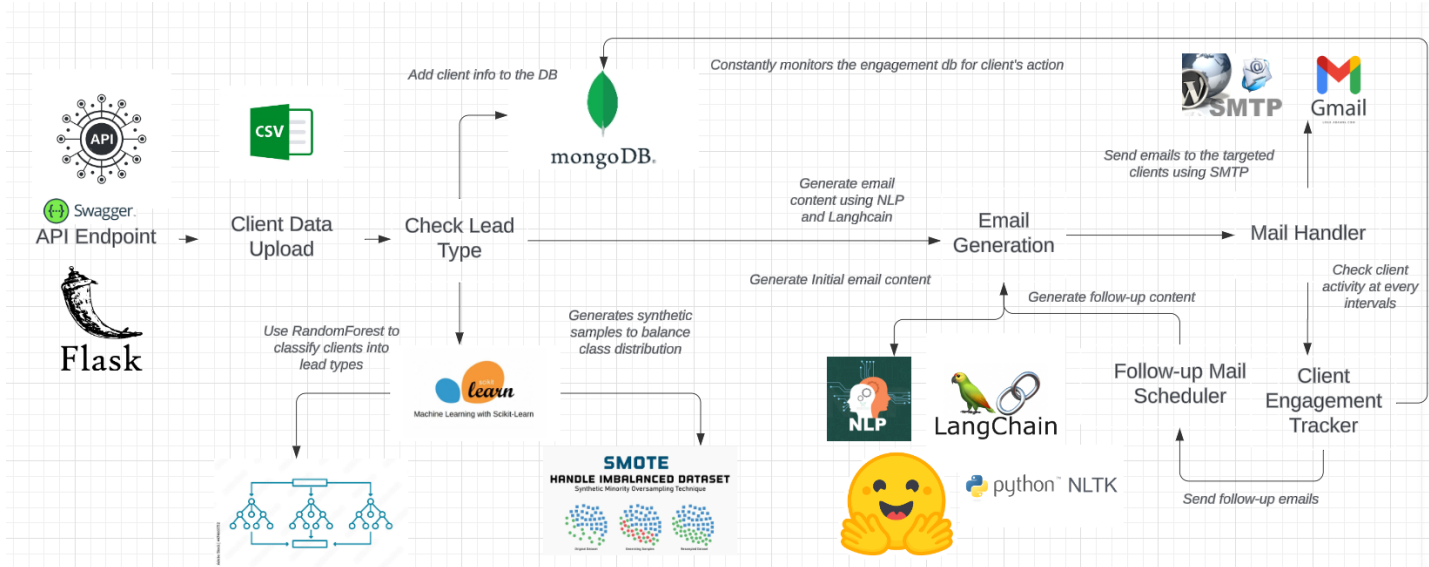


# AI- SMART LEADS

## A) Introduction



This project is an **AI-driven automation process** designed to analyze and engage potential clients based on company profiles. The process begins by feeding an Excel file into the program, which performs data analysis to identify potential clients based on company description, size, and revenue.

The analyzed data is then passed to a trained "smart leads" model to make predictions, classifying companies as hot, warm, or cold leads. Once the predictions are made, the information is logged into a MongoDB database for each client.

Next, an automation module is triggered asynchronously, allowing the main program flow to continue without interruption, enabling parallel and efficient execution. This module calls the machine learning model, which generates personalized emails based on the company's industry and specific pain points identified in the description.

Personalized emails are generated using an LLM model from Hugging Face, orchestrated through LangChain. Various React agents are incorporated to handle different tasks within the email generation process, leveraging NLP for text processing.

Emails are sent to clients via SMTP with SSL for secure communication. The automation module also tracks client engagement, such as email opens and website clicks, and sends personalized follow-up emails. Once a threshold for email frequency is reached, the job is terminated.

In summary, this system performs end-to-end automation: from lead analysis to personalized email generation and client engagement tracking, all while ensuring secure and efficient communication.

## B) Technologies

- **Flask:** Utilized for creating REST APIs to manage and interact with the application.
- **MongoDB:** Employed for storing client information and tracking engagement details.
- **Natural Language Processing (NLP):** Applied for analyzing and processing text data from client information.
- **Machine Learning:** Implemented using the Random Forest Regressor for multi-variable classification, categorizing leads into Hot, Warm, and Cold.
- **NumPy:** Used for numerical computations and handling arrays.
- **Pandas:** Utilized for processing and manipulating dataframes read from CSV files.
- **Scikit-learn:** Leveraged for importing and utilizing machine learning models.
- **Docker:** Used for containerizing the application to ensure consistency and ease of deployment.
- **Python:** The primary language used for both machine learning tasks and Flask application development.
- **SMTP:** Employed for sending emails through the SMTP protocol.
- **HTML:** Utilized for rendering and formatting HTML content in emails and web pages.
- **SSL (Secure Sockets Layer):** Implemented to secure connections and ensure data privacy during communication.
- **Hugging Face:** Used for implementing large language models (LLMs) and fine-tuning text-based tasks.
- **LangChain:** Employed for building and managing agents and chaining LLMs for advanced workflows and automation.

## C) Machine Learning Workflow

- **Data Loading:** The dataset is loaded from a CSV file using Pandas, with the encoding specified to handle character encoding issues.
- **Data Preprocessing: Missing Value Handling:** Empty values in the 'Description' column are filled with empty strings to avoid issues during processing.
- **Text Vectorization:** The TfidfVectorizer converts text data from the 'Description' column into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency), considering both unigrams and bigrams.
- **Feature Scaling:** Numerical features, such as 'Company Size' and 'Company Revenue', are scaled using StandardScaler to standardize them, improving the model's performance.
- **Feature Combination:** The processed numerical features and TF-IDF features are combined into a single feature matrix (X\_preprocessed).
- **Target Variable:** The target variable, 'Leads Score', is extracted as the output to be predicted.
- **Data Splitting:** The dataset is divided into training and test sets using train\_test\_split, ensuring that the split maintains the proportion of each class using stratify.
- **Class Imbalance Handling:** The Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data to balance the class distribution, addressing any class imbalance issues.
- **Model:** A RandomForestClassifier is initialized with class\_weight='balanced' to handle class imbalance, and random\_state=39 to ensure reproducibility.
- **Training:** The Random Forest model is trained on the balanced training data.
- **Prediction:** The trained model is used to make predictions on the test set.
- **Evaluation:** Model performance is assessed using accuracy and a classification report, which includes metrics such as precision, recall, and F1-score.

## D) Model Accuracy

```
===== Building the Model =====
Accuracy of the Random Forest model: 0.9285714285714286
Classification Report:
      precision    recall  f1-score   support

     0       1.00      1.00      1.00         6
     1       1.00      0.75      0.86         4
     2       0.80      1.00      0.89         4

   accuracy       0.93
  macro avg       0.93
 weighted avg       0.93

===== Model built =====


===== Evaluating the Model =====
Cross-Validation Accuracies: [0.5 0.5 0.75 0.75 1.  1.  1.  1.  1.  1.  0.75 1.  1.
 1.  1.  1.  1.  1. ]
Average Accuracy: 0.9078947368421053

===== Model evaluated =====
```

An accuracy of 93% for multiclass classification on 70 records, along with a cross-validation accuracy of 91%, indicates strong and consistent model performance. However, there is still potential for improvement as more data becomes available for training.

## E) Swagger API for Testing Endpoints:

1. `/smart_lead_predictor`: Tests if a company is a hot, warm, or cold lead by analyzing uploaded company profiles.
2. `/ai_lead_generator`: Analyzes potential clients and automatically sends personalized emails based on lead classification.

 **Swagger**  
OPEN API SPECIFICATION

[Explore](#)

## A swagger API 0.0.1

[/apispec\\_1.json](#)

powered by [Flasgger](#)

[Terms of service](#)

default

▼

POST

`/api/ai_lead_generator` Upload a file and read CSV content to analyze and target potential clients

post\_api\_ai\_lead\_generator

POST

`/smart_lead_predictor` Predict lead type based on SME data.

post\_smart\_lead\_predictor

[Powered by [Flasgger](#) 0.9.7.1]

#### F) **Integration and Scalability Plan:**

##### 1. Integration with CRM Systems (e.g., HubSpot):

- API integration using HubSpot's RESTful API with OAuth 2.0 for authentication.
- Data synchronization and field mapping between systems.
- Real-time updates using webhooks.
- Automated workflows for lead scoring and email sequences.
- Comprehensive testing, documentation, and ongoing support.

##### 2. Scalability Plan:

- Use cloud infrastructure with auto-scaling and load balancing.
- Database sharding and indexing for optimization.
- Asynchronous processing with task queues and batch processing.
- Caching frequently accessed data to improve performance.
- Implement microservices architecture, API rate limiting, and robust monitoring.

#### G) **Performance Metrics:**

- Key metrics include Lead Quality Score, and Engagement Metrics.
- With Increase in engagement level the clients are targeted more for sending personalized business solutions.
- Leads are categorized into Hot, Warm, and Cold based on lead quality scores.

#### H) **Ethical Considerations:**

- Data privacy through encryption, environment variables for secret storage, and strict access controls.
- Bias mitigation and fairness in AI-driven lead scoring models.
- Transparency in data usage and AI decision-making.
- Regular security audits and updates for code and libraries.

#### **I) Analysis and Future Scope:**

The model was trained on a dataset of 70 company profiles, including their descriptions, company size, and revenue. The analysis showed a trend where companies in the healthcare sector are adopting AI strategies regardless of their size and revenue. For instance, even if a healthcare company is small in terms of size and revenue, it's likely to be classified as a warm lead due to the industry's focus on AI. Larger healthcare companies with higher revenue and more employees are categorized as hot leads, as they are in a stronger position to adopt AI solutions.

Additionally, companies with descriptions emphasizing terms like "data-driven" or "technology" are more inclined to adopt AI strategies, even if their revenue and employee count are lower. On the other hand, companies that focus less on technology, regardless of their size and revenue, are generally considered cold leads.

The goal was to build a model that generalizes well rather than overfitting to the data. However, due to the small dataset, the model's performance is limited. In the future, when larger datasets are available, the model is expected to perform significantly better. Given that this is a multi-class classification problem, a Random Forest classifier was chosen for its effectiveness in handling multiple classes through its ensemble scoring technique.

Based on current trends, sectors like e-commerce and healthcare are increasingly adopting AI solutions. In the future, this model can be further utilized to analyze emerging trends and filter potential clients, making the entire process of lead generation and client engagement more automated and efficient.