# Lead Deduplication & Email Validation Tool

**Author:** Yash Mane
**Submission for:** Caprae Capital – AI Readiness Challenge
**Project Duration:** ~5 hours

---

## Approach

The goal was to improve lead generation data quality by building a lightweight, easy-to-use tool that cleanses uploaded contact data by:

- Removing duplicates
- Validating email formats
- Presenting results in a streamlined, user-friendly UI

I opted for a functionally focused approach (Quality First) to build a high-impact feature within the time limit.

---

## Model Selection

This version of the tool does not use a machine learning model but leverages:

- Regex pattern matching for email validation
- Rule-based filtering for deduplication
  This choice was intentional to ensure speed, simplicity, and reliability within 5 hours.

---

## Data Preprocessing

1. Input is a user-uploaded .csv file with leads.
2. Data is read using pandas, and rows with duplicate emails are dropped.
3. Each email is tested using a regular expression:

python

CopyEdit

```
r"^[\w\.-]+@[\w\.-]+\.\w{2,4}$"
```

This checks for proper structure: username@domain.extension.

---

## Performance Evaluation

As this tool uses rule-based logic, performance is measured by:

- % of emails flagged as valid (manual spot-check for correctness)
- Clean UI/UX in previewing and downloading cleaned leads
- Successfully removing all duplicate entries by email

The tool correctly handled all edge cases in sample test data including:

- Invalid email endings (e.g., @invalid)
- Duplicate domains and names
- Empty fields

---

Instead of building a complex AI model, this tool delivers real-world impact by solving a core B2B sales pain point: bad lead data. It's scalable, easy to integrate, and requires zero training — ideal for lean teams and startups.