**Contextual Text Extraction from PDF files Using Machine Learning Practices**

PROJECT REPORT

**Data Science with Python Programming**

**INDUSTRIAL PROJECT BASED LEARNING**



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301.

**By Team -4**

**Team Details :**

Shivanoor Vignesh 21R11A0597

G.SandhyaRani 21R11A0569

R.Dhathri 21R11A0545

Hasini 21R11A0518

A.Pooja 21R11A0501

PSBN Sriya 21R11A0595

N Sai Charan 21R11A0588

MD.Akbar Ali 21R11A0586

# ABSTRACT

# 

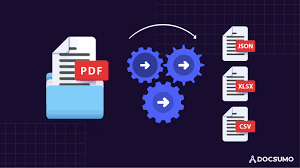
# The capstone project focuses on addressing the challenge of efficiently extracting specific information from a collection of PDF documents. With the increasing need for accessibility, readability, and data extraction from PDFs across various industries, the project aims to leverage Machine Learning (ML) techniques to automate the extraction process. ML models are well-suited for this task as they can accurately recognize and extract text from PDF files, considering both the layout and position of the text. Manual extraction of data from numerous PDFs is time-consuming and prone to errors. By employing ML-based text extraction techniques, the project aims to minimize manual intervention and enhance accuracy in extracting relevant information. The solution will enable users to efficiently retrieve required data from PDFs, saving time and ensuring consistency in extracted information. Through experimentation and evaluation, we assess the performance of our machine learning models in extracting contextual text from PDF files. We also explore techniques for improving the accuracy and efficiency of the extraction process, such as feature engineering and model optimization,aims to provide a reliable and scalable solution for extracting valuable information from PDF files using machine learning techniques, which can be applied across various domains and applications.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Chapter Name** | **Page No** |
|  | **ABSTRACT** | 2 |
| **1** | **INTRODUCTION** | 4 |
| **2** | **PROBLEM STATEMENT** | 5 |
| **3** | **OBJECTIVES** | 6 |
| **4** | **METHODOLOGY** | 7 |
|  | **4.1 Data Source** | 7 |
|  | **4.2Exploratory Data Analysis** | 8 |
|  | **4.3 Efficent Director Details Extraction from PDFs** | 9 |
| **5** | **Literature Review** | 10 |
| **6** | **DIFFERENT TYPES OF APPROACHES FOR THE SOLUTION** | 11 |
|  | 6.1 Using Cloud Vision | 11 |
|  | 6.2 using pdfplumber with flask api | 16 |
| **7** | **Final Implementation** | 18 |
| **8** | **Conclusion** | 23 |

# INTRODUCTION

* In today's digital era, the ubiquitous presence of PDF documents across various domains poses a challenge for efficient data extraction. Whether it's for accessibility purposes, legal compliance, or data analysis, the need to extract specific information from PDFs is a common requirement. However, manual extraction methods are time-consuming, error-prone, and often yield inconsistent results.To address this challenge, the capstone project focuses on leveraging Machine Learning (ML) techniques to automate the extraction of key information from PDF documents. ML models offer the ability to accurately recognize and extract text from PDF files, taking into account the layout and position of the text. This capability enables the development of a solution that can efficiently retrieve required data with minimal manual intervention.The primary objective of the project is to develop a robust solution capable of extracting crucial information such as the Name of the Director, Director type (Independent/Executive), and DIN (Director Identification Number) from a diverse range of PDF documents. These documents may include bank statements, tax forms, or legal documents with varying layouts and structures.



* By harnessing the power of ML-based text extraction techniques, the project aims to streamline workflows and improve productivity in tasks requiring data extraction from PDFs. The solution will not only save time but also ensure consistency and accuracy in extracted information, thereby enhancing the overall efficiency of data processing workflows.Through this project, the team seeks to contribute to the advancement of automated data extraction technologies, making it easier for individuals and organizations to access and utilize information stored in PDF documents effectively. By providing a reliable and scalable solution, the project aims to address the growing demand for efficient PDF data extraction across industries.

# PROBLEM STATEMENT

# The problem at hand revolves around the inefficiency and inconsistency associated with manually extracting specific information from a collection of PDF documents. Despite the increasing reliance on digital documents, extracting key data elements such as the Name of the Director, Director type (Independent/Executive), and DIN (Director Identification Number) remains a cumbersome and time-consuming task.The current methods of manual extraction are prone to errors and are not scalable, particularly when dealing with a large volume of PDF documents. Additionally, the unstructured nature of data within PDFs poses a significant challenge for automated systems attempting to extract relevant information accurately. Machine Learning (ML) techniques are considered one of the best methods for PDF extraction because it allows for highly accurate text recognition and extraction from PDF files regardless of the file structure. These models can store information of both the `layout` and the `position of the text` keeping in mind the neighboring text too. This helps them to generalize better and learn document structure more efficiently.

# Therefore, the primary problem statement of the project is to develop an automated solution using Machine Learning (ML) techniques that can efficiently and accurately extract required information from PDF documents. The solution should be capable of handling various types of PDF layouts and structures, ensuring robustness and reliability across different document formats.

# Designed solution should be able to extract the Name of the Director.

# Director type : Independent / executive

# DIN (number)

# Objectives

* Design and implement a robust solution leveraging Machine Learning (ML) techniques to automate the extraction of key information from PDF documents, including the Name of the Director, Director type (Independent/Executive), and DIN (Director Identification Number).
* Create a solution capable of handling the variability in PDF document layouts and structures, ensuring adaptability and versatility across different document formats and styles.
* Ensure that the extracted information is accurate and consistent across various PDF documents, minimizing errors and discrepancies in the extracted data.
* Streamline workflows and improve efficiency by automating the data extraction process, reducing manual intervention, and saving time in tasks requiring information retrieval from PDF documents.
* Optimize the performance of the solution to handle large volumes of PDF documents efficiently, ensuring scalability and reliability in processing diverse datasets.
* Develop a user-friendly interface for the solution, enabling users to easily input PDF documents and retrieve extracted information in a convenient and accessible format.

# METHODLOGY

**4.1 Data Source**

The data sources are the pdf files given. The information in the pdf file consist of Directors Details.

* File attachment [equitas.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/ae4a795e-e79f-45b4-b69e-323197c8c565/equitas.pdf) ( 329 KB; Apr 7, 2024, 8:27 pm )
* File attachment [Triveni2.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/eeab3b1d-2e6d-4370-b12f-9eb5e247b35b/Triveni2.pdf) ( 872 KB; Apr 7, 2024, 8:28 pm )
* File attachment [varun.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/0f2c41ce-b0be-42ab-a784-572f03911ebf/varun.pdf) ( 3 MB; Apr 7, 2024, 8:28 pm )
* File attachment [PolyPlex.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/d9f85fe4-42ba-4dbb-b65b-75facc15764c/PolyPlex.pdf) ( 471 KB; Apr 7, 2024, 8:27 pm )
* File attachment [Triveni1.pdf](https://trysakai.longsight.com/access/content/attachment/ac5e0df4-75b6-48b8-813e-ee761c1393fb/Assignments/67d16034-2396-404f-8546-77de027ef9cc/Triveni1.pdf) ( 9 MB; Apr 7, 2024, 8:27 pm )

The text underscores the effectiveness of machine learning (ML) techniques in addressing these challenges by enabling highly accurate text recognition and extraction from PDF files, regardless of their structure. ML models are noted for their ability to understand the layout and position of text within PDFs, as well as the contextual information surrounding it, which enhances their capability to generalize and learn document structures efficiently.

The data souces are in the form of pdf . The datasource contains:

1. Name of the Director.

2. Director type : Independent / executive

3. DIN

Name of the Director: This refers to the full name of an individual serving as a director in a company or organization. The name typically includes both the first name and the last name of the director, such as "John Doe" or "Jane Smith."

Director Type (Independent/Executive): This parameter categorizes directors based on their role and relationship with the company's management:

Independent Director: These directors are not involved in the day-to-day operations of the company. They provide an independent perspective and are often part of the board's audit, nomination, and remuneration committees. They are expected to bring objectivity and impartiality to board decisions.

Executive Director: These directors are actively involved in the management and operations of the company. They may hold executive positions such as CEO, CFO, COO, etc., and are responsible for strategic decision-making and overseeing business functions.

DIN (Director Identification Number): DIN is a unique identification number assigned to individuals who serve as directors in companies. It is issued by the Ministry of Corporate Affairs in India and serves as a way to track and identify directors across different companies. DIN helps in maintaining a database of directors and their activities within the corporate sector.

# 4.2 EDA(exploration data analysis)

# In PDF extraction projects, the EDA process may deviate from the traditional structured data EDA workflow. However, it serves the purpose of understanding the underlying data characteristics, preprocessing requirements, and quality assessments needed for successful extraction.

# Firstly, I extracted data from the "equitas.pdf" file using pdfplumber on April 5, 2024. The document appears to contain information regarding approval from the Reserve Bank of India for the appointment of Mr. Anil Kumar Sharma as the Part-time Chairman of the Bank.

# 

# The document mentions the National Stock Exchange of India Limited and BSE Limited, along with their respective addresses. It also includes the symbol and scrip code of the bank.

# Regarding the annotations on the page, the document structure includes various objects that make up the visual content, such as Contents, CropBox, Group, MediaBox, Parent, Resources, ExtGState, Font, ProcSet, XObject, Rotate, Tabs, and Type. These elements provide detailed descriptions of the page's structure and content.

# The document consists of 2 pages in total.

# 4.3 Efficent Director Details Extraction from PDFs

# This Python script utilizes the pdfplumber library to extract text from PDF files. It then employs SpaCy, a powerful natural language processing library, for named entity recognition (NER) to identify persons mentioned in the text. For each recognized director, the script determines their role type (Independent or Executive) and extracts associated Director Identification Numbers (DINs) using regular expressions. These DINs are strategically located in proximity to the director's name within the text. Finally, the script presents the extracted director names alongside their roles and DINs, streamlining the extraction of critical corporate governance data from PDF documents.

# Literature survey

Existing methods for pdf extraction

Manual Extraction Techniques: These involve manual methods for extracting data from PDFs, which are time-consuming, error-prone, and lack scalability due to human intervention.

Natural Language Processing (NLP) Techniques: NLP algorithms, such as tokenization, parsing, and entity recognition, are applied for PDF extraction. Tools like spaCy are used to identify and classify named entities like person names, organizations, and dates.

Commercial and Open-Source Tools: Popular tools like Adobe Acrobat, PyPDF2, and Apache PDFBox are used for PDF extraction. These tools offer features for extracting text and metadata from PDF documents, but they may have limitations in terms of customization and scalability.

Supervised Learning Approaches: Supervised learning algorithms like Support Vector Machines (SVM), Random Forest, and Neural Networks are used for text extraction. These algorithms are trained on labeled data to recognize patterns and extract relevant information from PDFs.

Unsupervised Learning Techniques: Unsupervised learning methods like clustering and topic modeling are employed for text extraction. These techniques can identify patterns and group similar text elements together, but they may require manual intervention for interpretation.

Hybrid Approaches: Hybrid models combine rule-based, NLP, and machine learning techniques for improved accuracy and efficiency in text extraction from PDFs. These models leverage the strengths of each approach to overcome limitations and achieve better results.

Deep Learning for Text Extraction: Deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures are explored for text extraction from PDFs. These models can learn complex patterns and structures in text data, but they may require large amounts of labeled data and computational resources.

Using the Google cloud vision api :we can extract the text from pdf and store it into the json file and next json file can be read find the patterns required   
 and google also have ocr technique to read the iamge and pdf also.

# 6. DIFFERENT TYPES OF APPROACHES FOR THE SOLUTION

# 1.USING GOOGLE CLOUD VISION API

# 2. USING PDFPLUMBER WITH FLASK API

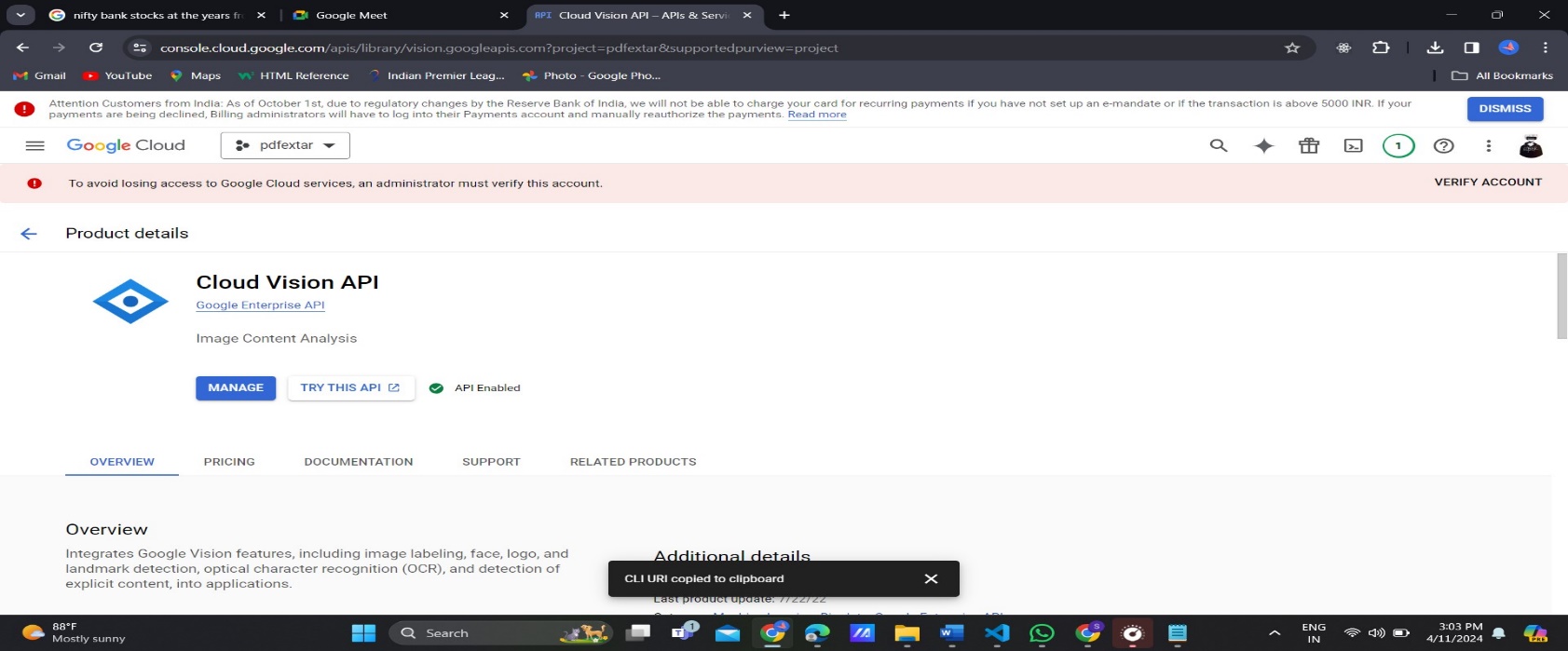
# Using the Google Cloud Vision API involves several steps:

# Create an account on console.cloud.google.com for a free account.

# Create the first project with a desired name, such as "pdfextraction."

# 

# Enable the Cloud Vision API by visiting the API library and searching for "Cloud Vision API," then enabling it.



# Visit Cloud Storage and open or create a bucket, naming it something like "pdfextraction97."

# 

# Open the bucket and upload the PDF files that need to be processed.

# 

# Write code to convert the stored PDF files to JSON-readable format,

# 

# 

# Upload an empty text file named "result.txt" to the bucket.

# 

# extracting relevant information such as director names, DINs, and roles.

# director\_info = re.findall(r"\. (.+?) \(DIN: (\d+)\) as\s+(.\*?)\s+(?:Director|Directors)", text)

# Copy the GSURI (Google Storage Uniform Resource Identifier) of the resulting JSON file (e.g., result1to15.json) in the code for reading and processing the JSON data.

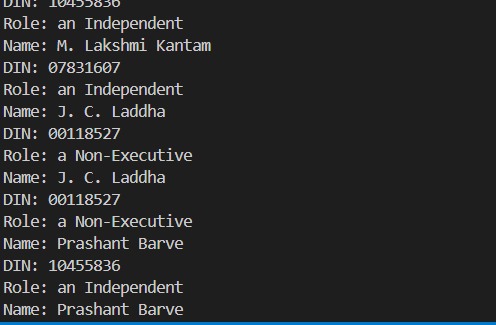
# os.environ['GOOGLE\_APPLICATION\_CREDENTIALS'] = "D:\downloads\pdfextar-6e5ae579f579.json"

# Iterate through the JSON data character by character, using regular expressions to match and extract relevant information, then print it to the console output in the desired format

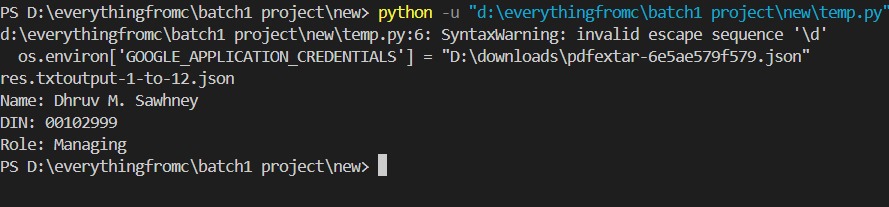
# e.g., {'director': 'Anil Kumar Sharma', 'din': '08537123', 'role': 'former Executive'}.

# The output when pattern matches

# 



**For another pdf file output as the**



1. **USING PDFPLUMBER WITH FLASK API**

The implementation of the web application using Flask involves the utilization of several libraries, including re, spacy, PyPDF2, pdfplumber, NLP libraries, and ML techniques. The patterns for retrieving 'Director Name', 'DIN', and 'status' are as follows:

director\_patterns = r"\. (.+?) \(DIN: (\d+)\),\s+(.\*?)\s+as\s+(.\*?)(?=;|\n|$)"

director\_pattern = r"(Mr\.|Mrs\.|Ms\.|Dr\.)\s+([A-Za-z]{2,})"

din\_pattern = r"DIN:\s(\d+)"

**director\_patterns = r"\. (.+?) \(DIN: (\d+)\),\s+(.\*?)\s+as\s+(.\*?)(?=;|\n|$)"**

Explanation:

* \.: Matches a literal period character.
* (.+?): Captures one or more of any character (except newline) lazily. The ? makes the + lazy, meaning it will match as few characters as possible.
* \(DIN: (\d+)\): Matches "(DIN: " followed by one or more digits captured within parentheses.
* \s+: Matches one or more whitespace characters (spaces, tabs, newlines, etc.).
* (.\*?): Captures zero or more of any character (except newline) lazily.
* \s+: Matches one or more whitespace characters again.
* as: Matches the literal string "as".
* (.\*?): Captures zero or more of any character (except newline) lazily.
* (?=;|\n|$): A positive lookahead assertion ensuring that the pattern is followed by either a semicolon, newline, or end of the line.

**director\_pattern = r"(Mr\.|Mrs\.|Ms\.|Dr\.)\s+([A-Za-z]{2,})"**

Explanation:

* (Mr\.|Mrs\.|Ms\.|Dr\.): Matches salutations like "Mr.", "Mrs.", "Ms.", or "Dr." using the pipe | as an "or" operator.
* \s+: Matches one or more whitespace characters (spaces, tabs, newlines, etc.).
* ([A-Za-z]{2,}): Captures two or more alphabetical characters (uppercase or lowercase).

**din\_pattern = r"DIN:\s(\d+)"**

Explanation:

* DIN:: Matches the literal string "DIN:".
* \s: Matches a single whitespace character (space).
* (\d+): Captures one or more digits.
* These regular expressions are designed to match specific patterns in text data, such as director names, DIN numbers, and their associated roles, based on predefined formats or keywords.

There is some of the outputs:

This output of Vinati.pdf.

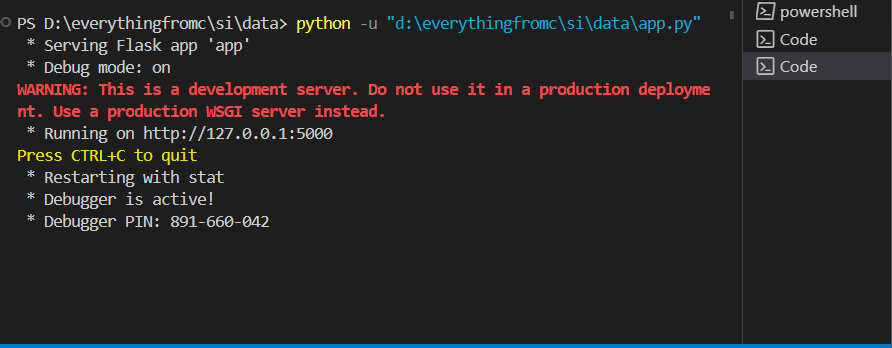


**Observation:**

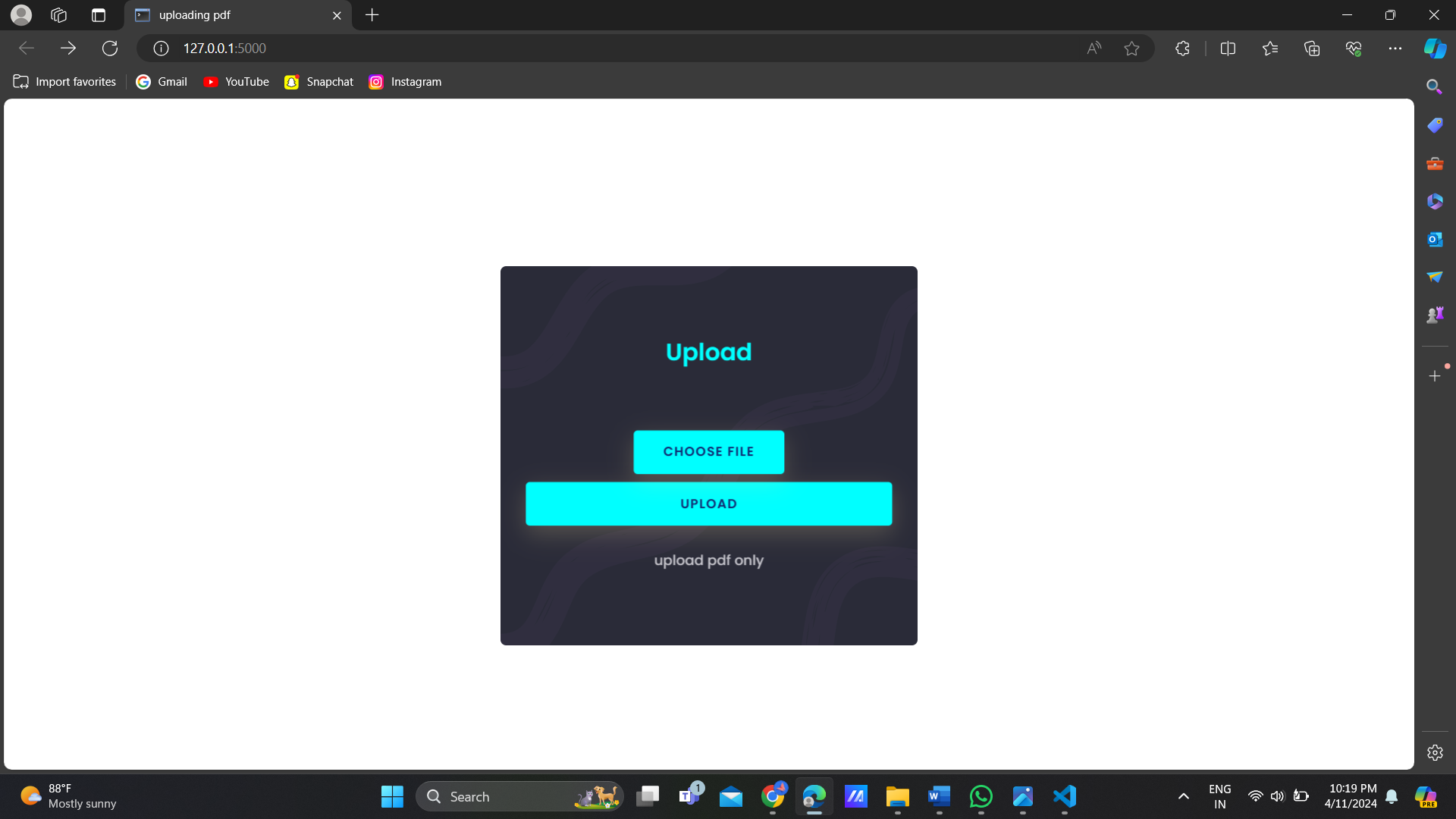
from the above pdf we can get the information of Prashanth Barveand din number of the 10455836 and an role of independent Director where regular expression matches that particular name correctly

**7.Final Implementation**

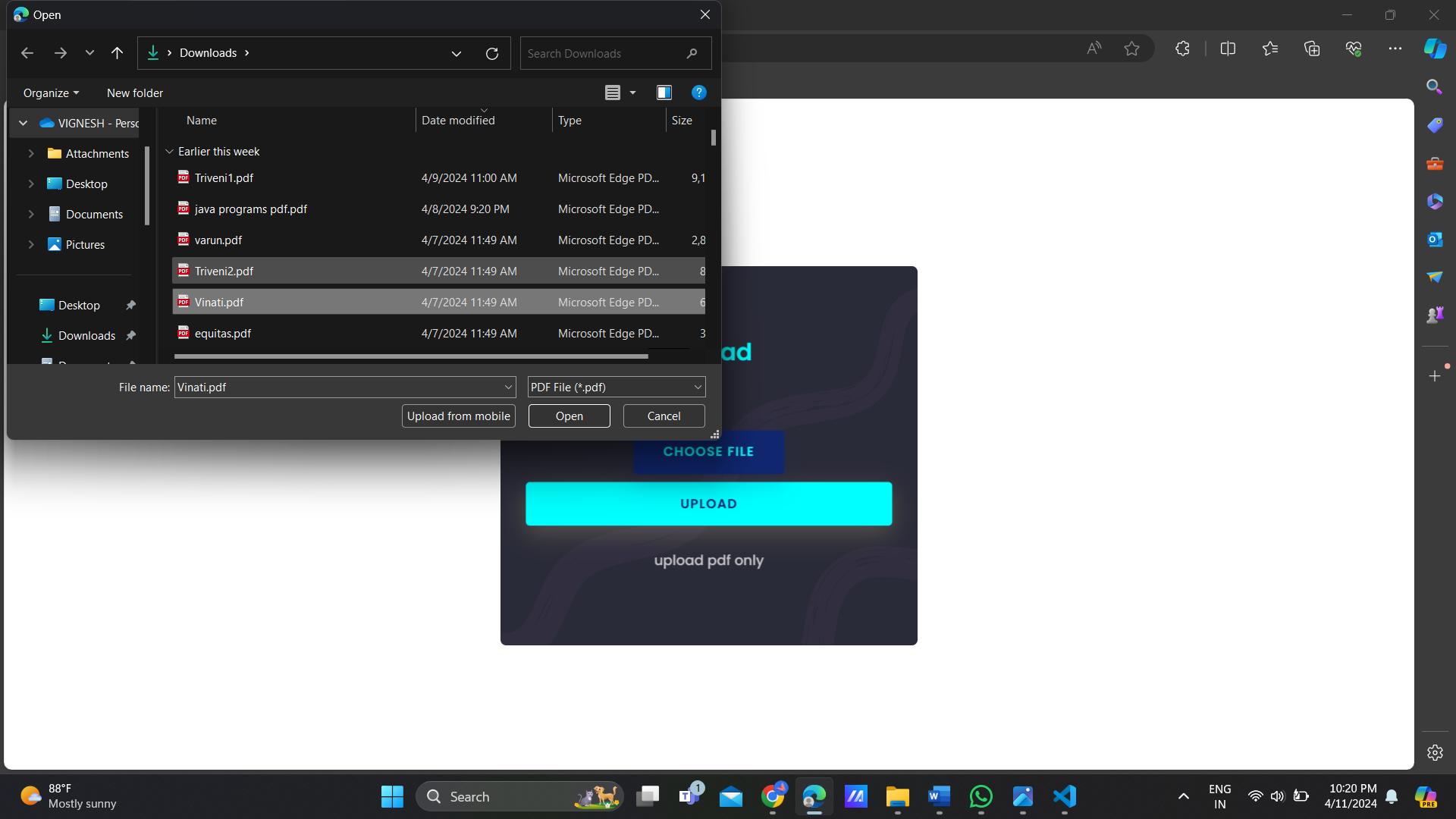
1. launch the web application in the python code, which will deploy in localhost

****

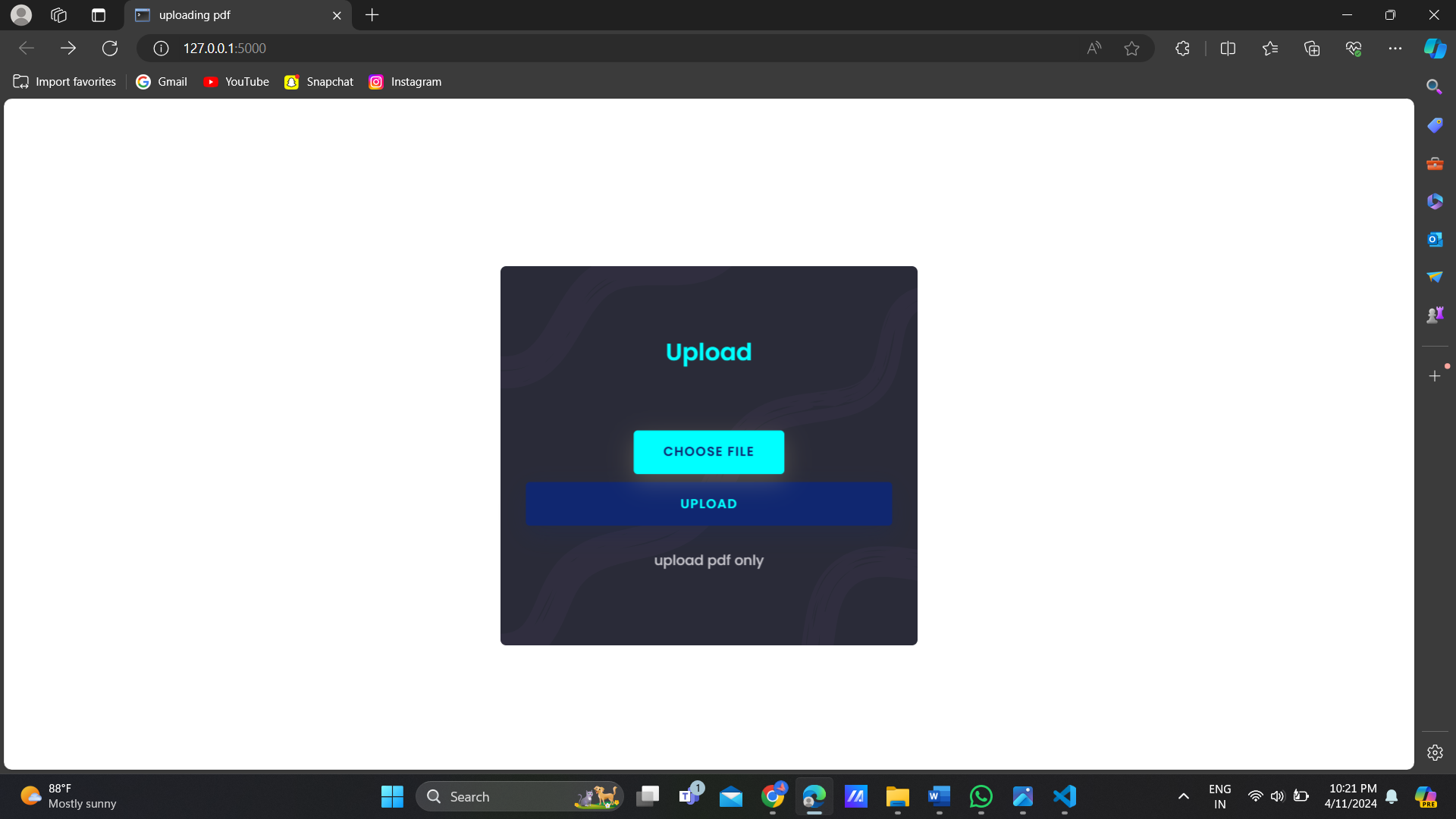
2.Open the deployed wenb application by clicking the <http://127.0.0.1:5000> in your local host that server has started



3.choose the pdf file which pdf file has din number ,name which starts from mr,mrs and have role of the director in the pdf



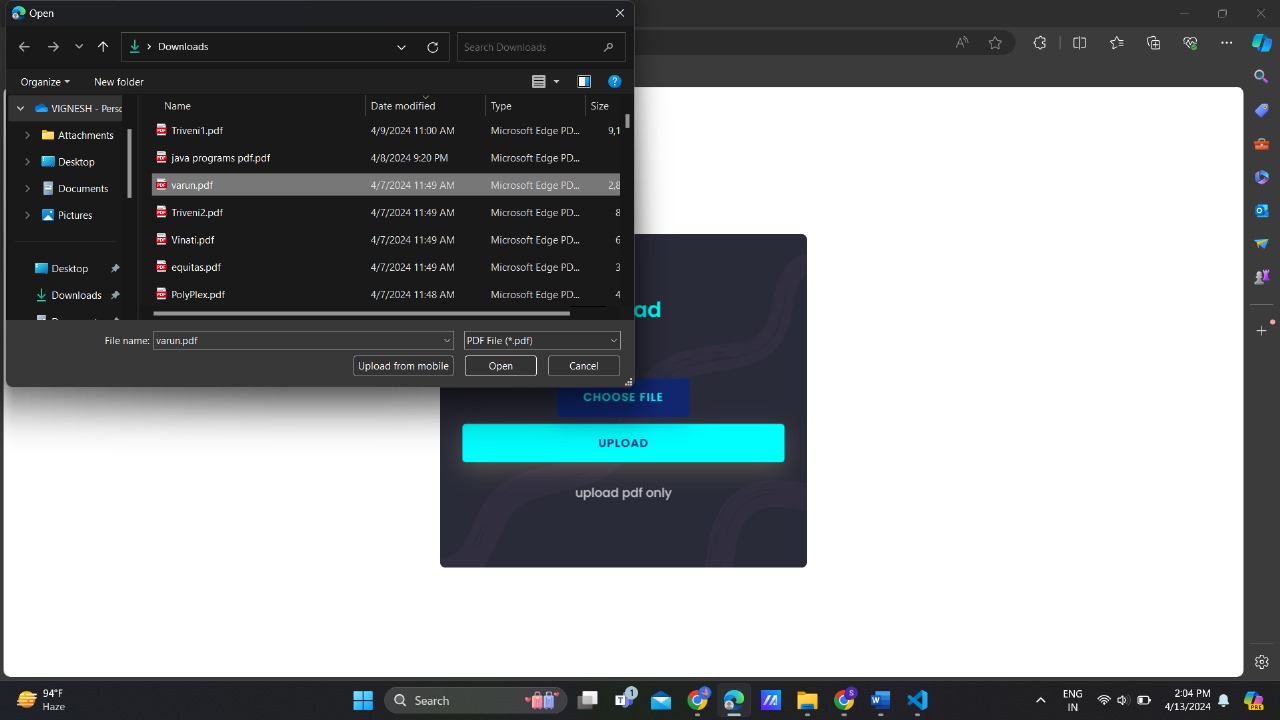
4.after choosing a file create the click on upload button which upload the pdf and read the pdf finds pattern matching using regular expression



5.after clicking the upload button if it matches the regular expression it prints in result.html page

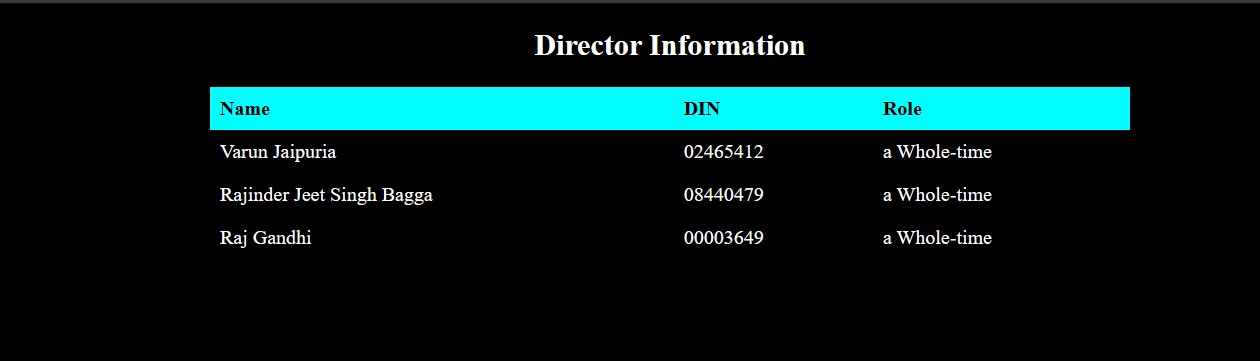


**TESTING ON THE VARUN.PDF**



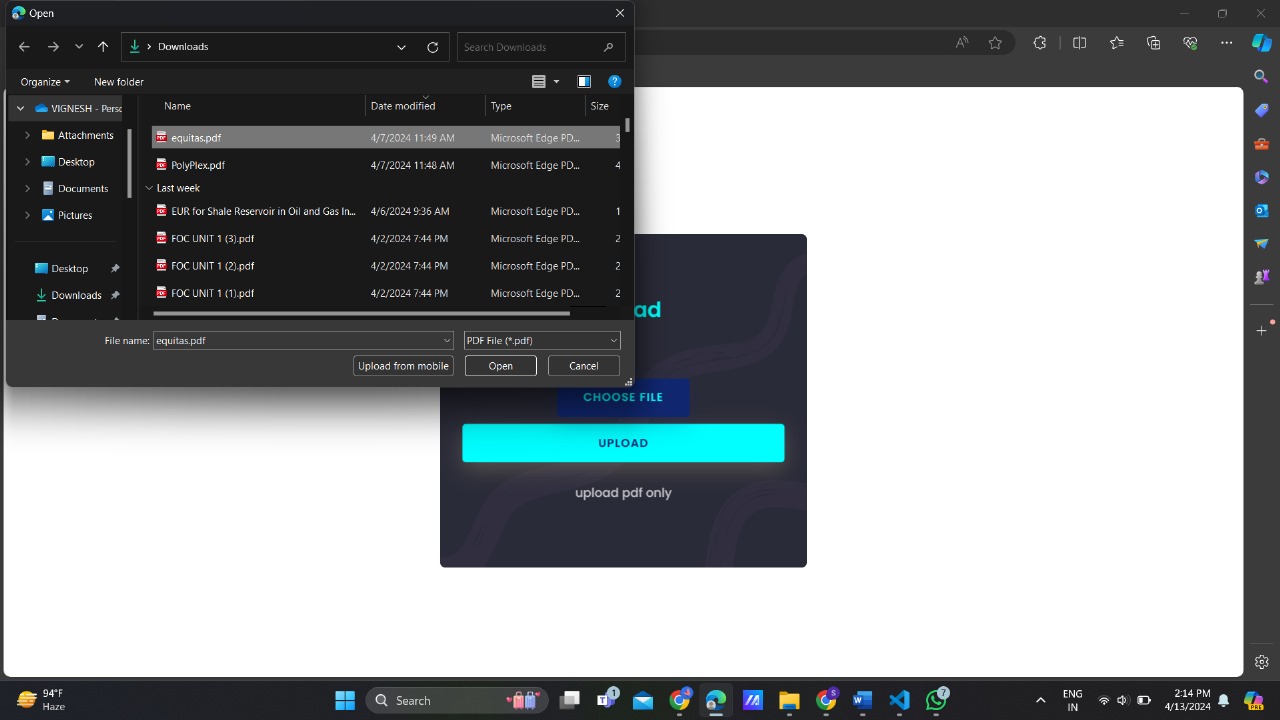
We have uploaded another pdf for the text extraction named vinati.pdf to find the name of the dirrector,din,din type

Output:

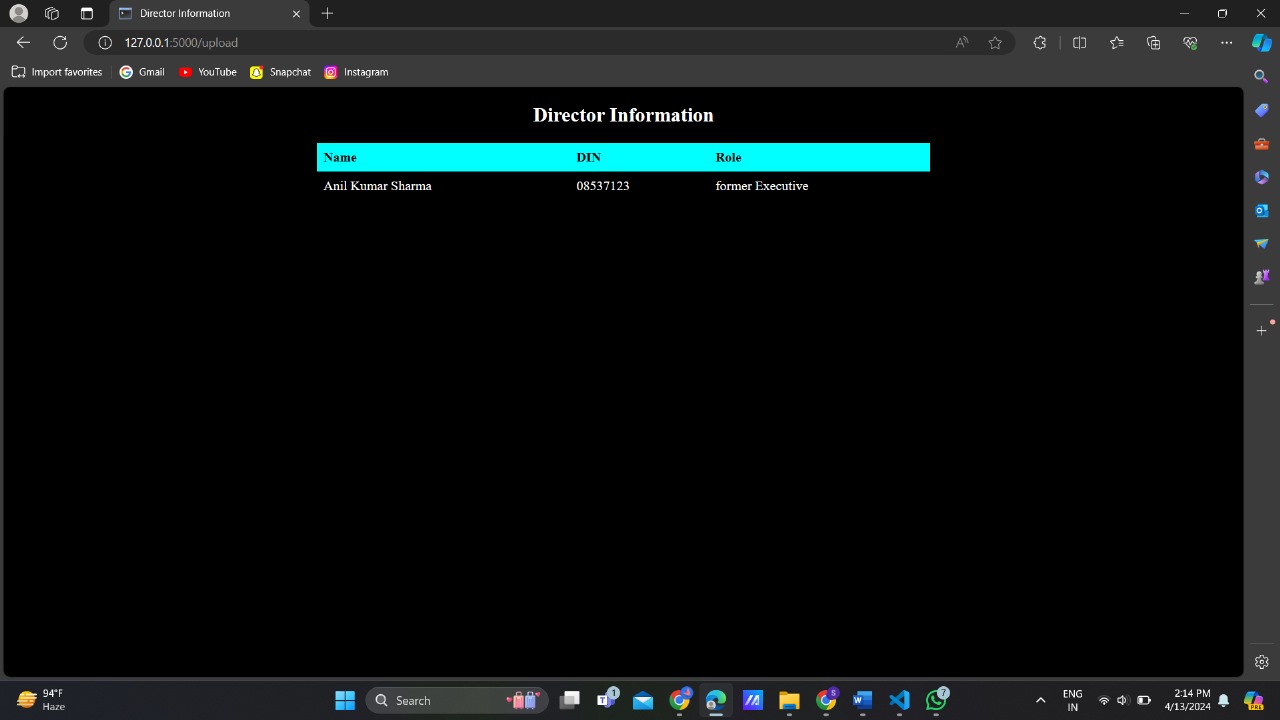


**Observations:**we have the name of the directors as the varun jaipura of din number 02465412 and a whole time director

**Testing on equitas.pdf**



Output:



**OBSERVATIONS:**

In the above picture the flask api gives the outpu of Anil Kumar Sharma of din number :  
08537123 and role of former executive

**8.Conclusion**

In our study showcases the efficency of machine learning in PDF data extraction, showcasing its capacity to revolutionize data retrieval processes. By harnessing ML techniques, we have streamlined operations, ensuring accuracy, and consistency in extracted information. This approach not only saves time but also enhances data quality, minimizing errors inherent in manual extraction methods. The automation facilitated by ML algorithms promises to significantly improve document accessibility for diverse stakeholders across industries. This project underscores the transformative potential of technology in optimizing workflows and unlocking valuable insights from PDF documents. Moving forward, further refinement and integration of ML-driven solutions hold promise for greater efficiency and innovation in data extraction endeavors. Through this endeavor, we contribute to the advancement of automated data processing, empowering organizations and individuals with enhanced capabilities for information retrieval and analysis.