Library-pyPdf
for page separating,
fitz used as PDF editor/viewer. And pytesseract used as OCR for extracting text. First converts the pdf into image and then extracts the text.
The text is written into different text files. (one txt file for one page).

Library- fitz,pytessract
Findings:

- Blurred image doesnot give accurate text output.(Page 1)
- Text immedately before image and after the image is not detected.
- Page number is not written in any pages.
- Minor mistakes in well formatted pages of text. For example “academic” is read as “academe”. Frequency of such spelling error is 2-4 %.
- Signature and Designation is not detected.
- Vertical handwriiten line in pg4 is detected but with very poor accuracy.
- Roman numerals are not detected properly. Roman numerals are often read as english alphabets to fit into a word.
- Pages with different background than whole page is not detected properly.
- It can distinguish between different columns of text and arranges them in the correct order.
- Text which are not exactly horizontal or vertical are not detected.
- No headings are detected (probably because they are not in the same background as the rest of the page).
- The text which is continous and selectable is detected in an image.
- While reading italics font(in sanskrit language) the accuracy is about 60%.
- Poor results for images. The labelling of different things is not detected.

Observations for page 13(table)

- Initially the text file has 35 blank lines.
- The is read very poorly and only some words are understandable.
- Text is read satisfactorily with 2-4% error.

- Table which is not separated by lines of rows and columns is read quite satisfactorily.

Library – Doctr
Findings:

- Very high rate of accuracy with almost no spelling mistakes (less than 1% error).
- Can also identify blurred text and text of images
- Very high accuracy in reading the data from tables including numbers.
- Can read roman numerals and is unaffected by different background color.
- Takes around 10 X more time than pytesseract.
- Only flaw is that it reads row wise. So text formatted in different columns is read altogether in one line.
- Orientation sensitive . Only the horizontal lines are read.

Library – Easyocr
Findings:

- Very low accuracy. Lots of speeling mistake. Error percentage of 20-30%.
- Only text with large font size are recognized properly.
- Poor results for images and tables as well.