ENSIAS

IT Certifications Pipeline

Student: **Zakaria Ait Hssain** *Supervisor:* **Pr. M. Lazaar**

End of Year Project

June 2, 2025

Contents

1. Introduction

2. Data Collection (Scraping)

3. Data Cleaning

4. Machine Learning Enrichment

5. Data Analysis

6. Streamlit UI

7. Conclusion

Introduction

Problem & Objectives

Problem

- IT certifications are critical in hiring and skill demonstration.
- No unified source or pattern across providers.
- Thus: difficulty to compare and choose between certifications.

Objectives

Create a unified and accessible certification dataset to enable data-driven analysis and choice.

Data Flow CCM (Actor/Flow)

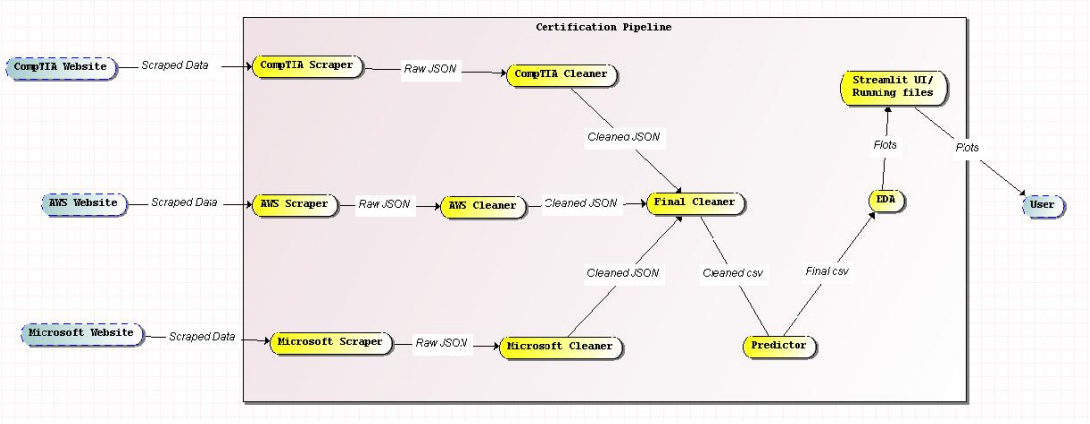
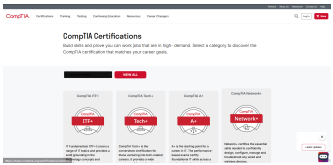


Figure: Communication Conceptual Model

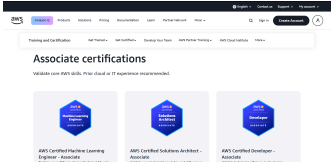
Data Collection (Scraping)

Certifications Boxes Page

CompTIA



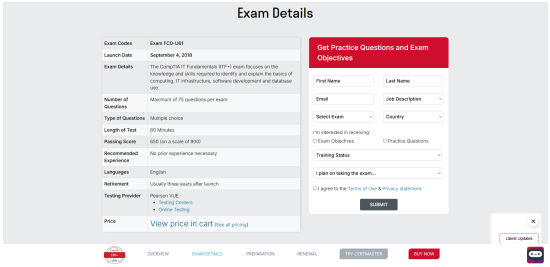
AWS



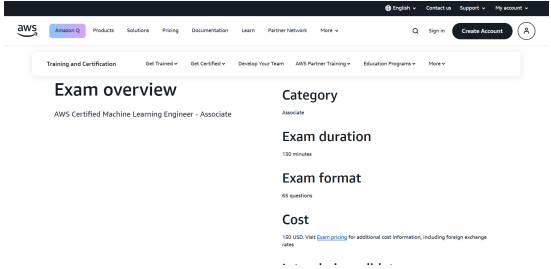
Microsoft

Certification Details Page

CompTIA



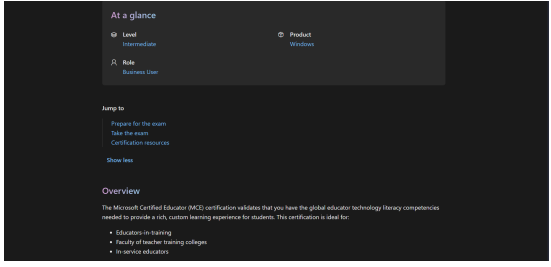
AWS



Certification Details Page

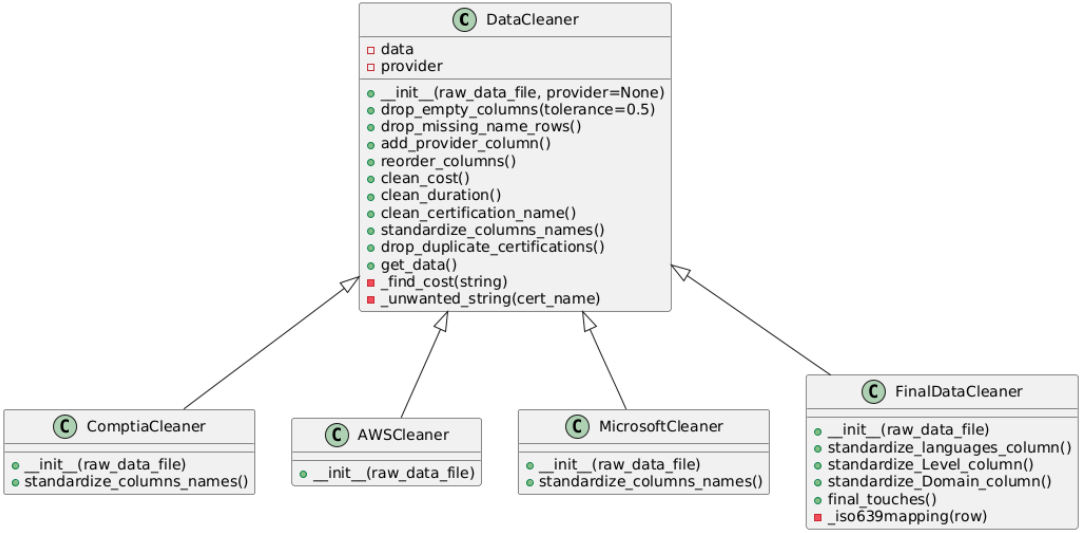
Microsoft First Layout

Microsoft Second Layout



Data Cleaning

Cleaning Classes Diagram



Machine Learning Enrichment

Machine Learning Enrichment

- Fields predicted: Level, Domain, Cost, and Duration
- Models used: Decision Trees and Random Forests (for duration)
- Performed: One-Hot encoding, Feature Selection, and Class weighting
- Dimensionality Reduction (Specifically PCA) was not necessary

Table: Best cross-validation scores for each predicted field

Target Field	Used Model	Best CV Score
Level (classifier)	Decision Tree	0.869
Domain (classifier)	Decision Tree	0.459
Cost (regressor)	Decision Tree	0.93
Duration (regressor)	Random Forest	0.837

Data Analysis

Key Insights and Analysis

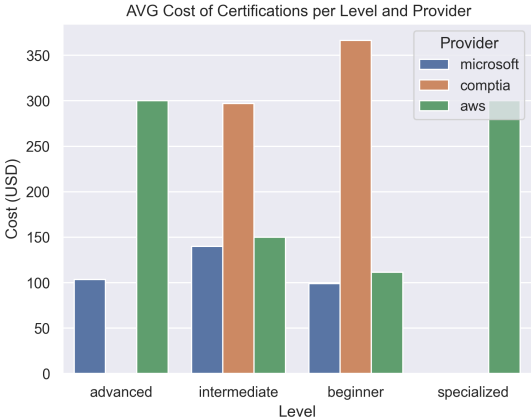


Figure: Average Cost per Level and Provider

- CompTIA certifications presents the highest average costs, surprisingly at the Beginner level.
- AWS shows a clear cost progression by level.
- Microsoft maintains relatively stable and moderate pricing.

Key Insights and Analysis

General Insights:

- Cost range: $100 – $392
- Duration modality: 60 & 100 min
- Dominant level: Intermediate
- Most common domains: Productivity, Business Apps & Data.
- Most common languages: English, Japanese & Spanish

More plots and analysis are present in the project report.

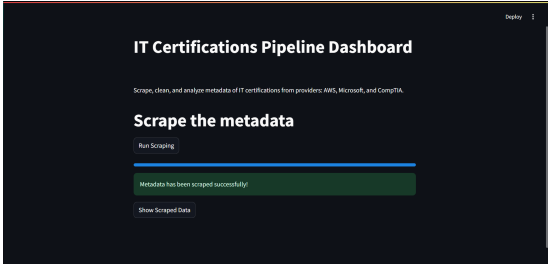
Streamlit UI

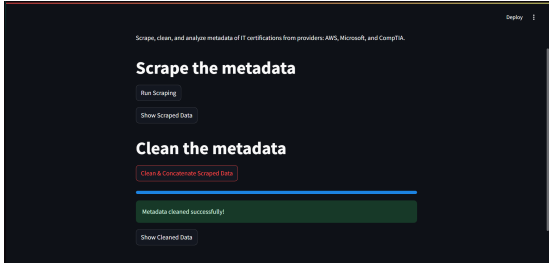
Key Features

Key features

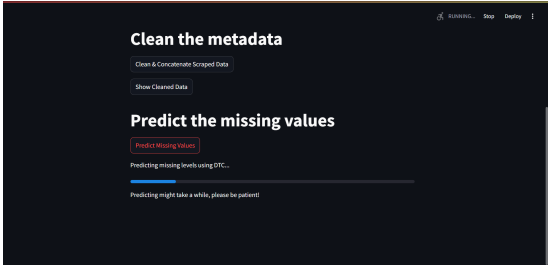
- Run full pipeline step by step.
- Track progress.
- View datasets and plots.
- Download datasets.

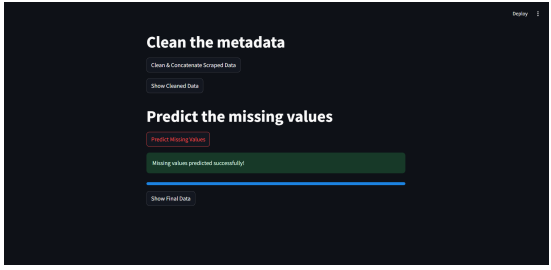
UI Overview


Deploy
IT Certifications Pipeline Dashboard
Scrape, clean, and analyze metadata of IT certifications from providers: AWS, Microsoft, and CompTIA.
Scrape the metadata
Run Scraping
Metadata has been scraped successfully!
Show Scraped Data

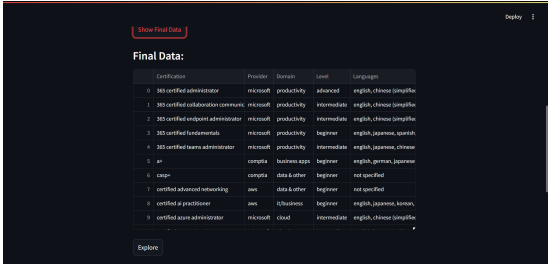

Deploy
Scrape, clean, and analyze metadata of IT certifications from providers: AWS, Microsoft, and CompTIA.
Scrape the metadata
Run Scraping
Show Scraped Data
Clean the metadata
Clean & Concatenate Scraped Data
Metadata cleaned successfully!
Show Cleaned Data

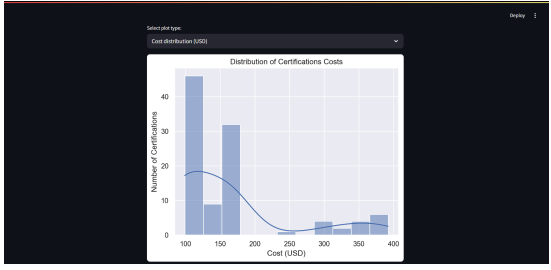
UI Overview


RUNNING...
Stop
Deploy
Clean the metadata
Clean & Concatenate Scraped Data
Show Cleaned Data
Predict the missing values
Predict Missing Values
Predicting missing levels using DTC...
Predicting might take a while, please be patient!


Deploy
Clean the metadata
Clean & Concatenate Scraped Data
Show Cleaned Data
Predict the missing values
Predict Missing Values
Missing values predicted successfully!
Show Final Data

UI Overview


Deploy
Show Final Data
Final Data:
Certification
Provider
Domain
Level
Languages
0 365 certified administrator microsoft productivity advanced english, chinese (simplifie
1 365 certified collaboration communic microsoft productivity intermediate english, chinese (simplifie
2 365 certified endpoint administrator microsoft productivity intermediate english, chinese (simplifie
3 365 certified fundamentals microsoft productivity beginner english, japanese, spanish
4 365 certified teams administrator microsoft productivity intermediate english, japanese, chinese
5 a+ comptia business apps beginner english, german, japanese
6 casp+ comptia data & other beginner not specified
7 certified advanced networking aws data & other beginner not specified
8 certified ai practitioner aws it/business beginner english, japanese, korean,
9 certified azure administrator microsoft cloud intermediate english, chinese (simplifie
Explore


Deploy
Select plot type:
Cost distribution (USD)
Distribution of Certifications Costs
Number of Certifications
0
10
20
30
40
100
150
200
250
300
350
400
Cost (USD)

Conclusion

Conclusion

This project offers

- Automated, trackable pipeline
- Unified, rich certification dataset
- Valuable tool for certification seekers

Possible project extensions

- Add more providers (e.g., Cisco, Google Cloud, Oracle)
- Job market integration (salary, demand)
- ...

Thank you!

zakaria_aithssain@um5.ac.ma

https://github.com/zakariaaithssain/it-certifications-pipeline