# Smart Lead Scoring – Internship Task Report

**Position Applied:** Machine Learning Engineer Intern
**Candidate Name:** *Hanana*

## I. Objective

To develop a lightweight, explainable machine learning tool that evaluates and ranks business leads based on features like seniority, company type, and contact presence. The goal is to support fast and transparent lead qualification.

## II. Method

- **Data Generation:** Created 150 synthetic leads with realistic B2B features including Title, Industry, Company Size, Email/LinkedIn presence, and Domain Score.
- **Labeling:** Leads were labeled based on business rules — for example, senior titles with verified contact information were marked as "good" leads.
- **Feature Engineering:** Applied fixed mappings to encode categorical features in a way that preserves interpretability.
- **Modeling:** Trained a Random Forest classifier using six selected features. The model outputs a probability-based lead score scaled from 0–100.
- **Explainability:** Used SHAP to explain feature importance and make the model decisions interpretable.

## III. Results

- Achieved **100% accuracy** on the rule-labeled data.
- Top features: **Title**, **Domain Score**, and **Email Present**.
- Developed a **Streamlit app** that accepts CSV input, scores leads, visualizes top-ranked leads, and shows SHAP-based feature impact.

## IV. Tech Stack

Python, scikit-learn, pandas, SHAP, Streamlit, Google Colab

## V. Business Relevance

This tool can help Caprae quickly assess scraped leads and identify high-priority contacts. It supports faster outreach, filters low-value entries, and provides transparency via feature explainability.

## VI. Next Steps

If extended, this project could be integrated with Caprae's real scraping pipeline and CRM feedback loop to further improve prediction quality.