**INTERNSHIP COMPLETION REPORT**

|  |  |
| --- | --- |
| **Submitted by** | **Submitted to** |
| **A.Hemalatha** | **Internship Organization:** Codsoft |
| B.Tech Information Technology (Graduated: July 2024) | **Internship Domain:** Machine Learning |
| IFET College of Engineering, Villupuram | **Internship Duration**: July 2025 (1 Month) |

**Projects Completed:**

1. Movie Genre Classification
2. Credit Card Fraud Detection
3. Spam Message Classifier

Submitted in partial fulfillment of personal upskilling and professional development.

**Date of Submission:** August 2025

**Declaration**

I, **A. Hemalatha**, a graduate in **B.Tech – Information Technology** (Class of 2024), hereby declare that the internship report titled **“Internship Completion Report – CodSoft (Machine Learning)”** is a **record of my original work** carried out during my one-month internship at **CodSoft in July 2025**.

This report has been prepared as part of my **self-learning and professional development** and has not been submitted elsewhere for any academic or professional credential.

I confirm that the content of this report is genuine and based on the tasks and projects I personally completed during the internship period.

**Place:** Villupuram  
**Date:** August 2025

**Signature of the Student**  
( *A. Hemalatha* )

**Acknowledgement**

I would like to express my sincere gratitude to **CodSoft** for providing me with the opportunity to undertake this internship in the domain of **Machine Learning** during **July 2025**. This internship has given me valuable hands-on experience and enhanced my skills in working with real-world datasets and models.

I am especially thankful to the **CodSoft team and mentors** for their support, timely feedback, and technical guidance throughout the internship. Their insights played a crucial role in helping me complete the projects successfully.

I also thank everyone who encouraged and supported me during this journey of continuous learning and career development.

**A. Hemalatha**  
B.Tech – Information Technology  
(Graduated in July 2024)

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Title** | **Page No.** |
| 1. | Introduction | 5 |
| 2. | Company Profile | 6 |
| 3. | Internship Objectives | 7 |
| 4. | Work done/Project details | 8 |
|  | 4.1. Task 1: Movie Genre Classification | 8 |
|  | 4.2. Task 2: Spam SMS Classifier | 10 |
|  | 4.3. Task 3: Credit Card Fraud Detection | 12 |
| 5. | Challenges faced | 14 |
| 6. | Key Learnings | 17 |
| 7. | Conclusion | 19 |
| 8. | References | 20 |
| 9. | Annexure | 21 |

### ****Introduction****

After completing my **B.Tech in Information Technology** in July 2024, I have been consistently working on enhancing my technical skills and preparing for a career in the fields of **Artificial Intelligence**, **Machine Learning**, and **Data Science**. As part of this self-learning journey, I undertook a **1-month internship** at **CodSoft** in **July 2025**.

The primary objective of this internship was to gain practical experience in **machine learning model development**, **data preprocessing**, and **problem-solving using real-world datasets**. Through this internship, I had the opportunity to work on three industry-relevant projects that helped me bridge the gap between academic knowledge and practical application.

Each project was designed to strengthen different aspects of machine learning:

* **Natural Language Processing** (Spam Message Classifier, Movie Genre Classification)
* **Binary Classification & Fraud Detection** (Credit Card Fraud Detection)
* **Model training, evaluation, and deployment using tools like Gradio and Scikit-learn**

This report summarizes the projects completed, methodologies followed, tools used, challenges faced, and the learnings gained during the internship.

### ****Company Profile****

### 

**CodSoft** is an **IT services and consulting firm** founded in **2022**, headquartered in **Kolkata, West Bengal**. The company operates with a compact team of **around 11–50 employees**, focused on delivering innovative software solutions to businesses

**Mission:** To continuously improve software through agile, feedback-driven development while staying updated with industry trends.

**Vision:** To create impactful, forward-looking software solutions using emerging technologies.

**Services Offered:**

* Web Development (e-commerce, responsive, custom sites)
* Mobile App Development (Android)
* AI, Machine Learning, and Data Science Solutions

CodSoft offers **4-week internship programs** in domains like Web, Android, Java, Python, Data Science, and AI/ML, focusing on real-time projects with mentorship. Interns receive an offer letter, certificate, and placement support.

With a small and supportive team, CodSoft provides a good environment for hands-on learning, flexible work-from-home options, and practical experience.

For more information, visit the official website: [www.codsoft.in](http://www.codsoft.in)

1. **Internship Objectives**

#### The key objectives included:

* To apply machine learning techniques to solve real-world problems using Python and Scikit-learn.
* To gain practical exposure to the end-to-end workflow of an ML project — from data preprocessing to model deployment.
* To understand and implement classification algorithms such as **Logistic Regression**, **Naive Bayes**, and **multi-label classification**.
* To improve my understanding of **Natural Language Processing (NLP)** for tasks like spam detection and text-based genre classification.
* To build interactive ML applications using **Gradio** for user-friendly deployment.
* To enhance coding, problem-solving, and project documentation skills through real-time project assignments.
* To develop a project portfolio that demonstrates my abilities to potential employers.

This internship helped me bridge the gap between theoretical learning and industry expectations while preparing for a full-time career in **AI, ML, and Data Science**.

### ****4. Work Done / Project Details****

## **4.1. Task 1: Movie Genre Classification**

**Project Title:** Movie Genre Classification using Multi-Label Logistic Regression  
**Objective:**  
To build a multi-label classification model that predicts one or more genres of a movie based on its description using Natural Language Processing (NLP).

**Dataset Used:**  
Custom text dataset from train\_data.txt containing movie descriptions and corresponding genres.

**Tools and Technologies:**

* Python Programming
* Pandas
* Scikit-learn
* TfidfVectorizer
* Logistic Regression
* MultiLabelBinarizer

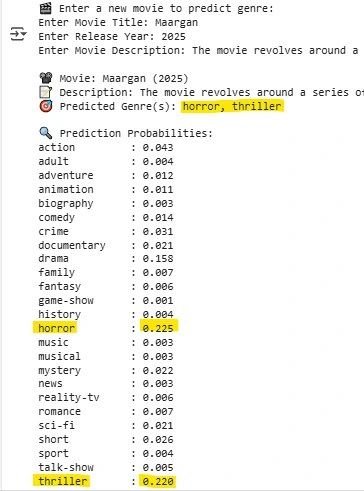
**Methodology:**

* Preprocessed movie descriptions using regex
* Vectorized text using TF-IDF
* Binarized genre labels for multi-label classification
* Trained a Logistic Regression model using OneVsRest strategy
* Set prediction threshold to 0.2 for better genre coverage

**Result:**  
Successfully predicted genres like Action, Drama, Sci-Fi for new movie descriptions based on text input. Model provided probability scores for each genre.

**GitHub Link:** [Movie Genre Classification](https://github.com/HEMALATHAifet/CODSOFT_Machine-Learning-Internship/tree/main/TASK1)

**LinkedIn Post:** [Movie Genre Classification](https://www.linkedin.com/posts/hemalatha-a-developer_day7-toabrshow-movieabrgenreabrclassification-activity-7348985840330518528-dPQ6?utm_source=share&utm_medium=member_desktop&rcm=ACoAAFwkWN8B0ZcDLaTb1j2WX_pvC5NFm1Gl_aA)



**4.2. Task 2: Spam Message Classifier**

**Project Title:** SMS Spam Detection using Naive Bayes  
**Objective:**  
To create a model that classifies SMS messages as spam or not spam using Natural Language Processing and Naive Bayes algorithm.

**Dataset Used:** Public dataset from [kaggle](https://raw.githubusercontent.com/justmarkham/pycon-2016-tutorial/master/data/sms.tsv)

**Tools and Technologies:**

* Python Programming
* Pandas
* Scikit-learn
* CountVectorizer
* Multinomial Naive Bayes
* Gradio

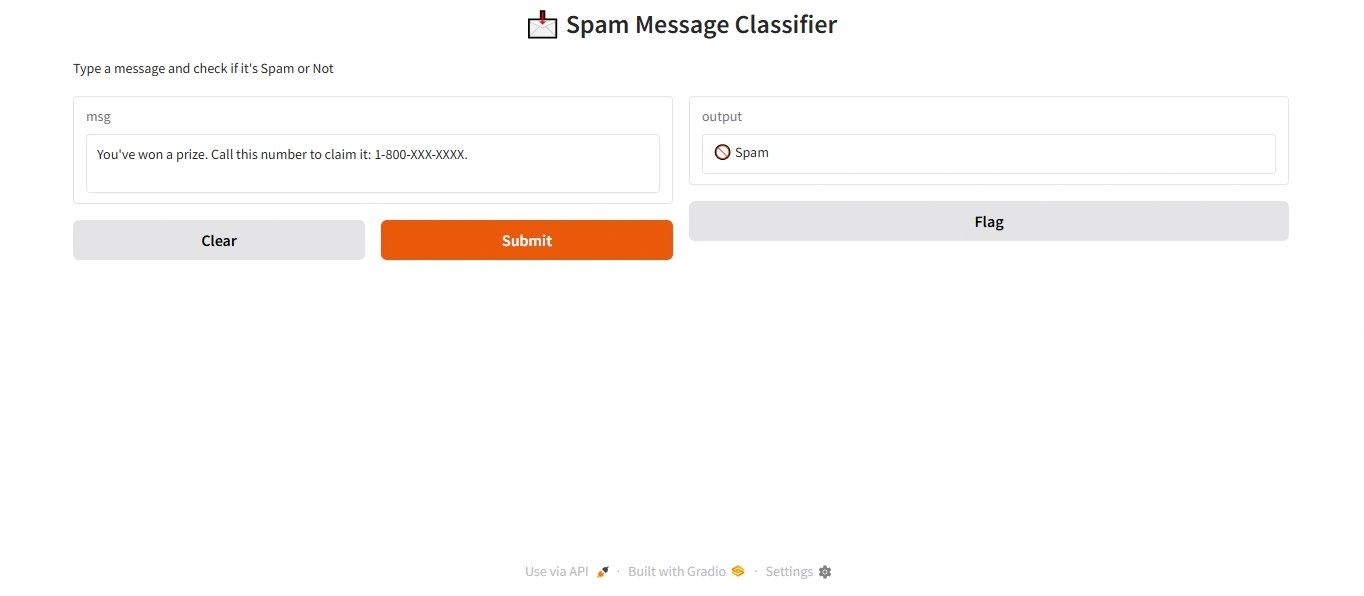
**Methodology:**

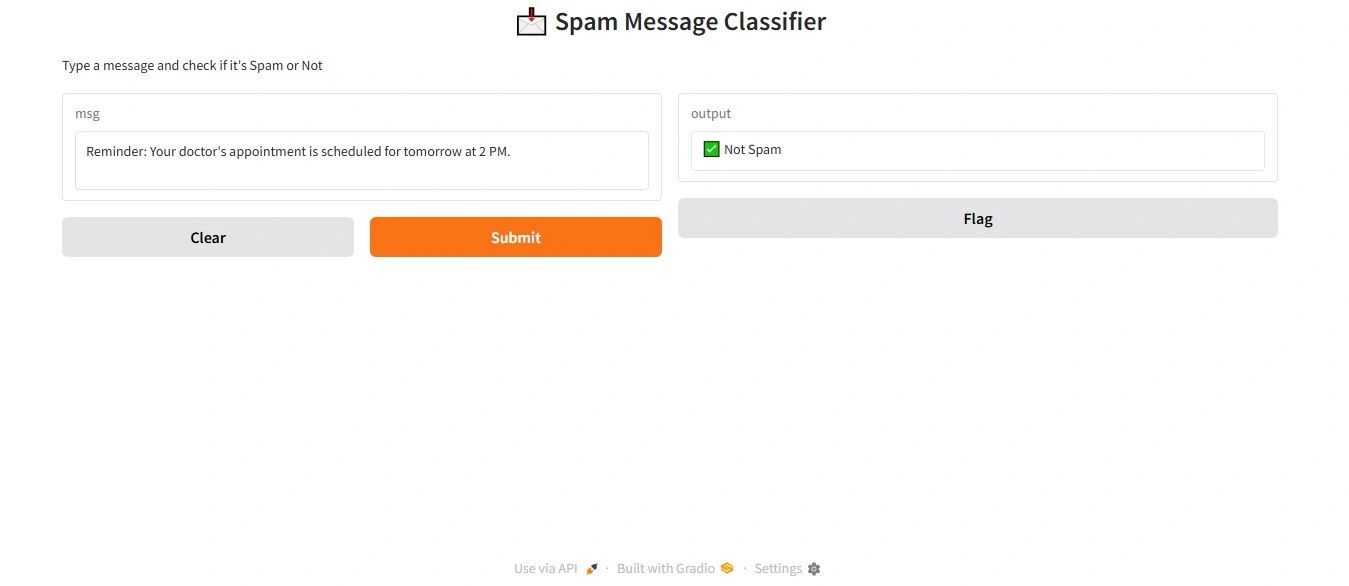
* Preprocessed SMS text data
* Created a pipeline using CountVectorizer and MultinomialNB
* Trained the model on 80% of the data
* Built a Gradio UI for user input and real-time classification

**Output / Results:**  
Model classified user-input messages as **SPAM** or **NOT SPAM** with high accuracy and a responsive UI.

**GitHub Link:** [SPAM SMS DETECTOR](https://github.com/HEMALATHAifet/PRODIGY_GenAI_Internship/tree/main/PRODIGY_GA_01)

**LinkedIn Post:** [SPAM SMS CLASSIFIER](https://www.linkedin.com/posts/hemalatha-a-developer_day8-codsoft-internship-activity-7349371182179082240-31yo?utm_source=share&utm_medium=member_desktop&rcm=ACoAAFwkWN8B0ZcDLaTb1j2WX_pvC5NFm1Gl_aA)





## **Task 3: Credit Card Fraud Detection**

**Project Title:** Credit Card Fraud Detection using Logistic Regression  
**Objective:**  
To develop a machine learning model that identifies fraudulent transactions based on transaction details like amount, location, device type, and international status.

**Dataset Used:**  
Credit\_card\_Fraud\_detection\_dataset.csv (custom dataset)

**Tools and Technologies:**

* Python Programming
* Pandas
* Scikit-learn
* Label Encoding
* Logistic Regression
* Gradio

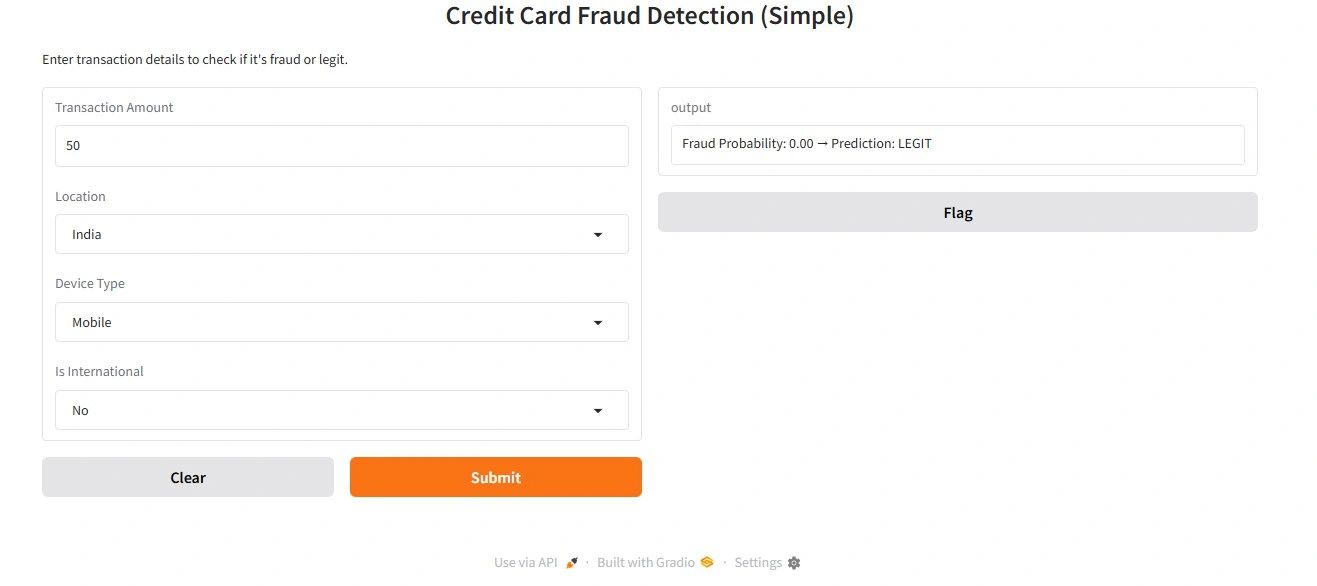
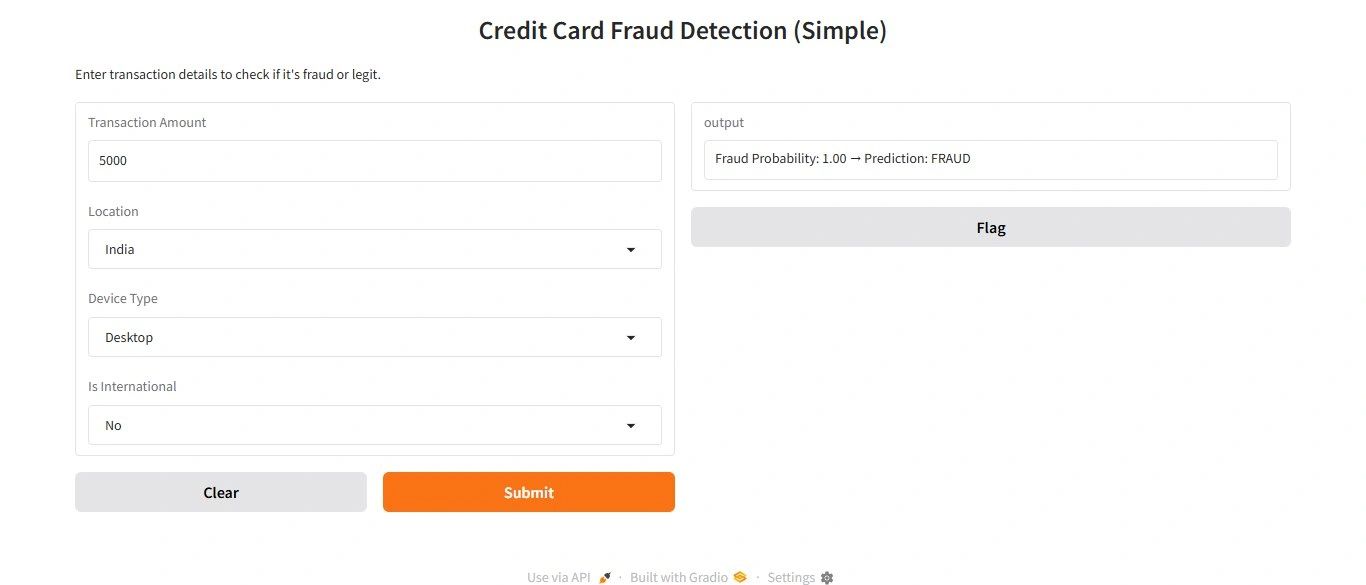
**Methodology:**

* Encoded categorical features (Location, DeviceType, IsInternational)
* Trained a Logistic Regression model to classify transactions
* Built a Gradio web interface for real-time fraud prediction

**Results:**  
Model classified transactions as **LEGIT** or **FRAUD** with associated probability score. Achieved effective binary classification with clean UI.

**GitHub link:** [Credit card fraud identification](https://github.com/HEMALATHAifet/CODSOFT_Machine-Learning-Internship/tree/main/TASK3)

**LinkedIn link:** [Credit card fraud identification](https://www.linkedin.com/posts/hemalatha-a-developer_day9-firstabrjob-creditabrcardabrfraudabrdetection-activity-7350148990111215616-CZGh?utm_source=share&utm_medium=member_desktop&rcm=ACoAAFwkWN8B0ZcDLaTb1j2WX_pvC5NFm1Gl_aA)



### 5. Challenges Faced

#### **Task 1: Movie Genre Classification**

**Technical Challenges:**

* Implementing **multi-label classification** using OneVsRestClassifier was new to me, and required learning how to adapt traditional classifiers to this format.
* Managing memory while working with **TF-IDF vectorization** on large text input required optimization.

**Learning Curve:**

* Understanding how genre probabilities work and choosing a suitable **probability threshold** for multi-label output was initially confusing.
* Fine-tuning text preprocessing (like cleaning, tokenizing) significantly affected model performance and required trial-and-error.

**Data Challenges:**

* The raw text file (train\_data.txt) had inconsistent formatting. Parsing it into structured data (description, genre pairs) required careful scripting.
* Genre labels varied across samples, so **binarizing multiple genres per movie** took effort using MultiLabelBinarizer.

#### **Task 2: Credit Card Fraud Detection**

**Technical Challenges:**

* First-time use of **Gradio** for creating UI interfaces caused issues like incorrect input types and encoding mismatches.
* Handling **LabelEncoder consistency** across training and prediction phases required proper structuring to avoid incorrect category mapping.

**Learning Curve:**

* I had to learn how to interpret **classification report metrics** (precision, recall, F1-score), especially for imbalanced datasets.
* Balancing model simplicity with performance in Logistic Regression, without overfitting to minority fraud cases, was a challenge.

**Data Challenges:**

* The dataset was **highly imbalanced** (more LEGIT than FRAUD cases), which impacted model predictions.
* Categorical fields like Location, DeviceType, and IsInternational required encoding and mapping for both training and real-time inference.

#### **Task 3: Spam Message Classifier**

**Technical Challenges:**

* Creating a **pipeline** with CountVectorizer and MultinomialNB took multiple attempts due to incorrect fit/transform steps.
* Ensuring that the Gradio app accepted free text inputs and produced consistent output during testing required UI refinements.

**Learning Curve:**

* Understanding the differences between **CountVectorizer vs. TfidfVectorizer** and when to use which was new to me.
* I also had to explore how MultinomialNB works with sparse matrix inputs and why it's suitable for text classification.

**Data Challenges:**

* Some messages had emojis, special characters, or non-standard text formats, requiring basic cleaning for better model learning.
* The dataset had a **class imbalance** (more ham than spam), affecting the classifier’s ability to detect spam initially.

### 6. Key Learnings

**Machine Learning internship** at CodSoft provided a strong foundation in applying machine learning to real-world problems. It helped me gain both **technical knowledge** and **professional skills** essential for a career in Artificial Intelligence and Data Science.

#### **Technical Tools Learned:**

* **Scikit-learn** – For model training, classification, and evaluation.
* **Pandas** – For data manipulation and preprocessing.
* **Gradio** – For building simple web-based interfaces for ML model deployment.
* **TfidfVectorizer & CountVectorizer** – For text representation in NLP projects.
* **LabelEncoder & MultiLabelBinarizer** – For handling categorical and multi-label data.
* **Pipeline (Scikit-learn)** – For streamlining text classification workflows.

#### **Concepts Understood:**

* **Natural Language Processing (NLP):** Text cleaning, feature extraction, spam detection.
* **Multi-label Classification:** Predicting multiple genres from movie descriptions.
* **Binary Classification:** Fraud detection using Logistic Regression.
* **Model Evaluation Metrics:** Precision, Recall, F1-Score, and interpreting classification reports.
* **Data Preprocessing:** Handling imbalanced data, label encoding, and normalization.

#### **Soft Skills Developed:**

* **Time Management:** Completing 3 projects within a fixed internship duration with self-discipline.
* **Problem Solving:** Debugging code, interpreting model behavior, and optimizing predictions.
* **Communication Skills:** Writing clear code and documentation, creating LinkedIn posts to share learning.
* **Adaptability:** Quickly learning new tools like Gradio and applying them effectively.
* **Self-motivation:** Managing this internship independently as a graduate aiming for job readiness.

**7. Conclusion**

Completing the Machine Learning internship at CodSoft has been an enriching and transformative experience. Through this internship, I had the opportunity to work on real-world projects that involved solving **classification problems** using **Machine Learning** and **Natural Language Processing techniques.**

I gained hands-on experience with tools like **Scikit-learn, Pandas, and Gradio**, and improved my understanding of **NLP, multi-label classification, and fraud detection**.

This internship strengthened my confidence, enhanced my technical and problem-solving skills, and added strong projects to my portfolio. It has brought me one step closer to my career goal of securing a job in **AI/ML or Data Science**.

As I continue to improve my skills and build experience, I plan to:

* Participate in more real-time projects and competitions.
* Contribute to open-source AI/ML projects.
* Keep updating my knowledge with the latest tools and techniques in the field.

Overall, the internship was a valuable step in my journey from a fresher to a job-ready ML professional.

### 8. References

#### GitHub Repository: [TASK1 -TASK2 - TASK3](https://github.com/HEMALATHAifet/CODSOFT_Machine-Learning-Internship/tree/main/TASK3)

* LinkedIn: [HEMALATHA. A](https://www.linkedin.com/in/hemalatha-a-developer/)

#### Documentation & Tutorials Referred:

|  |  |  |
| --- | --- | --- |
| 1. | Scikit-learn Documentation | <https://scikit-learn.org/stable/documentation.html> |
| 2. | Gradio Documentation | <https://www.gradio.app/docs> |
| 3. | NLP Guide | <https://www.gradio.app/docs> |
| 4. | Python for ML | <https://realpython.com/> |

**9. Annexure**

****