# Akaike's Data Science Internship Assignment

---

# Email Classification
# &
# PII Masking Tool

**Submitted By:**

**Mahammad Hishan K M**

**Applicant for Data Science Intern**

**hishankmd12@gmail.com**

**Submitted on: 19/04/2025**

**Submitted To:**

**Akaike - Talent Acquisition Team**

## 1. Introduction

In today's digital era, email remains a primary mode of communication between users and customer support teams. These emails frequently contain sensitive personal data, such as names, contact information, card details, and government-issued IDs. Improper handling of such information can result in legal violations and loss of user trust. Hence, organizations must deploy automated systems that not only protect PII but also ensure the efficient routing of messages based on their content.

This project aims to build an automated email processing pipeline that ensures secure handling of sensitive data through PII masking and leverages machine learning for intelligent email classification.

---

## 2. Problem Statement

Customer service teams often struggle with the dual challenge of maintaining data privacy and managing a high volume of incoming emails. These emails may include:

- Sensitive Personally Identifiable Information (PII) such as names, email addresses, phone numbers, and national IDs
- Financial information including credit/debit card details, CVV, and expiry dates

Moreover, classifying these emails into actionable categories (e.g., Billing, Technical Support, Account Issues) is time-consuming if done manually.

The key problems addressed in this project include:

- **PII Exposure Risk**: The unintentional retention or leakage of sensitive user data.
- **Manual Categorization**: Lack of scalable and accurate email classification.
- **Deployment Readiness**: Difficulty in integrating such systems into real-time environments.

Our solution solves these by implementing a regex-based PII masking system and a machine learning model to classify masked content efficiently.Personally Identifiable Information (PII) and Payment Card Industry (PCI) data such as full names, phone numbers, email addresses, credit/debit card numbers, Aadhar numbers, and more. Storing or processing such data without proper protection poses a significant security and compliance risk (e.g., GDPR, HIPAA).

The challenge is two-fold:

1. **Accurately detect and mask sensitive information** in a way that is non-reversible and compliant.
2. **Classify the masked email content** into predefined categories (such as Technical Support, Billing Issues, etc.) to route it to the appropriate support team.

This project addresses the problem by developing a complete pipeline for regex-based PII masking and machine learning-based email classification, deployed via a public API.

---

## 3. Approach for PII Masking and Classification

### 3.1 PII Masking (Without LLMs)

We used a **rule-based approach** utilizing regular expressions to detect and mask sensitive entities without relying on large language models. The fields masked include:

- full_name
- email
- phone_number
- dob
- aadhar_num
- credit_debit_no
- cvv_no
- expiry_no

Each detected field is replaced with a placeholder and a unique hash using hashlib.md5() to ensure consistent anonymization across the message. For example:

| |
|---|
| Rahul Sharma → [full_name]_abc123 |
| 1234 5678 9012 → [aadhar_num]_456def |

All masked fields are stored in an output structure that also includes their classification and position in the original email body.

### 3.2 Email Classification

We classified emails based on their context (even after PII is masked) into categories like:

- Technical Support
- Billing Issues
- Account Management

We vectorized the email content using TF-IDF and trained a **Logistic Regression** model to predict categories based on the masked body text.

---

## 4. Model Selection and Training Details

### 4.1 Dataset

We used the dataset combined_emails_with_natural_pii.csv which includes labeled examples of support emails.

### 4.2 Preprocessing

- PII fields were masked using regex-based detection.
- Stopwords were removed.
- Text was vectorized using TF-IDF.

### 4.3 Model Used

- **Classifier**: Logistic Regression (from sklearn.linear_model)
- **Vectorizer**: TfidfVectorizer
- Accuracy obtained: ~89% on validation set

The trained model was saved using joblib and integrated into the inference pipeline.

## 5. Challenges Faced and Solutions Implemented

### Challenge 1: Detecting All Types of PII

- Some patterns like Aadhar and card expiry dates are tricky due to varying formats.
- **Solution**: Carefully crafted regex patterns and tested them against many formats.

### Challenge 2: Model Generalization

- Initial models overfitted due to over-representation of certain categories.

- **Solution**: Balanced the dataset and tuned model hyperparameters (e.g., regularization).

**Challenge 3: API Output Format Compliance**

- Hugging Face Spaces required strict output JSON formatting.
- **Solution**: Carefully formatted the output to include:

```
{
  "input_email_body": ...,
  "list_of_masked_entities": [...],
  "masked_email": ...,
  "category_of_the_email": ...
}
```

**Challenge 4: Deployment Issues**

- Space build errors (e.g., relative import bugs, requirement issues)
- **Solution**: Adjusted import structure and ensured lightweight dependency list.

---

## 6. Conclusion

This project successfully demonstrates a practical pipeline for:

- Regex-based PII masking without LLMs
- Secure handling of PCI/PII data
- Accurate classification of email content
- Deployment-ready REST API for real-time support email processing

The solution meets the requirements for privacy, accuracy, performance, and deployability and is extensible for future enhancements.