|  |  |
| --- | --- |
| JOB DESCRIPTOR DATASET:  ===============================================  1. Content  [Real or Fake] : Fake Job Description Prediction  This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn the job descriptions which are fraudulent.  2. About this file   * job\_id- Unique Job ID * title-The title of the job ad entry. * Location-Geographical location of the job ad. * Department-Corporate department (e.g. sales). * salary\_range-Indicative salary range (e.g. $50,000-$60,000) * company\_profile- A brief company description. * Description-The details description of the job ad. * Requirements-Enlisted requirements for the job opening. * Benefits-Enlisted offered benefits by the employer. * Telecommuting-True for telecommuting positions. * has\_company\_logo-True if company logo is present. * has\_questions-True if screening questions are present. * employment\_type- Full-type, Part-time, Contract, etc. * required\_experience- Executive, Entry level, Intern, etc. * required\_education- Doctorate, Master’s Degree, Bachelor, etc. * industry-Automotive, IT, Health care, Real estate, etc. * function- Consulting, Engineering, Research, Sales etc. * fraudulent- target & Classification attribute.   3. PROBLEM STATEMENT  The dataset is very valuable as it can be used to answer the following questions:   1. Create a classification model that uses text data features and meta-features and predict which job description are fraudulent or real. 2. Identify key traits/features (words, entities, phrases) of job descriptions which are fraudulent in nature. 3. Run a contextual embedding model to identify the most similar job descriptions. 4. Perform Exploratory Data Analysis on the dataset to identify interesting insights from this dataset. 5. Run a contextual embedding model to identify the most similar job descriptions. 6. Perform Exploratory Data Analysis on the dataset to identify interesting insights from this dataset.   4. ROUGH APPROACH   * NLP used for feature extraction, text mining * SL CLASSIFICATION MODELS also can be used LogR, RF, NB with PIPELINES |  |