RAJALAKSHMI
ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

# YOUTUBE DATA HARVESTING AND WAREHOUSE MANAGEMENT SYSTEM

## A MINI PROJECT REPORT

**SUBMITTED BY**

| | |
|---|---|
| **JEVANANDHAM S** | **231501069** |
| **KRISHNA KUMAR N K** | **231501081** |
| **HARSHAVARDHAN S** | **231501504** |

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

RAJALAKSHMI ENGINEERING COLLEGE (AUTONOMOUS)

THANDALAM

CHENNAI-602105

RAJALAKSHMI
ENGINEERING COLLEGE

**2024-2025**

# BONAFIDE CERTIFICATE

CERTIFIED THAT THIS PROJECT REPORT "**YOUTUBE DATA HARVESTING AND WAREHOUSE MANAGEMENT SYSTEM**" IS THE BONAFIDE WORK OF **"JEVANANDHAM S [231501069], KRISHNA KUMAR N K [231501081] AND HARSHAVARDHAN S [231501504]"** WHO CARRIED OUT THE PROJECT WORK UNDER MY SUPERVISION.

**Submitted for the Practical Examination held on ________________**

SIGNATURE

**Mr. U. Kumaran,**
**Assistant Professor (SS)**
**AIML,**
**Rajalakshmi Engineering College,**
**(autonomous)**
**Thandalam, Chennai - 602 105**

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

# ABSTRACT

The rapid growth of digital content on platforms such as YouTube has necessitated the development of sophisticated data harvesting and warehouse management systems to efficiently handle vast amounts of user-generated data. This paper explores the design and implementation of a YouTube data harvesting system, which involves the extraction, aggregation, and storage of various content-related metrics, including video views, user interactions, metadata, and engagement patterns. The system leverages web scraping techniques, YouTube API integration, and machine learning algorithms to collect and categorize data in real-time, ensuring its relevance and accuracy. Once harvested, the data is processed and stored in a structured data warehouse, optimized for performance, scalability, and quick retrieval. The warehouse management system incorporates advanced indexing and query optimization methods, enabling businesses, content creators, and analysts to gain actionable insights into audience behavior, content performance, and trending topics. Furthermore, the paper highlights the challenges faced in managing large-scale data, such as data redundancy, latency, and privacy concerns, offering solutions to mitigate these issues. Through this integrated system, organizations can better understand the dynamics of the YouTube platform, drive targeted marketing strategies, and enhance content recommendations, contributing to the overall efficiency and growth of the digital ecosystem.