The most important of the given problem is data.

1. As we have scanned pdf documents, I have converted the pdf to image then image to text.
2. We will make a csv file containing text , form type for given pdf to make it supervised data.
3. We have only converted 1st page of pdf to text because form type will be displayed on 1st page only as per the data.
4. We will use regex to get the form type from the textual data.
5. We will save the dataframe to csv file so that we don't have to do above exercise again and again.
6. Once we have gathered data, we will make all the text and form type in lower case .
7. After making it lower the for form type I have corrected the wrongly classified form type and make it correct.
8. Then we will use label encoder to transform word to numbers for ML.
9. On text column of data, we will split it based on white space and remove stop words.
10. After removing stopwords we will use snowball stemmers to stem the data.
11. After all this we will use tfidf to vectorize the data. We will use l2 normalization in tfidf to normalize the data.
12. After all this we will use SVM for data modelling and prediction.
13. We use SVM because SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional spaces and relatively memory efficient.

NOTES:-

- While we convert pdf to text there is a memory leak that need to be fixed.
- We can make an API for people to classify the document to right type that's why we have saved the model we just need to write a small function for this.