SCOREME SOLUTIONS Hackathon Assignment: Detecting And Extracting Tables From Pdfs

Name: Ridhima

Class: MCA-II

Roll No: 202201032

CONTENTS

- Introduction
- Problem Statement
- Tools and Techniques Used
- Code Walkthrough
- Results
- Conclusion

INTRODUCTION

- This project aims to develop a tool for extracting tables from PDFs without relying on traditional extraction tools like Tabula or Camelot.
- I leveraged **Python** programming along with **PyMuPDF** for PDF parsing and **pandas** for data handling to achieve the extraction goals.

PROBLEM STATEMENT

Developing a tool to extract tables from PDFs and to overcome limitations of handling diverse table formats, ensure data integrity, and enhance document processing efficiency through automated extraction and export to Excel.

TOOLS AND TECHNIQUES USED

1. **Python Programming Language**: Primary language for developing the extraction tool, leveraging its libraries and flexibility.

2. **Pandas**: Utilized for data manipulation, converting extracted table data into structured Data Frames, and preparing data for export to Excel.

3. **PyMuPDF**: Used for parsing PDF documents, extracting text, and analyzing layout information.

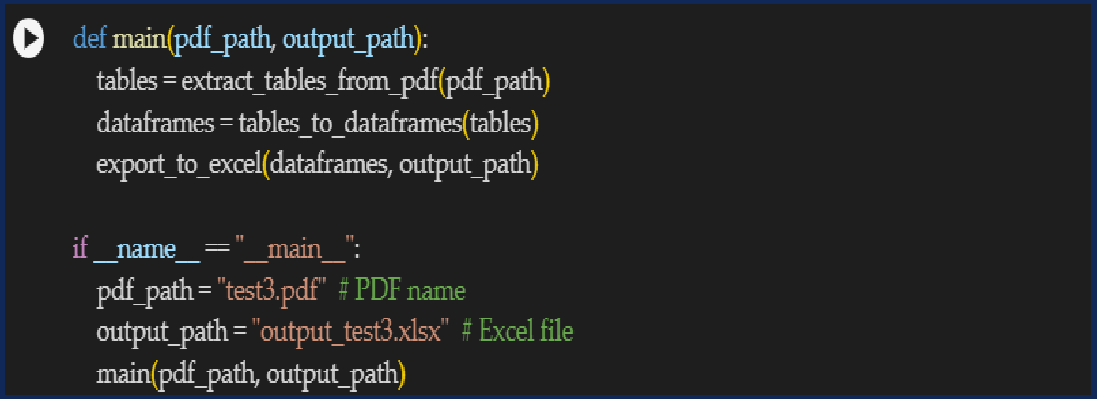
4. **Excel Export (openpyxl)**: Used for exporting extracted tables into Excel format while maintaining data integrity and structure.

5. **Text and Layout Analysis**: Techniques to analyze PDF text blocks, identify table structures, and handle irregularities such as merged cells and multi-line text.

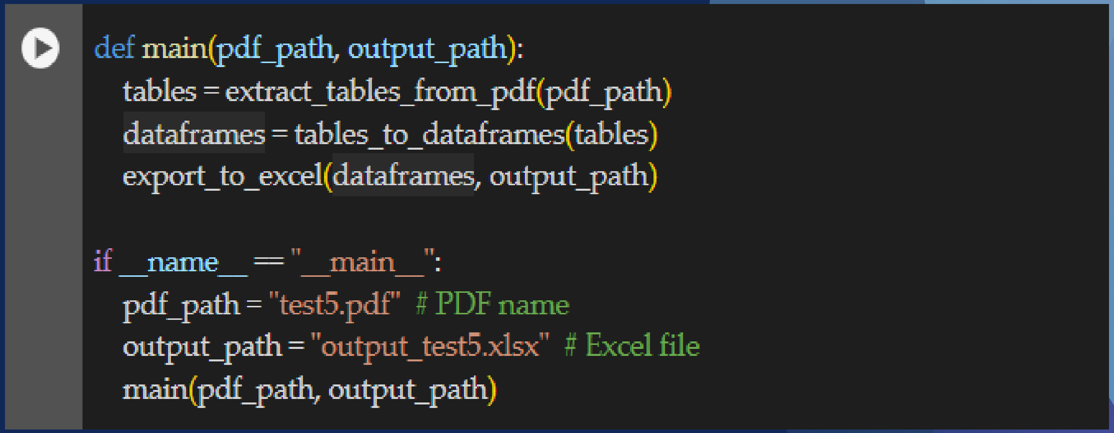
6. **Data Cleaning**: Addressing illegal characters within extracted text to ensure compatibility with Excel formatting requirements.

7. **Iterative Development**: Process involving testing with sample PDFs, refining extraction algorithms, and optimizing performance for efficiency.

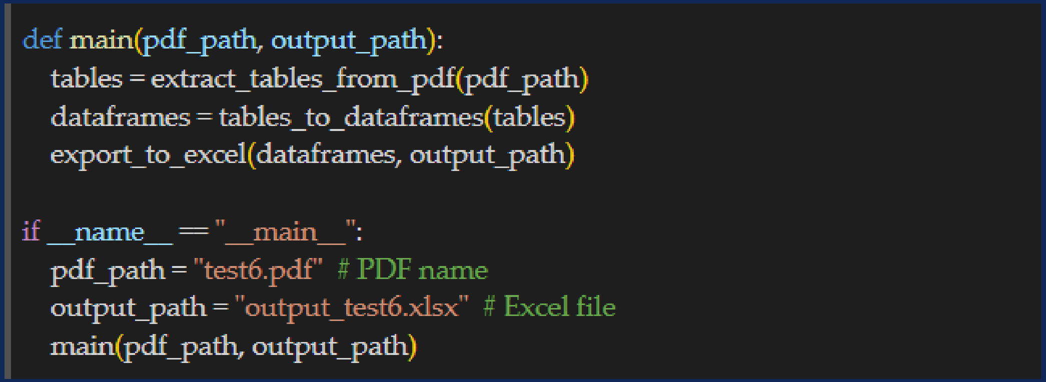
CODE WALKTHROUGH



```python
def main(pdf_path, output_path):
    tables = extract_tables_from_pdf(pdf_path)
    dataframes = tables_to_dataframes(tables)
    export_to_excel(dataframes, output_path)

if __name__ == "__main__":
    pdf_path = "test3.pdf"  # PDF name
    output_path = "output_test3.xlsx"  # Excel file
    main(pdf_path, output_path)
```

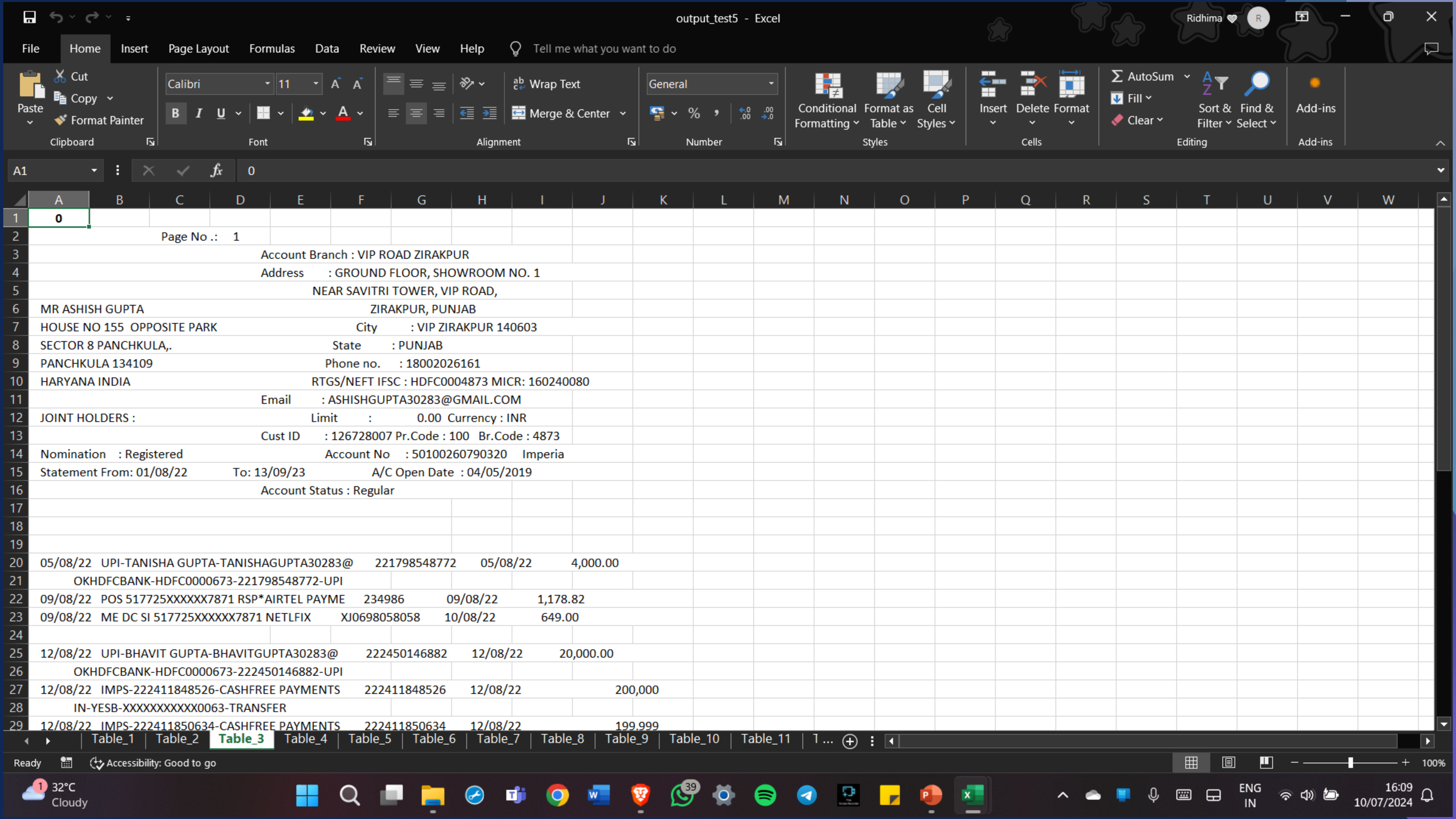


```python
def main(pdf_path, output_path):
    tables = extract_tables_from_pdf(pdf_path)
    dataframes = tables_to_dataframes(tables)
    export_to_excel(dataframes, output_path)

if __name__ == "__main__":
    pdf_path = "test5.pdf"  # PDF name
    output_path = "output_test5.xlsx"  # Excel file
    main(pdf_path, output_path)
```



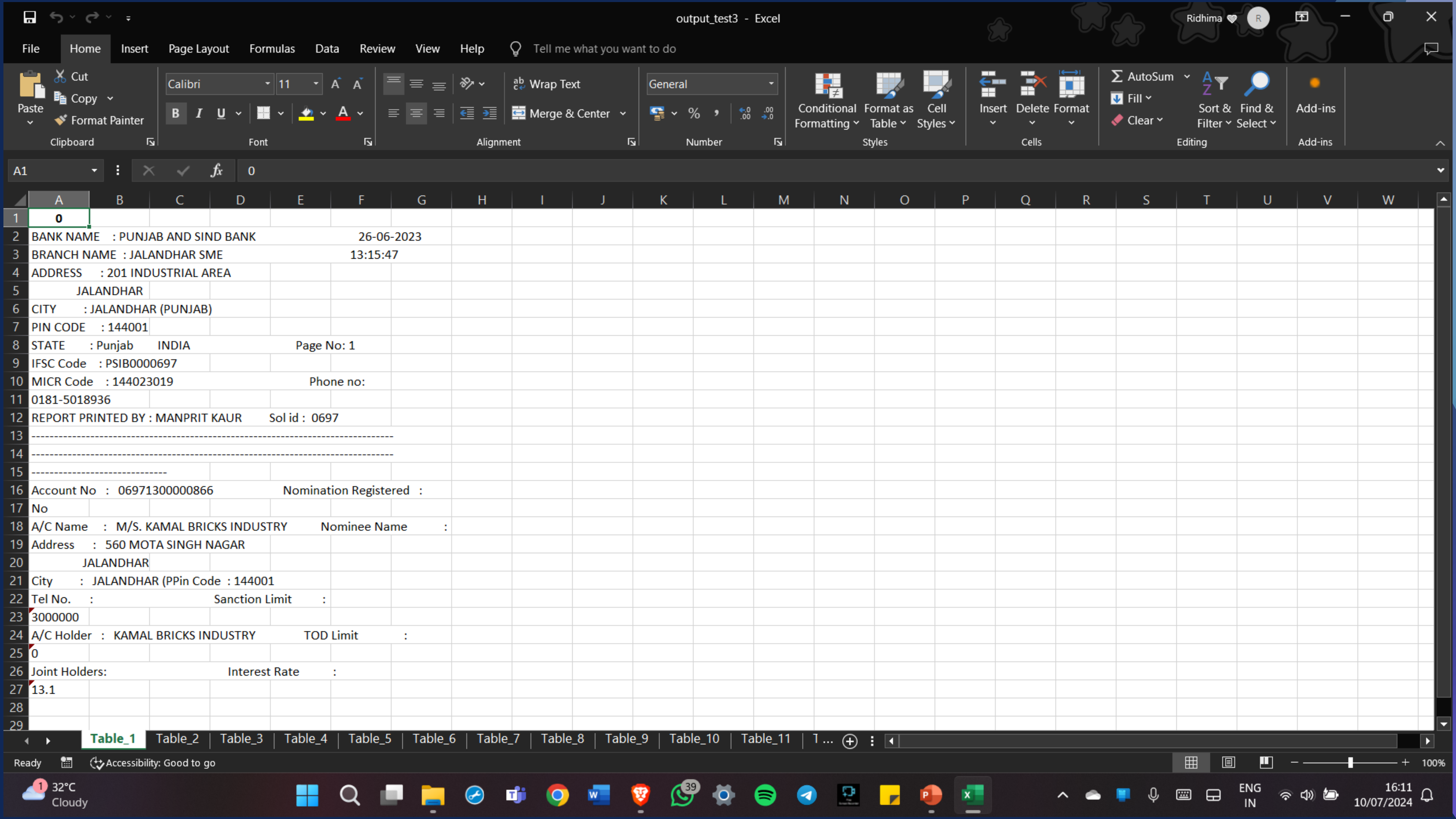
```python
def main(pdf_path, output_path):
    tables = extract_tables_from_pdf(pdf_path)
    dataframes = tables_to_dataframes(tables)
    export_to_excel(dataframes, output_path)

if __name__ == "__main__":
    pdf_path = "test6.pdf"  # PDF name
    output_path = "output_test6.xlsx"  # Excel file
    main(pdf_path, output_path)
```

RESULTS-I



```
0
                    Page No .:   1
                                   Account Branch : VIP ROAD ZIRAKPUR
                                   Address       : GROUND FLOOR, SHOWROOM NO. 1
                                           NEAR SAVITRI TOWER, VIP ROAD,
MR ASHISH GUPTA                                    ZIRAKPUR, PUNJAB
HOUSE NO 155  OPPOSITE PARK                      City          : VIP ZIRAKPUR 140603
SECTOR 8 PANCHKULA,.                           State         : PUNJAB
PANCHKULA 134109                              Phone no.      : 18002026161
HARYANA INDIA                              RTGS/NEFT IFSC : HDFC0004873 MICR: 160240080
                                   Email         : ASHISHGUPTA30283@GMAIL.COM
JOINT HOLDERS :                            Limit         :           0.00  Currency : INR
                                   Cust ID       : 126728007 Pr.Code : 100   Br.Code : 4873
Nomination    : Registered                   Account No     : 50100260790320     Imperia
Statement From: 01/08/22          To: 13/09/23          A/C Open Date  : 04/05/2019
                                   Account Status : Regular



05/08/22  UPI-TANISHA GUPTA-TANISHAGUPTA30283@     221798548772      05/08/22         4,000.00
          OKHDFCBANK-HDFC0000673-221798548772-UPI
09/08/22  POS 517725XXXXXX7871 RSP*AIRTEL PAYME    234986          09/08/22         1,178.82
09/08/22  ME DC SI 517725XXXXXX7871 NETLFIX      XJ0698058058      10/08/22          649.00

12/08/22  UPI-BHAVIT GUPTA-BHAVITGUPTA30283@      222450146882      12/08/22        20,000.00
          OKHDFCBANK-HDFC0000673-222450146882-UPI
12/08/22  IMPS-222411848526-CASHFREE PAYMENTS     222411848526      12/08/22                  200,000
          IN-YESB-XXXXXXXXXXX0063-TRANSFER
12/08/22  IMPS-222411850634-CASHFREE PAYMENTS     222411850634      12/08/22                  199,999
```

RESULTS-II



A1 | 0

```
1   0
2   BANK NAME    : PUNJAB AND SIND BANK                    26-06-2023
3   BRANCH NAME  : JALANDHAR SME                         13:15:47
4   ADDRESS      : 201 INDUSTRIAL AREA
5             JALANDHAR
6   CITY         : JALANDHAR (PUNJAB)
7   PIN CODE     : 144001
8   STATE        : Punjab        INDIA                       Page No: 1
9   IFSC Code    : PSIB0000697
10  MICR Code    : 144023019                                Phone no:
11  0181-5018936
12  REPORT PRINTED BY : MANPRIT KAUR        Sol id :  0697
13  ------------------------------------------------------------------------------
14  ------------------------------------------------------------------------------
15  -----------------------------
16  Account No   :   06971300000866                Nomination Registered   :
17  No
18  A/C Name     :   M/S. KAMAL BRICKS INDUSTRY          Nominee Name          :
19  Address      :   560 MOTA SINGH NAGAR
20              JALANDHAR
21  City         :   JALANDHAR (PPin Code  : 144001
22  Tel No.      :                          Sanction Limit        :
23  3000000
24  A/C Holder   :   KAMAL BRICKS INDUSTRY              TOD Limit             :
25  0
26  Joint Holders:                              Interest Rate         :
27  13.1
28
29
```

CONCLUSION

This project successfully demonstrates a custom Python-based solution for extracting tables from PDFs, addressing challenges of diverse table formats and ensuring data integrity. By leveraging PyMuPDF and pandas, the tool efficiently identifies, extracts, and exports tables into Excel files, enhancing document processing workflows. This approach provides a reliable alternative to traditional PDF table extraction tools, offering flexibility and accuracy in handling complex table structures.

THANK YOU

Name: Ridhima

Class: MCA-II

Roll No: 202201032