**Title: PDF Table Extractor Report**

Name: Pooja Sharma Registration Number: 2022ca059 College: Motilal Nehru National Institute of Technology

Introduction: This report provides an overview of the PDF Table Extractor project, which focuses on extracting tabular data from PDF documents, particularly bank statements. The project aims to accurately extract tables, even when they lack clear borders or consistent formatting, using advanced pattern matching and spatial analysis techniques.

Key Features:

Multiple Text-based Approaches:

* Pattern recognition to identify tables based on whitespace and text structure
* Special handling for financial data like dates and monetary values
* Line pattern similarity analysis to group related rows

Enhanced PDFPlumber Integration:

* Direct extraction of bordered tables
* Custom algorithms for non-bordered tables using spatial analysis
* Word position analysis to detect columns and alignment

Bank Statement Specialized Extraction:

* Custom regex patterns for transaction entries
* Column position identification by analyzing whitespace patterns
* Automatic detection of transaction date formats

Position-based Table Detection:

* Analyzes word coordinates to identify columns without relying on visual borders
* Clusters words with similar positions to detect table structure
* Validates potential tables using data type consistency checks

Libraries Required:

* pip install pandas PyPDF2 pdfplumber numpy

Conclusion: This solution is particularly well-suited for extracting tabular data from bank statements and financial documents. The combination of pattern recognition and spatial analysis allows for accurate table extraction, even when traditional methods fail.