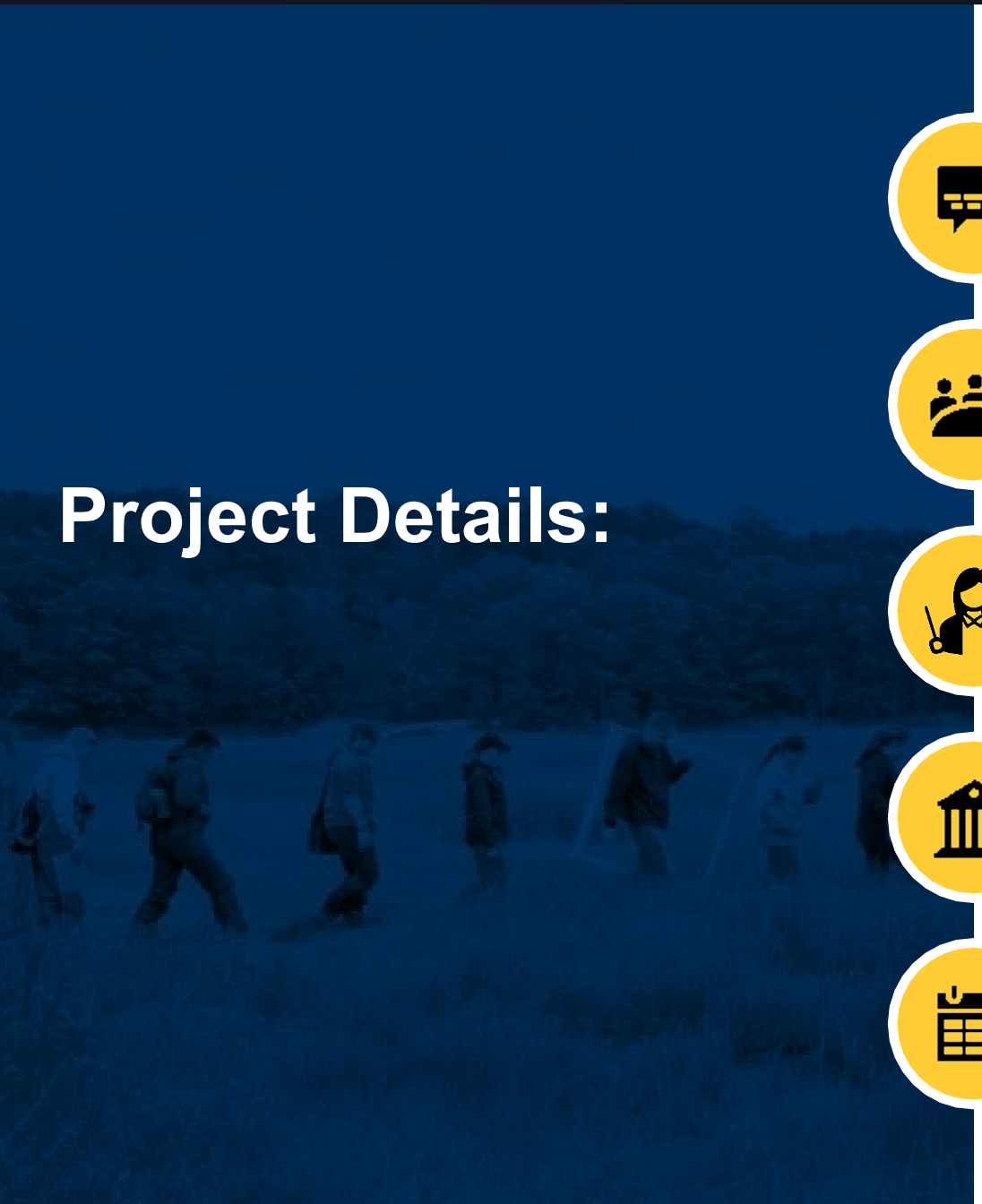
University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

EDU PREDICT TOOL
PREDICTING ENROLLMENT TRENDS IN HIGHER EDUCATION IN THE USA



Project Details:

Project Title: EDU PREDICT TOOL

Project Team: 01

Project Advisor: Dr. Ardiana Sula

MSDS Capstone – SP 25 DSCI 6051-07

Date of Presentation: 04-02-2025

EduPredict – Forecasting Higher Education Enrollment Trends

- **Goal:** Analyze and predict future enrollment patterns for international students in U.S. universities using data-driven techniques.
- **Outcome:** A **Power BI-based interactive dashboard** that provides enrollment trend forecasts under different scenarios (**Baseline, Growth, and Decline**).
- **Core Features:**
  - Machine learning-driven insights for **strategic academic planning**.
  - Customizable filters for **region, study level, and time period** to explore data dynamically.
- **Significance:**
  - Equips university officials with insights to **optimize resource planning and policy decisions**.
  - Supports institutions in **adapting to demographic and economic shifts** in student enrollment.
  - Understanding how technological shift effects the field of education

TEAM - 01

- Koteswar Enamadni

- Ifra Naaz Mohammed

- Krishnaveni Peesapati

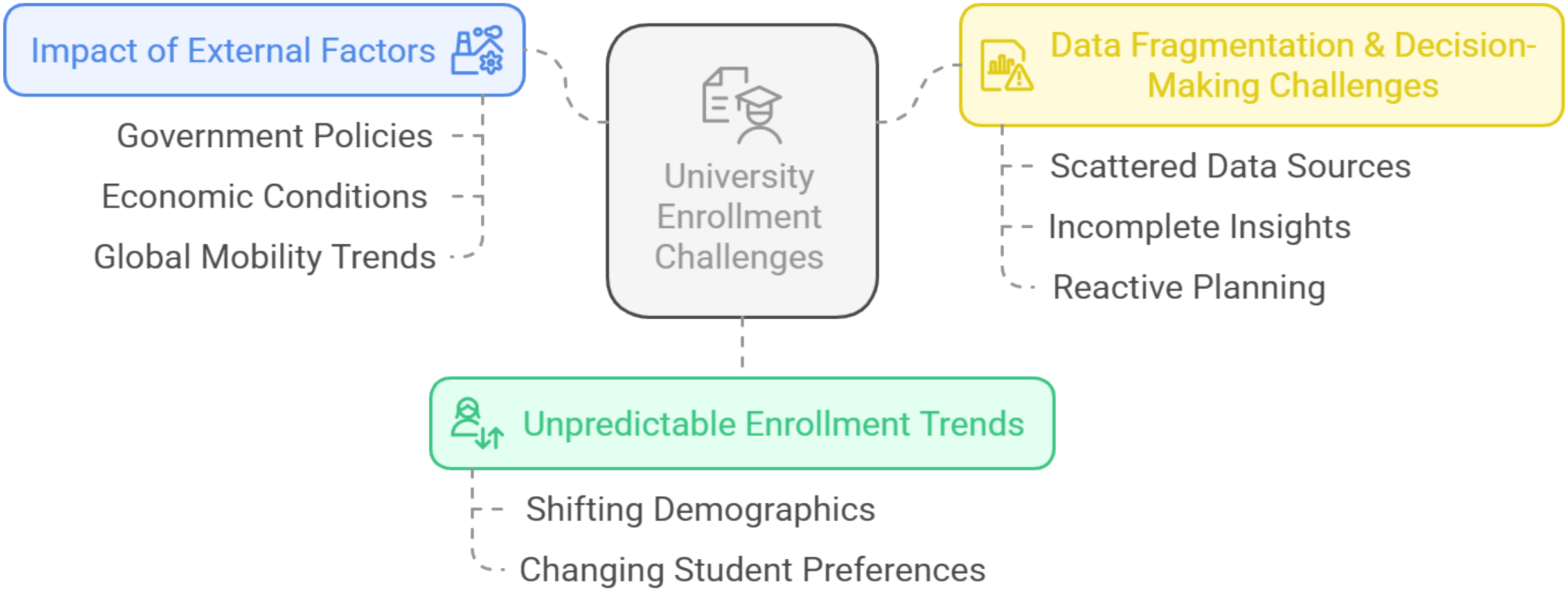
- Gnaneswari Vaddepalli

- Chethan Chakradhar M

- Karthik Vinnakota

The Problem

University Enrollment Challenges: Trends, Factors & Data



Project Goals & Objectives

Primary Goal:
Develop a predictive model that estimates future student enrollments based on multiple factors.

Data Collection & Sources : Dataset_link

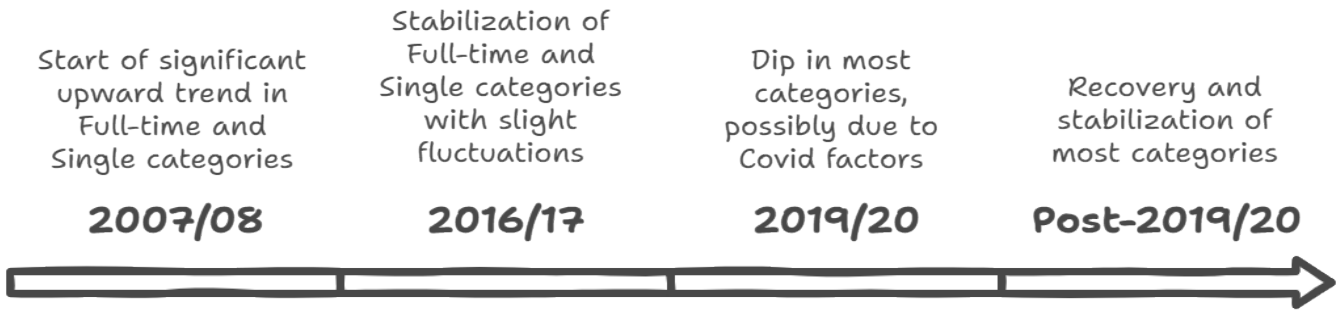
- Dataset Breakdown:
- Student Enrollment Trends (By year, region, academic level).
- Demographics (Gender, marital status, visa type).
- Funding Sources (Self-funded, government-sponsored, institution grants).

Key Objectives:

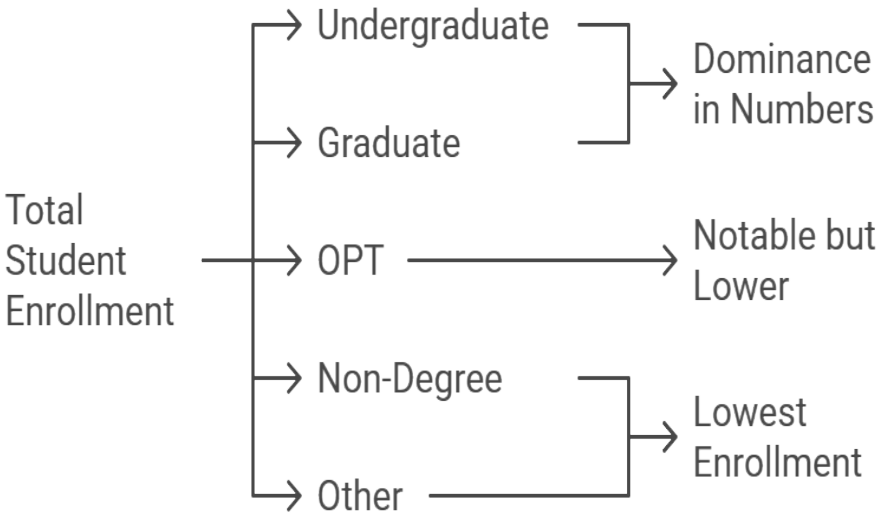
- Integrate historical datasets on enrollment, demographics, and funding sources.
- Identify patterns & trends in student admissions across various academic levels.
- Build a forecasting model using machine learning.
- Deploy results via a user-friendly Power BI dashboard for dynamic analysis.

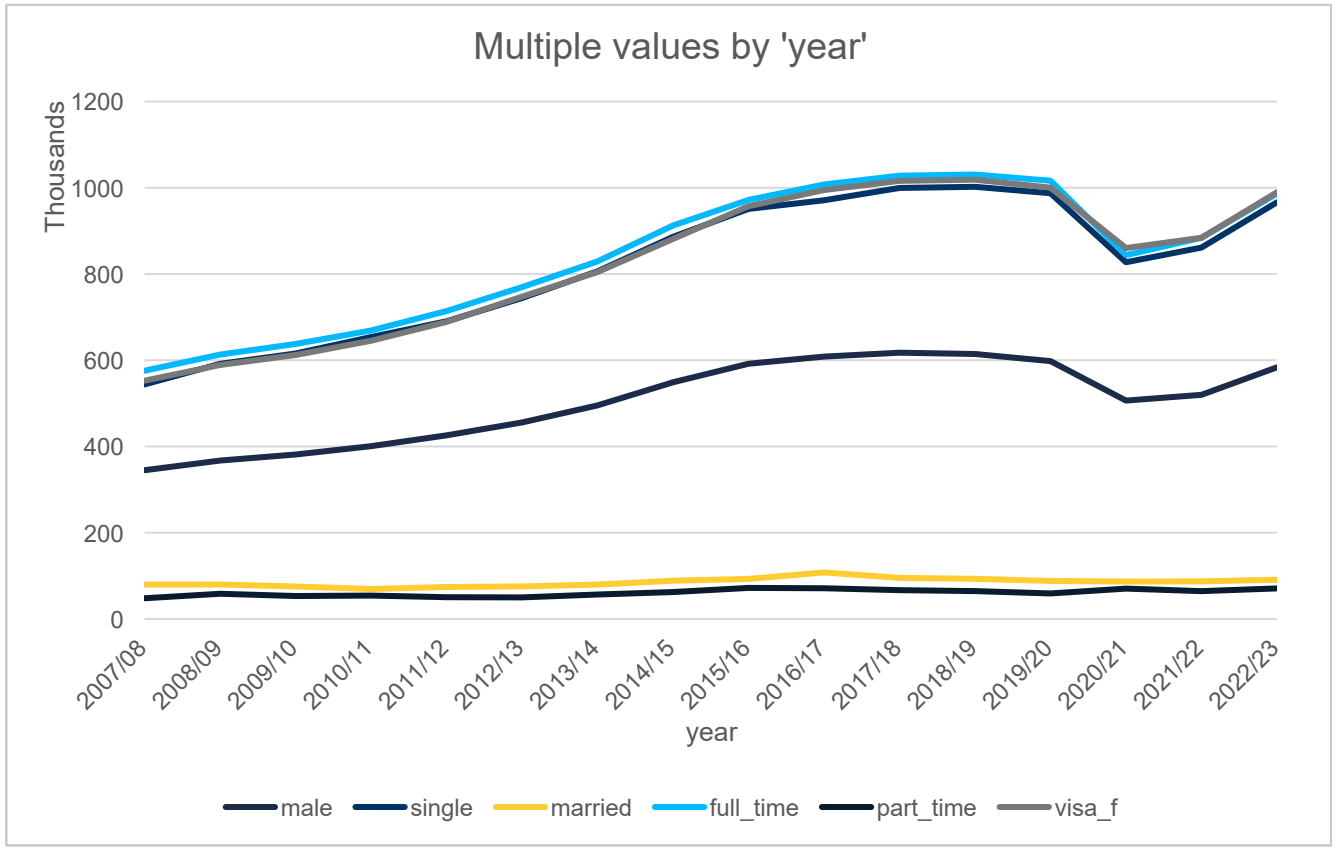
Key EDA Visualizations

Trends in Demographic Categories Over Time



Student Enrollment Trends and Patterns





Trends in Employment

Full-time Growth

Significant upward trend observed from 2007/08 to 2016/17.

Male Growth

Slower upward trend compared to full-time and single categories.

Married Stability

Remains relatively flat throughout the observed period.

Observations Summary

Highlights notable growth and external factors affecting trends.

Single Growth

Notable increase from 2007/08 to 2016/17, stabilizing afterward.

Part-time Growth

Shows an upward trend but at a slower rate.

Visa_f Increase

Slight increase over the years, lowest among all categories.

Handling Missing Data

Title: Ensuring Data Accuracy

- **Challenges:**
  - Some datasets had missing values, particularly in **older academic years**.
- **Solutions Implemented:**
  - **2007 <= data: selected data from and after 2007 to ensure connectivity in data.**
  - **Feature engineering: Cleaned all the data types and created new columns to include in dashboard.**
- **Outcome:**
  - A well-structured, **high-quality dataset** for training predictive models.

Business Questions for Analysis

Student Demographics & Enrollment

- What is the distribution of students by academic type (Undergraduate, Graduate, Non-Degree, OPT)?
- How has the number of international vs. U.S. students changed over the years?
- What are the trends in student enrollment (full-time vs. part-time, marital status, gender distribution)?

Financial Insights

- What are the primary sources of funding for international students?
- How does funding distribution vary across academic types?

Regional & Visa Analysis

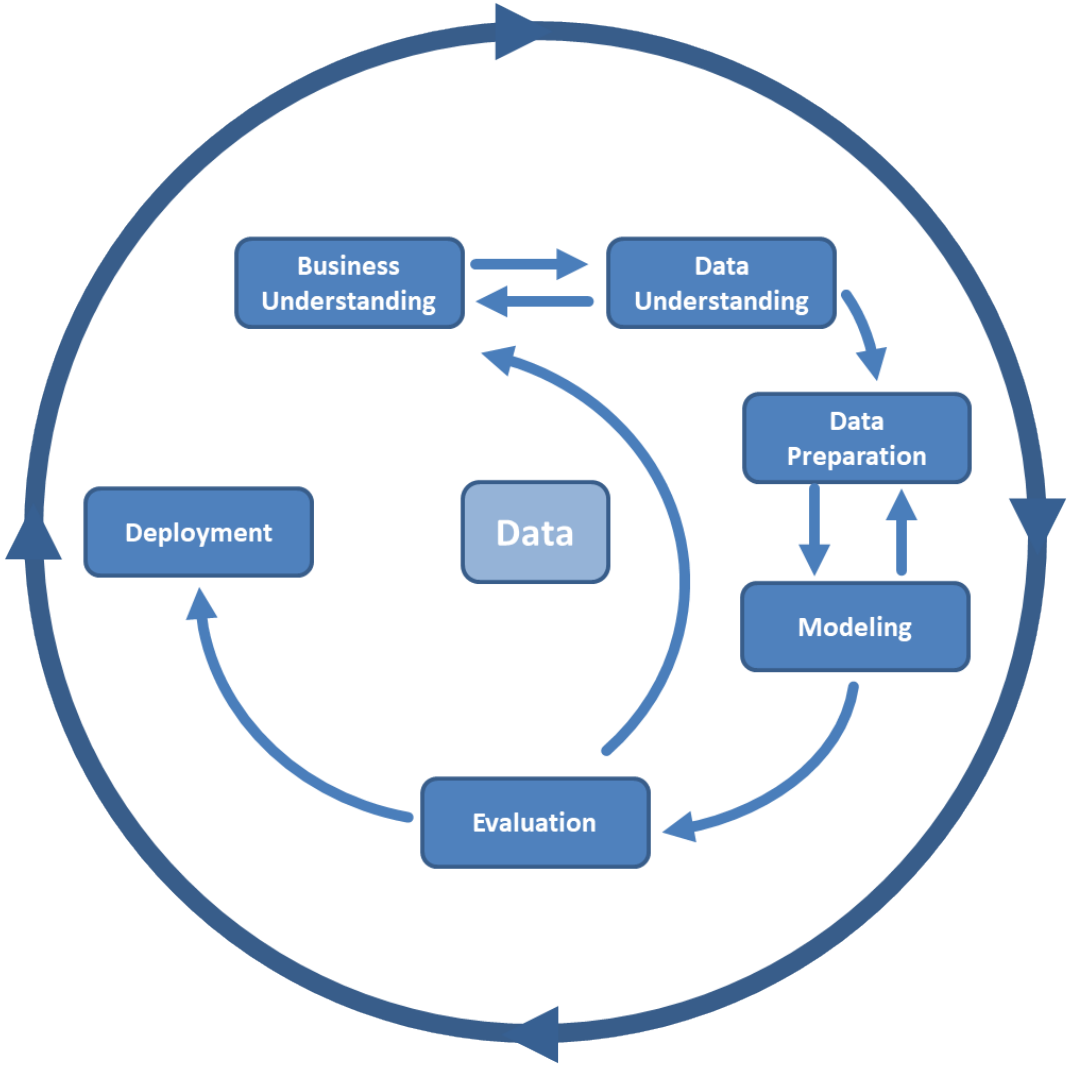
- Which regions and countries contribute the most international students?
- What are the trends in visa types (F, J, and other categories) over the years?

Field of Study Trends

- Which fields of study and majors have the highest enrollment?
- How has the popularity of different majors changed over time?

CRISP DM Methodology

- **Data-Centric Approach:** Focuses on data understanding and preparation to ensure robust model building.
- **Flexibility:** Its iterative nature allows for updates and refinements with changing data or goals.
- **Stakeholder Focus:** Aligns insights with business objectives for actionable decision-making.
- **Broad Applicability:** Suitable for diverse industries, including forecasting international student demographics.



Model Comparisions

Random Forest:

- We used it as an initial benchmark.
- Accurately predicted past data but **lacked forecasting capability**.
- Served as a reference for evaluating other models.
  - MAE: 6259.04
  - R^2 Score: 0.8303

Polynomial Regression: Regression-based approach, but **prone to high variance**.

- Training MAE: 41,133.15
- Testing MAE: 88,058.07
- Training R^2 Score: 0.8190
- Testing R^2 Score: 0.5114

ARIMA: A time-series model that **effectively captures trends** and long-term patterns.

- In-Sample MAE: 38,395.77
- In-Sample R^2 Score: 0.7513

ARIMA Model

Why ARIMA?

- Specifically designed for time-series forecasting.
- Captures long-term enrollment trends effectively.
- More stable than regression-based models for forecasting.

Strengths of ARIMA:

- Identifies trends without overfitting.
- Provides reliable enrollment projections based on past patterns.
- More interpretable than complex machine learning models for stakeholders.

Final Decision: ARIMA was selected as the best model for student enrollment forecasting due to its ability to provide stable, interpretable, and data-driven predictions over time.

ARIMA Modeling & Forecasting

- **Parameter Selection**
  Selected (p, d, q) values using ACF, PACF, and grid search.
- **Model Fitting & Refinement**
  Trained ARIMA models on historical data with iterative improvements.
  Ensured residuals had no autocorrelation, indicating a well-fitted model.
- **Forecasting & Confidence Intervals**
  Generated forecasts for each dataset column.
  Included confidence intervals to quantify prediction uncertainty.

Evaluation & Insights

- **Model Performance Assessment**
  Evaluated accuracy using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).
  Compared performance across datasets to validate robustness.

Baseline & Progress [03-26-2025 to 04-02-2025]

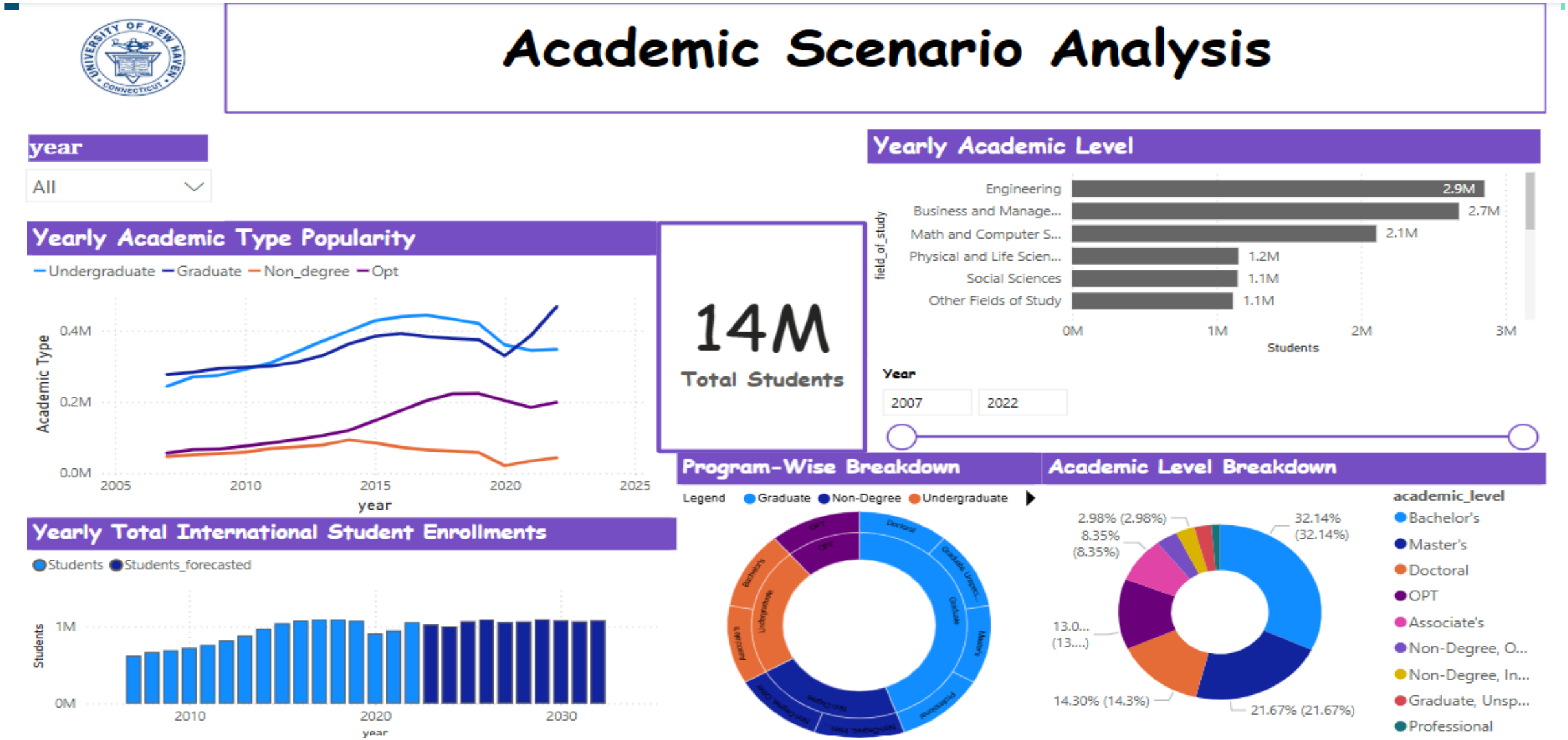
Progress achieved so far:

- **Business Insights & Dashboard Design:** Created intuitive visualizations, calculated essential metrics, and organized the dashboard into main sections—**Demographics, Program Insights, Funding, and Visa Trends**—to facilitate informed decision-making.
- **Predictive Modeling & Trend Analysis:** Time-based features were engineered, and datasets were prepared for forecasting future student enrollments. ARIMA, Linear Regression, and Polynomial Regression models were applied to analyze historical trends and generate predictions for future enrollments.
- **Dashboard Finalization and Reporting:** The dashboard has been fully created, showcasing key insights across demographics, program details, funding sources, and visa trends. Preparation for the final report is underway to document findings comprehensively.
- **Designed interactive dashboards** to visualize enrollment trends.

Bi-Weekly Key Milestones Report

- **Week 4 – 6 Progress:**
- To enhance model accuracy, we employed techniques like **feature engineering**, creating time-based features. Multiple models, including ARIMA, Polynomial Regression and Linear Regression, were applied and compared for optimal performance.
  - Explored **alternative models** for better forecasting.
  - Integrated insights into **Power BI dashboard**.
- **Next Steps:**
- **Dashboard Optimization**: Enhance interactivity and visualization by refining the layout and improving UI/UX for clarity and a better user experience.
- **Insights and Reporting**: Document insights for the final project submission while experimenting with alternative forecasting models to enhance accuracy.

Dashboard Analysis



Koteswar's Contributions

- **Data Cleaning and Transformation**: Loads, cleans, and prepares enrollment data (including creating a 'total' enrollment column and handling datetime).
- **Data Validation**: Removed duplicate records and cross-verified totals to maintain data accuracy and reliability for analysis.
- **Feature Engineering**: Modified and enhanced features to improve the dataset's quality and relevance, optimizing it for more effective analysis and modeling.
- **Dashboard Design Contribution**: Created a rough hand-drawn layout to conceptualize the dashboard's structure and functionality, facilitating initial design visualization and team collaboration.

Koteswar's Contributions

- **Data Handling:** Loads, cleans, and prepares enrollment data (including creating a 'total' enrollment column and handling datetime).
- **Stationarity:** Analyzes data stationarity using the Dickey-Fuller test and applies differencing to achieve stationarity if needed.
- **Model Selection:** Employs ACF/PACF analysis to guide ARIMA model order selection and uses RMSE to identify the best-performing model.
- **Forecasting:** Forecasts future total enrollment using the chosen ARIMA model, producing predictions with confidence intervals for a 10-year horizon.
- **Multi-Variable Forecasting & Consolidation:** Forecasts various enrollment categories individually, combines them into a comprehensive table, and visualizes the results.

Ifra's Contributions

Data Forecasting & Integration:

- Forecasted all columns in the *academic.csv* file using time series models.
- Integrated forecasted data with *academic details* and *field of study* from historical records.
- Ensured consistency and accuracy in data mapping across multiple datasets.

ARIMA Model Development:

- Implemented ARIMA-based time series forecasting in Jupyter Notebook (Colab).
- Analyzed historical student enrollment trends and projected future values.
- Optimized model parameters to enhance forecasting accuracy.

Dashboard Development & Insights:

- Designed interactive Power BI dashboard incorporating forecasted trends.
- Visualized Gender Ratio, Visa Type Distribution, and Enrollment Type Trends.
- Developed a year-over-year breakdown for full-time vs. part-time student enrollment.

Gnaneswari's Contributions

Dataset Cleaning & Preprocessing: Processed the source_of_fund.csv dataset by handling missing values, structuring the data, and transforming necessary columns for analysis. Grouped data by year to calculate total student enrollments for trend analysis.

Model Development & Training: Built a Linear Regression Model to predict student enrollments for future years based on historical trends. Extracted numeric year values for regression and trained the model using past data.

Model Evaluation & Performance Metrics: Evaluated model accuracy using R^2 Score and RMSE (Root Mean Squared Error) to measure prediction reliability. Ensured no negative predictions by adjusting forecasted values.

Predictions & Future Analysis: Generated student enrollment projections from 2022 to 2032 and saved them in CSV format for further analysis. Created a combined dataset with both historical and predicted enrollments for reference.

Visualization & Reporting:

- Plotted historical vs. predicted enrollment trends using Matplotlib for clear insights.
- Exported the visualization and prediction results as CSV files for integration into Power BI.
- Designed a Power BI Dashboard to present key findings interactively.

Chetan's Contributions

Dataset Handling:

- Loaded dataset 'cleaned_origin_data.csv' for trend analysis.
- Used pandas for efficient data manipulation.

Data Type Conversion:

- Ensured 'year' column was numeric.
- Handled errors during conversion.

Data Organization:

- Grouped data by 'origin' for trend analysis.
- Aggregated students per year for structured time-series data.

Validation:

- Checked for missing values and anomalies.
- Skipped origins with insufficient data points.

Storage Optimization:

- Created directory 'origin_forecast_plots' for results storage.

Model Selection:

- Used Linear Regression from sklearn for trend forecasting.
- Chose for simplicity, interpretability, and efficiency.

Feature Engineering:

- Converted 'year' values into numeric format.
- Prepared structured dataset for model input.

Forecasting:

- Predicted student enrollment trends for 2023-2032.
- Applied linear regression model for each origin.

Error Analysis:

- Evaluated model using mean squared error (MSE) and R^2 score.
- Analyzed accuracy of predictions.

Visualization:

- Generated and saved trend plots for actual vs. predicted enrollments.

Krishnaveni's Contributions

- **Data Processing:**
- Reads student enrollment data from academic_detail.csv.
- Extracts and formats academic years into a numeric feature.
- Aggregates total student count per year.
- **Machine Learning Modeling:**
- Uses **Linear Regression** and **Polynomial Regression** to predict future student enrollment.
- Implements sklearn pipelines to streamline preprocessing and modeling.
- Evaluates models using **Mean Squared Error (MSE)** and **R^2 Score**.
- **Visualization & Trend Analysis:**
- Plots historical student data.
- Visualizes model predictions to assess accuracy and trends.

Accuracy:

The model achieves an R^2 score of **0.7038**, indicating that it explains **70.38%** of the variance in student enrollment trends, while the remaining **29.62%** may be influenced by other factors not captured in the model.

Karthik's Contributions

1. Data Preprocessing & Cleaning:

Cleaned the cleaned_field_of_study_data.csv file by converting the year column to datetime format. Handled missing values using forward fill (ffill) for consistency. Grouped the data by field of study and year, summing student counts for accurate forecasting.

2. Time Series Forecasting with ARIMA:

Applied ARIMA models individually for each field of study using (5, 1, 0) parameters.

Split the data into:

Training set (80%) → Model fitting.

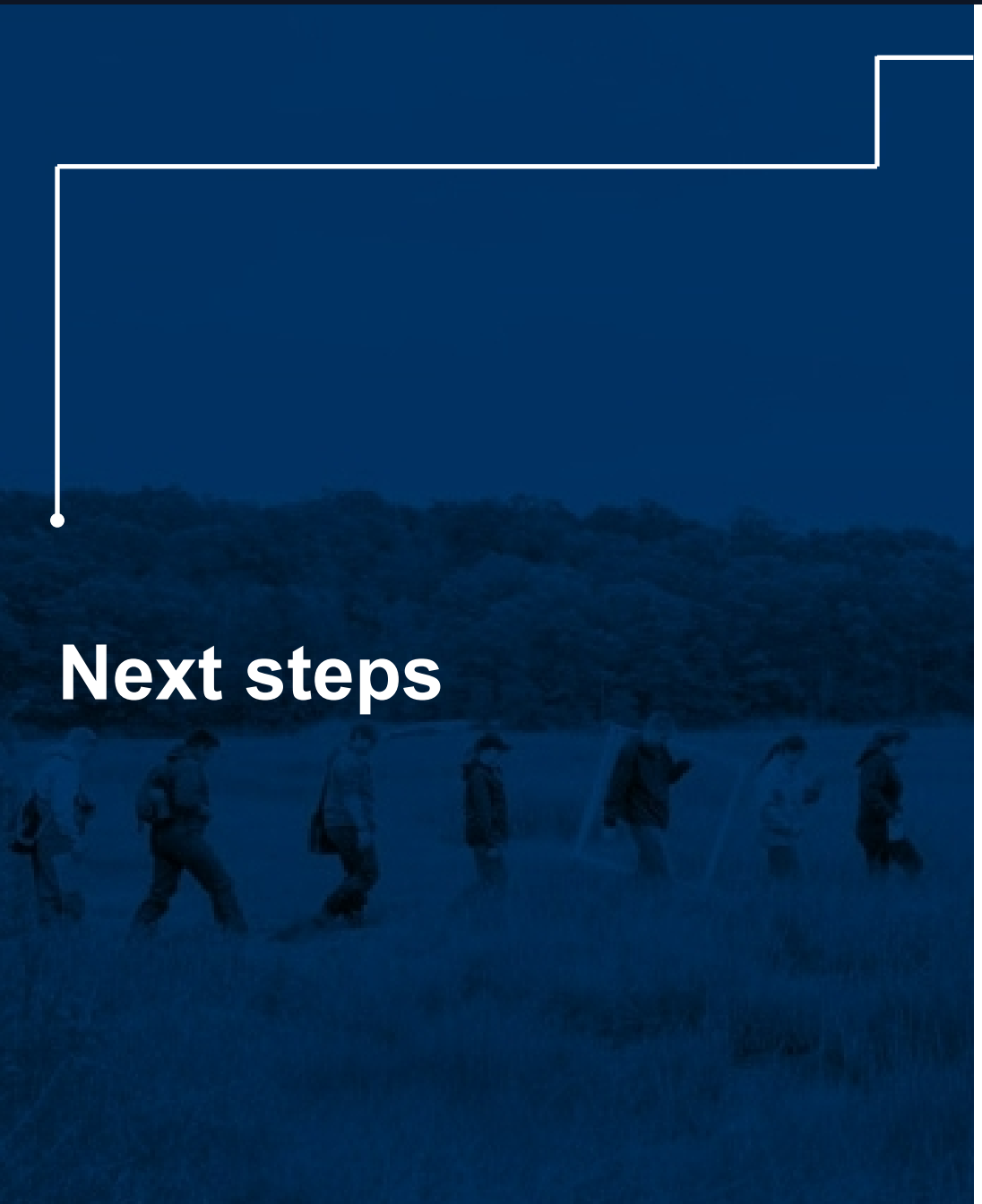
Testing set (20%) → Model evaluation.

Forecasted student enrollment for 2023-2032 (10 years).Saved the forecasted values in forecast_2023_2032.csv.

3. Model Evaluation & Accuracy:

Measured accuracy using:

MAE, RMSE, MAPE, and R^2 Score. Computed the overall accuracy for the entire model. Saved the metrics in arima_model_accuracy.csv.



Next steps

- Experiment with alternative forecasting models to **improve accuracy**.
- **Improve UI/UX** for better clarity and enhance the user experience.
- **Refine the dashboard** for interactive data visualization based on the layout.
- **Document insights** for reporting and final project submission.
- **Feedback and further insights:** Seeking insights to enhance the dashboard and usability.

THANK YOU

Please let us know if you have any further suggestions or improvements.

University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

POWER ON