Resume-categorization prediction pipeline:

Dataset Analysis → Data Preprocessing → Data Exploration → Data Cleaning → Model Training → Evaluation

1. Dataset Analysis

- df. head() - To get a glimpse of the dataset which will print the first 5 rows of the data by default
- df. isnull().sum() - Checks to if there are any null values in the columns
- df['Category'].value_counts() - To check the total number of categories that exist in the dataset
- df**.** drop(columns **=** ['ID', 'Resume_html'], inplace **= True**) - To drop the meaningless columns that might negatively impact the model's performance

2. Data preprocessing

- text**.** lower() - To convert into lowercase letters
- text **=** re**.** sub('[^a-zA-Z]',' ', text) - By using the regular expression, remove integer, punctuations
- Tokenization using word tokenizer
- Remove stop words from the corpus
- Stemming to convert words into the base form

3. Data Exploration

- Visualization using matplotlib

4. Data Cleaning

- Removing punctuation and stop words from the corpus using nltk libraries
- Splitting the datasets into Train_Validation_Test

5. Model Training

- RandomForestClassifier - It's a baseline model to quickly assess the performance of the model. The model helps to provide relatively simple interpretability with limited data.
- GridSearchCV - Systemically search for the optimal combination of hyperparameters for a given model

6. Model Evaluation

- Accuracy
- Classification report - Consists of precision, F1-score, recall