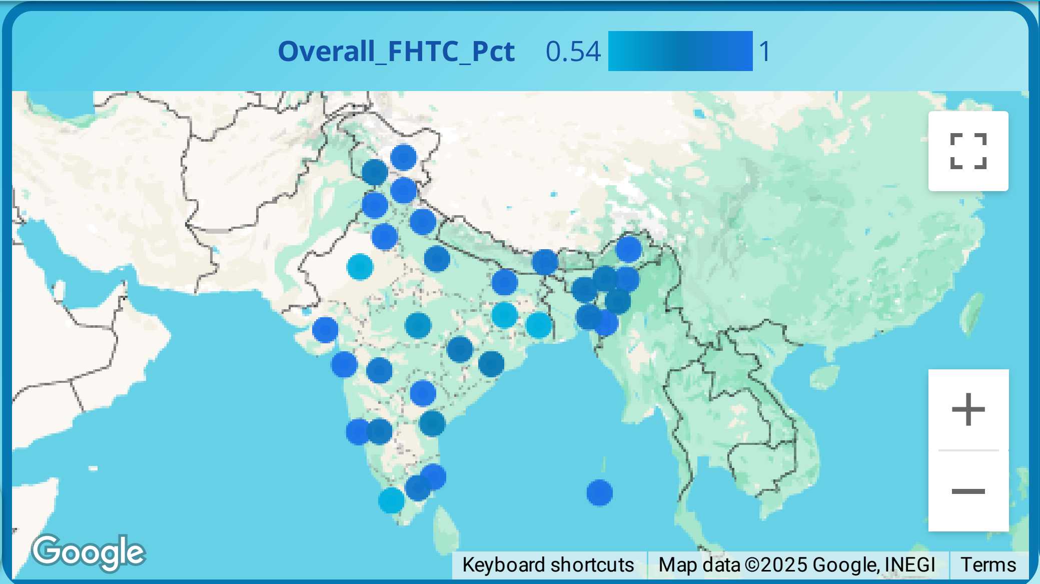
34

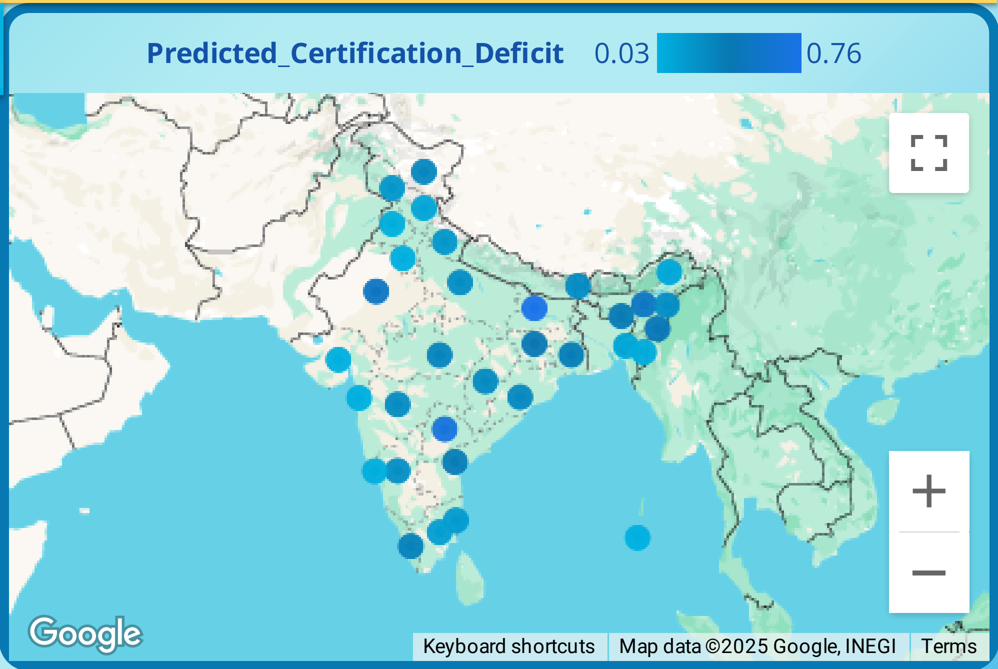
National Average FHTC %

0.86

Average Certification Deficit

0.25




Predicted_Certification_Deficit 0.03 0.76
Google
Keyboard shortcuts
Map data ©2025 Google, INEGI
Terms

Which factors predict "Implementation Risk"?

Our AI model was trained to predict the "Implementation Integrity Gap" - the difference between reported and certified progress in the Har Ghar Jal mission. Our initial hypothesis was that on-the-ground infrastructure issues would be the most critical factor. However, the AI discovered a more profound and insightful story. The model revealed that the most powerful drivers of implementation success are, surprisingly, human capital and demonstrated momentum.

Literacy Rate (2011) Importance: 41.9% The single biggest predictor is a state's historical literacy rate. This suggests that the success of modern government schemes is less about pipes and taps, and more about the administrative capacity, public awareness, and human infrastructure of a region.

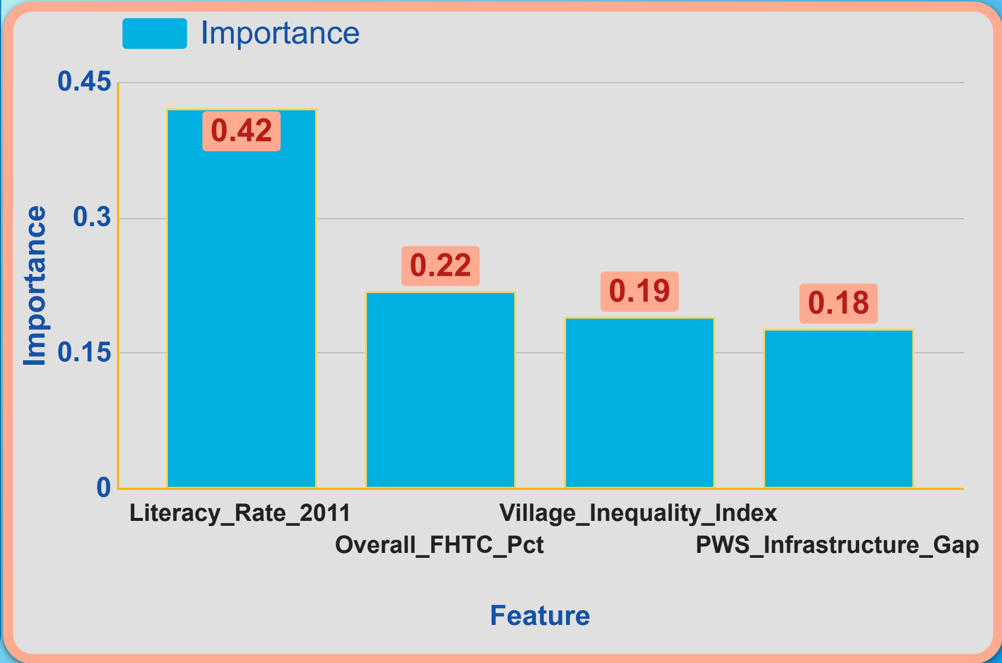
Overall FHTC Coverage % Importance: 21.7% Counter to our initial belief, the model proved that the overall FHTC percentage is not a vanity metric. It's a genuine indicator of a state's ability to execute. States that have successfully rolled out broad coverage have a proven operational capacity that directly translates to better performance in the final certification stage.

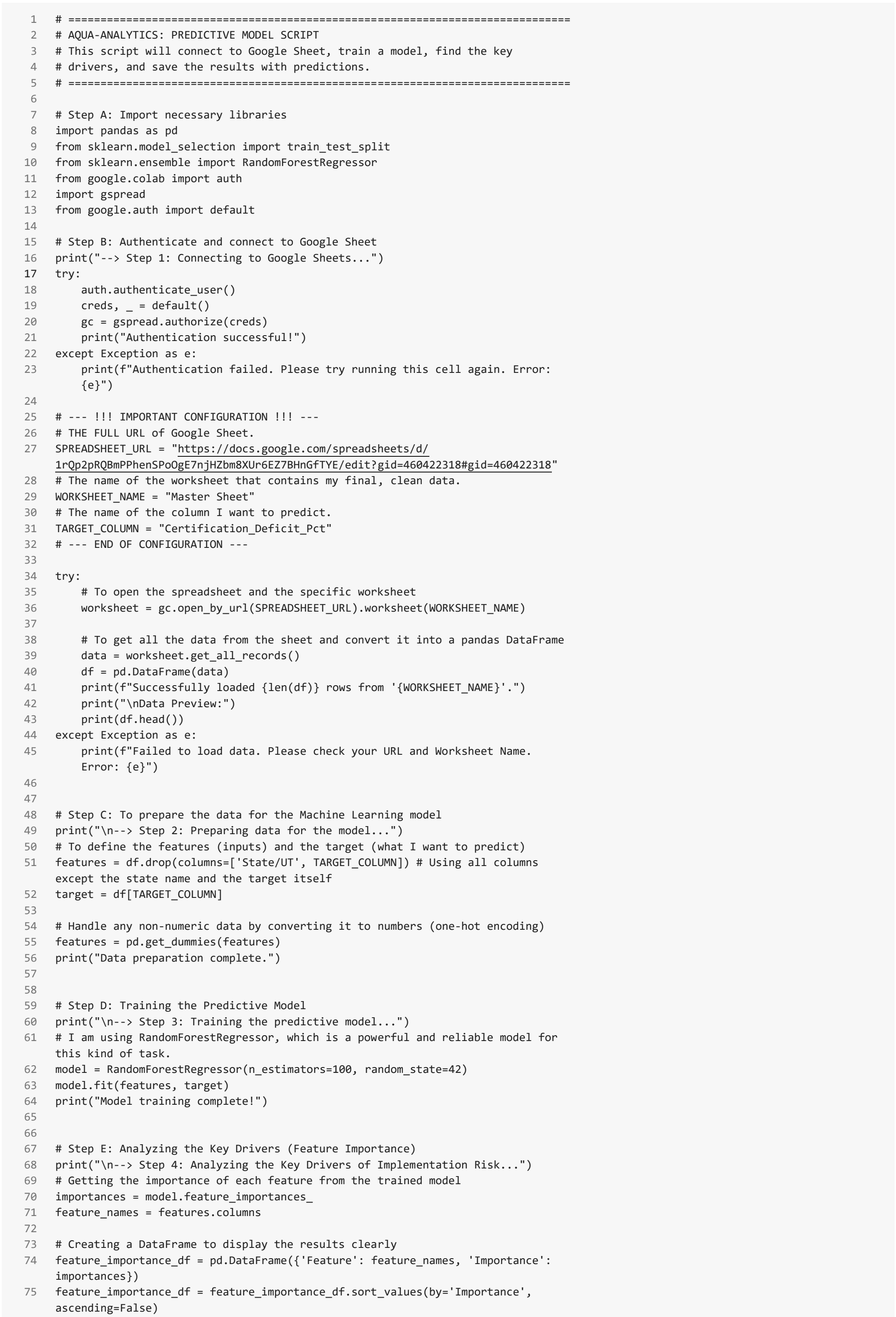
Village Inequality Index Importance: 18.9% While still a relevant factor, the disparity in coverage between villages is less predictive than the state's overall human capital and its demonstrated operational success.

PWS Infrastructure Gap Importance: 17.5% The lack of foundational Piped Water Supply is a contributing factor to the problem, but it is the least predictive of the four.

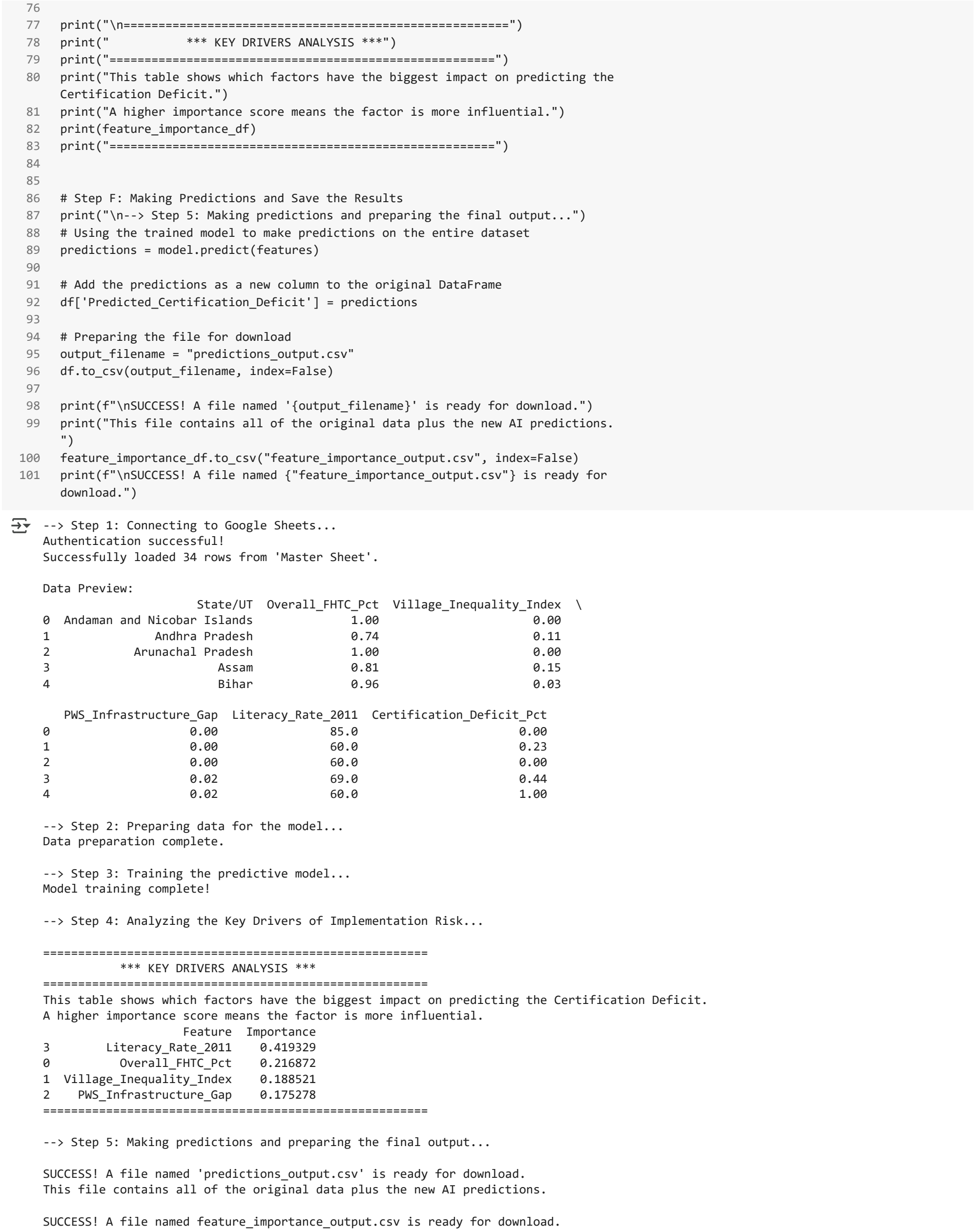
Key Takeaway: *Our project's core insight is that for future government projects to succeed, investment in socio-economic development and administrative efficiency is the most critical factor for ensuring true, verifiable impact.*

Drivers of Implementation Risk





```python
# =============================================================================
# AQUA-ANALYTICS: PREDICTIVE MODEL SCRIPT
# This script will connect to Google Sheet, train a model, find the key
# drivers, and save the results with predictions.
# =============================================================================

# Step A: Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from google.colab import auth
import gspread
from google.auth import default

# Step B: Authenticate and connect to Google Sheet
print("--> Step 1: Connecting to Google Sheets...")
try:
    auth.authenticate_user()
    creds, _ = default()
    gc = gspread.authorize(creds)
    print("Authentication successful!")
except Exception as e:
    print(f"Authentication failed. Please try running this cell again. Error: {e}")

# --- !!! IMPORTANT CONFIGURATION !!! ---
# THE FULL URL of Google Sheet.
SPREADSHEET_URL = "https://docs.google.com/spreadsheets/d/1rQp2pRQBmPPhenSPoOgE7njHZbm8XUr6EZ7BHnGfTYE/edit?gid=460422318#gid=460422318"
# The name of the worksheet that contains my final, clean data.
WORKSHEET_NAME = "Master Sheet"
# The name of the column I want to predict.
TARGET_COLUMN = "Certification_Deficit_Pct"
# --- END OF CONFIGURATION ---

try:
    # To open the spreadsheet and the specific worksheet
    worksheet = gc.open_by_url(SPREADSHEET_URL).worksheet(WORKSHEET_NAME)

    # To get all the data from the sheet and convert it into a pandas DataFrame
    data = worksheet.get_all_records()
    df = pd.DataFrame(data)
    print(f"Successfully loaded {len(df)} rows from '{WORKSHEET_NAME}'.")
    print("\nData Preview:")
    print(df.head())
except Exception as e:
    print(f"Failed to load data. Please check your URL and Worksheet Name. Error: {e}")


# Step C: To prepare the data for the Machine Learning model
print("\n--> Step 2: Preparing data for the model...")
# To define the features (inputs) and the target (what I want to predict)
features = df.drop(columns=['State/UT', TARGET_COLUMN]) # Using all columns except the state name and the target itself
target = df[TARGET_COLUMN]

# Handle any non-numeric data by converting it to numbers (one-hot encoding)
features = pd.get_dummies(features)
print("Data preparation complete.")


# Step D: Training the Predictive Model
print("\n--> Step 3: Training the predictive model...")
# I am using RandomForestRegressor, which is a powerful and reliable model for this kind of task.
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(features, target)
print("Model training complete!")


# Step E: Analyzing the Key Drivers (Feature Importance)
print("\n--> Step 4: Analyzing the Key Drivers of Implementation Risk...")
# Getting the importance of each feature from the trained model
importances = model.feature_importances_
feature_names = features.columns

# Creating a DataFrame to display the results clearly
feature_importance_df = pd.DataFrame({'Feature': feature_names, 'Importance': importances})
feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=False)
```



```python

print("\n=======================================================")
print("           *** KEY DRIVERS ANALYSIS ***")
print("=======================================================")
print("This table shows which factors have the biggest impact on predicting the Certification Deficit.")
print("A higher importance score means the factor is more influential.")
print(feature_importance_df)
print("=======================================================")


# Step F: Making Predictions and Save the Results
print("\n--> Step 5: Making predictions and preparing the final output...")
# Using the trained model to make predictions on the entire dataset
predictions = model.predict(features)

# Add the predictions as a new column to the original DataFrame
df['Predicted_Certification_Deficit'] = predictions

# Preparing the file for download
output_filename = "predictions_output.csv"
df.to_csv(output_filename, index=False)

print(f"\nSUCCESS! A file named '{output_filename}' is ready for download.")
print("This file contains all of the original data plus the new AI predictions.")
feature_importance_df.to_csv("feature_importance_output.csv", index=False)
print(f"\nSUCCESS! A file named {"feature_importance_output.csv"} is ready for download.")
```

```
--> Step 1: Connecting to Google Sheets...
Authentication successful!
Successfully loaded 34 rows from 'Master Sheet'.

Data Preview:
                     State/UT  Overall_FHTC_Pct  Village_Inequality_Index  \
0  Andaman and Nicobar Islands              1.00                      0.00
1               Andhra Pradesh              0.74                      0.11
2            Arunachal Pradesh              1.00                      0.00
3                        Assam              0.81                      0.15
4                        Bihar              0.96                      0.03

   PWS_Infrastructure_Gap  Literacy_Rate_2011  Certification_Deficit_Pct
0                    0.00                85.0                       0.00
1                    0.00                60.0                       0.23
2                    0.00                60.0                       0.00
3                    0.02                69.0                       0.44
4                    0.02                60.0                       1.00

--> Step 2: Preparing data for the model...
Data preparation complete.

--> Step 3: Training the predictive model...
Model training complete!

--> Step 4: Analyzing the Key Drivers of Implementation Risk...

=======================================================
           *** KEY DRIVERS ANALYSIS ***
=======================================================
This table shows which factors have the biggest impact on predicting the Certification Deficit.
A higher importance score means the factor is more influential.
                   Feature  Importance
3       Literacy_Rate_2011    0.419329
0         Overall_FHTC_Pct    0.216872
1  Village_Inequality_Index    0.188521
2    PWS_Infrastructure_Gap    0.175278
=======================================================

--> Step 5: Making predictions and preparing the final output...

SUCCESS! A file named 'predictions_output.csv' is ready for download.
This file contains all of the original data plus the new AI predictions.

SUCCESS! A file named feature_importance_output.csv is ready for download.
```