US VISA APPROVAL

Machine Learning Project

Overview

This project presents a ML based solution using ML Algorithm to predict which visa applications will be approved and thus recommend a suitable profile for applicants whose visa have a high chance of approval.

Github-link

Objective:

- Develop a predictive model to determine the likelihood of a US visa application's approval.
- Provide actionable insights to visa applicants and consultants based on historical data.

Problem Statement:

- Visa approval processes can be complex and time-consuming. Understanding the factors influencing approval can help streamline preparation.

Key Technologies & Tools:

- Machine Learning: Python, Scikit-learn, XGBoost
- Data Management: MongoDB
- Cloud & DevOps: AWS (EC2, ECR), Docker, GitHub Actions
- Web Framework: Flask (for the web interface)

Data Collection & Preprocessing

Data Sources:

- **Publicly available datasets from government and immigration websites.**
- **Historical visa application data, including applicant demographics, education, job information, and visa type.**

Data Collection Methods:

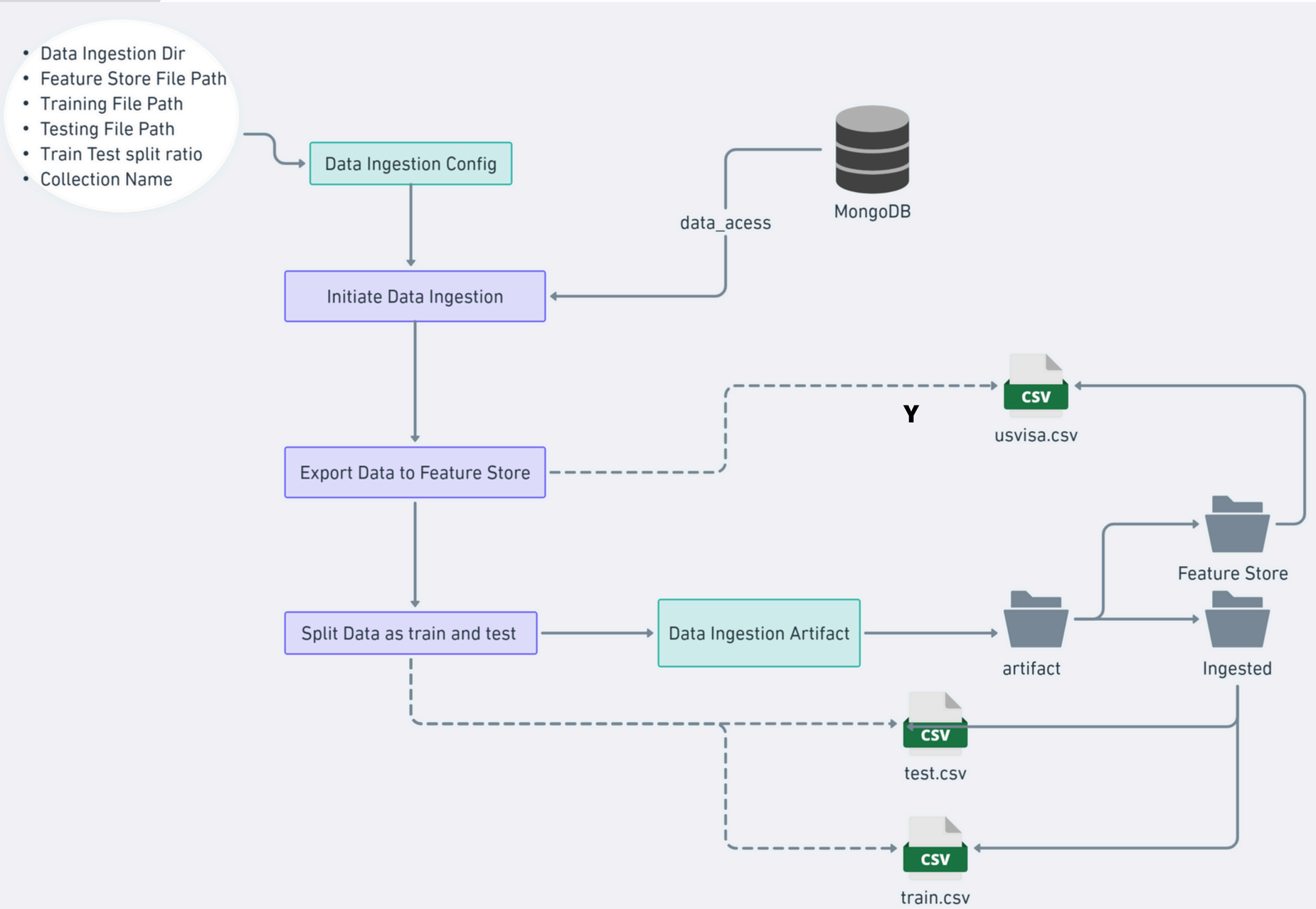
- **Web scraping and API integrations.**
- **Data ingestion pipeline using Python scripts and MongoDB for storage.**

Data Validation & Cleaning:

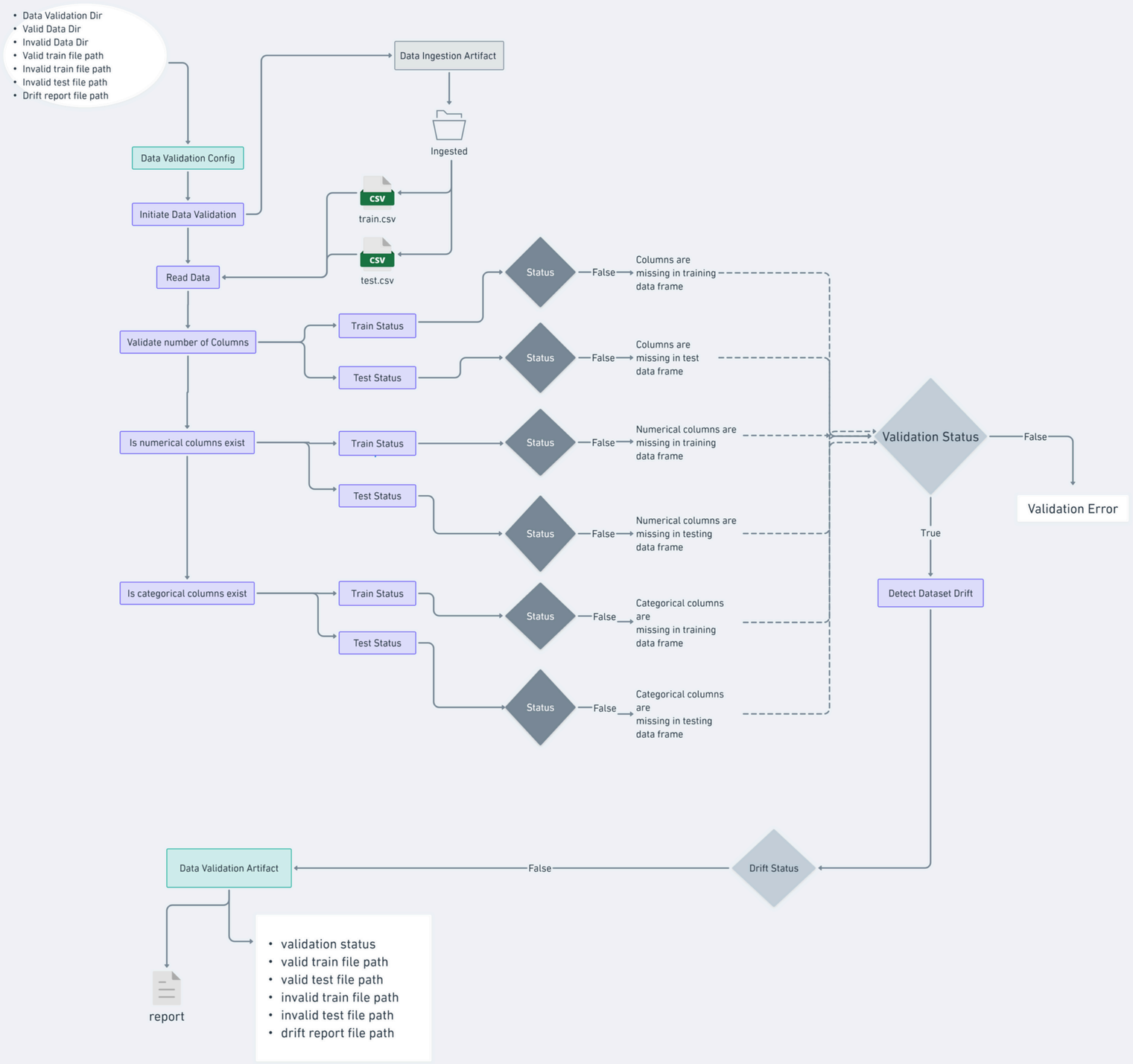
- **Removal of duplicates and null values.**
- **Standardization of data formats (e.g., dates, categorical variables).**
- **Handling of missing values through imputation or removal.**

Feature Engineering:

- **Creation of new features such as "years of experience" or "visa category."**
- **Encoding of categorical features (one-hot encoding, label encoding).**
- **Normalization and scaling of numerical features.**



- **Data is ingested by exporting visa application data from MongoDB into a DataFrame, saving it in a feature store, and splitting it into training and testing datasets for model development.**



Data is validated by checking for missing columns, ensuring schema compliance, and detecting data drift between training and testing datasets. This process ensures data integrity and consistency for model development.

Model Training and Evaluation

Model Selection:

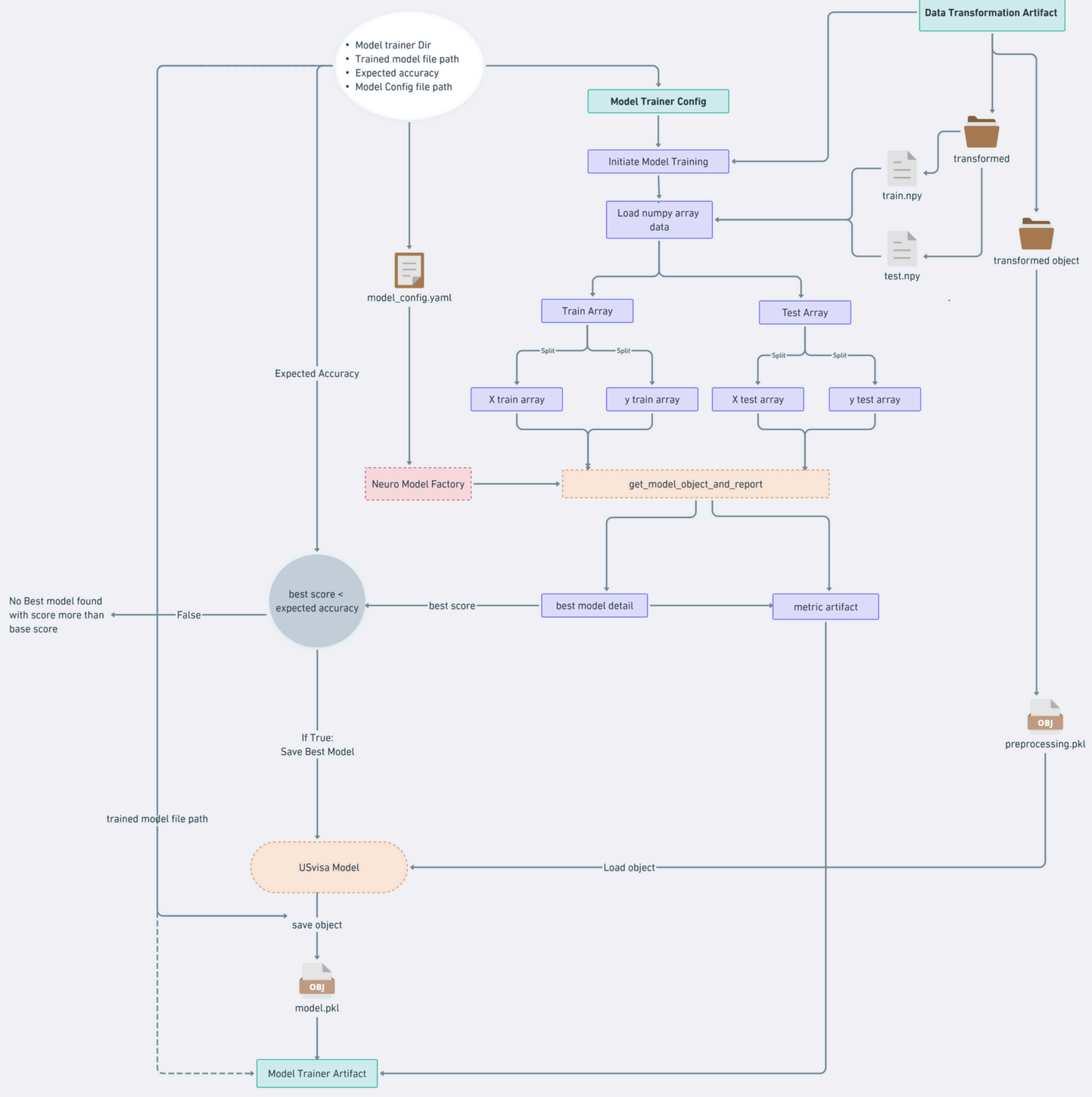
- **Initially tested multiple algorithms, including Logistic Regression, Decision Trees, Random Forest, and XGBoost.**
- **Chose the Random Forest model based on its interpretability, accuracy, and performance.**

Training and Hyperparameter Tuning:

- **Split the data into training and testing sets (80/20 split) and performed hyperparameter tuning using GridSearchCV.**
- **Employed cross-validation to ensure the model's generalizability.**

Evaluation Metrics:

- **Evaluated models using metrics like accuracy, precision, recall, and F1 score. The Random Forest model achieved the best results, with an accuracy of 85%.**



The model is trained by using neuro_mf to select the best model based on expected accuracy. The selected model is evaluated on test data using metrics like accuracy, precision, recall, and F1 score.

4o

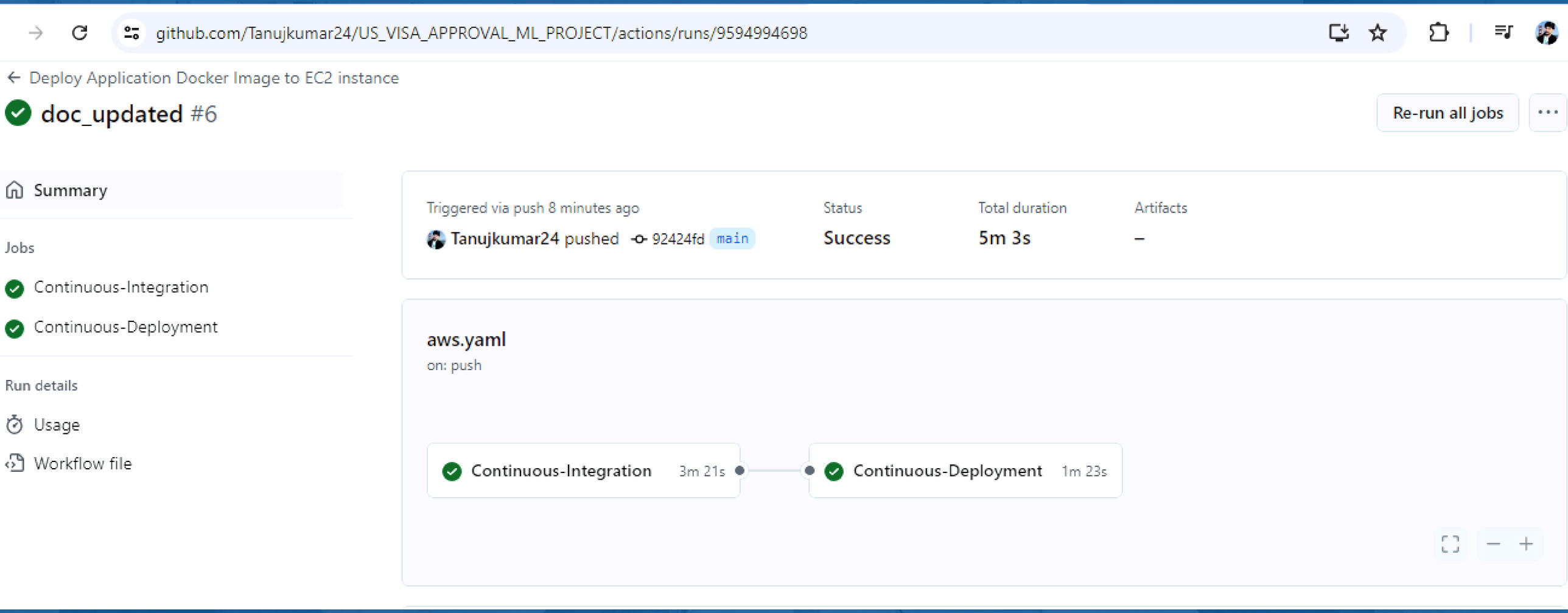
Model Deployment

Dockerization:

- **Created a Dockerfile to package the application and its dependencies into a Docker image, ensuring consistent environments across different stages (development, testing, production).**

CI/CD Pipeline:

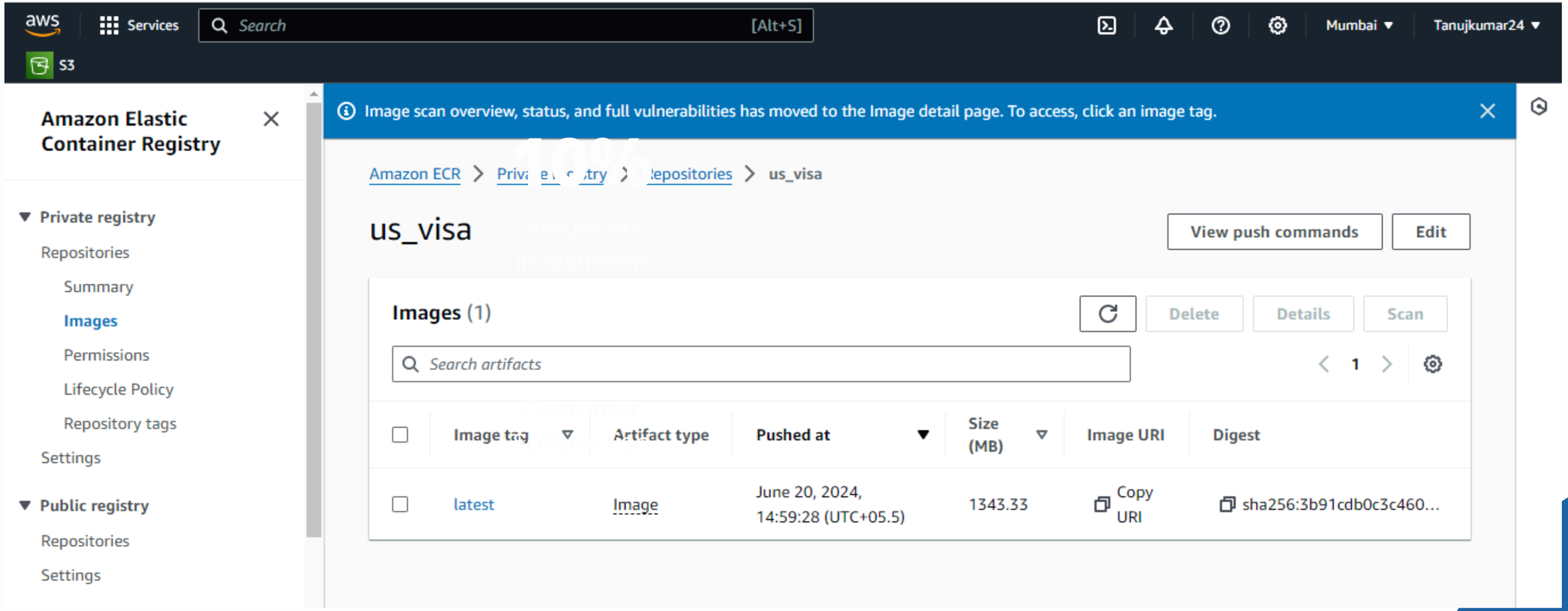
- **Set up a continuous integration and deployment (CI/CD) pipeline using GitHub Actions. This automated the testing and deployment process, ensuring that changes were quickly and reliably deployed to production.**



continuous integration and deployment (CI/CD)

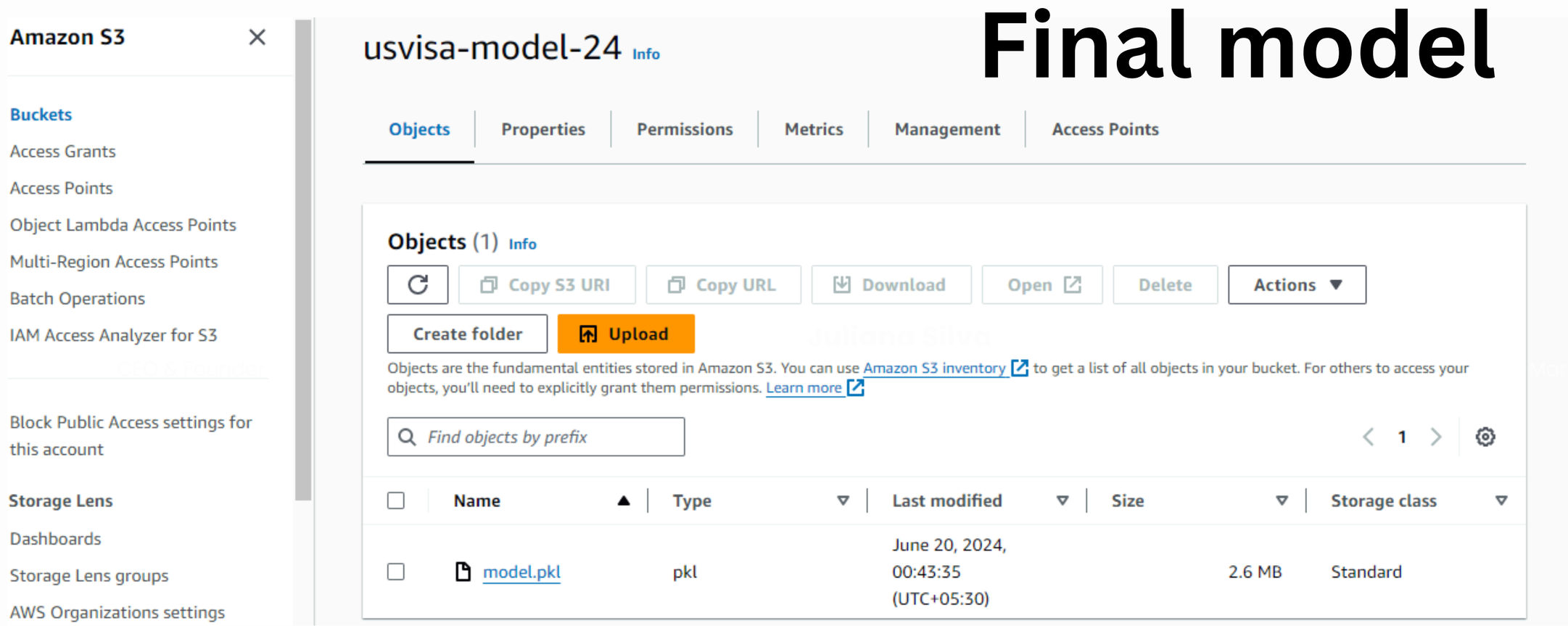
AWS Infrastructure Setup:

- **Elastic Container Registry (ECR): Used for secure storage of Docker images.**
- **EC2 Instances: Hosted the application on virtual machines.**
- **IAM Roles and Policies: Managed access and permissions for EC2 and ECR.**



Deployment Process:

1. Built the Docker image and pushed it to ECR.
2. Launched an EC2 instance and pulled the image from ECR.
3. Deployed the Docker container on the EC2 instance.



Final model

entered the values

Requires Job Training

N

Number of Employees

14500

Select Region

Northeast

Prevailing Wage

601

Contract Tenure

Hour

Full or Part Time

Y

Age of the company

15

Predict Visa-Status

Visa Prediction Status: Rendering

continent

Select Continent

Education of employee

Select Education

Has Job Experience

Has Experience

Requires Job Training

Required Training

Number of Employees

select between 15000 to 40000

Select Region

Region

Prevailing Wage

select between 700 to 7000

Contract Tenure

Contract Tenure

Full or Part Time

Full Time

Age of the company

select between 10 to 180

Predict Visa-Status

Result

Visa Prediction Status: Visa-approved