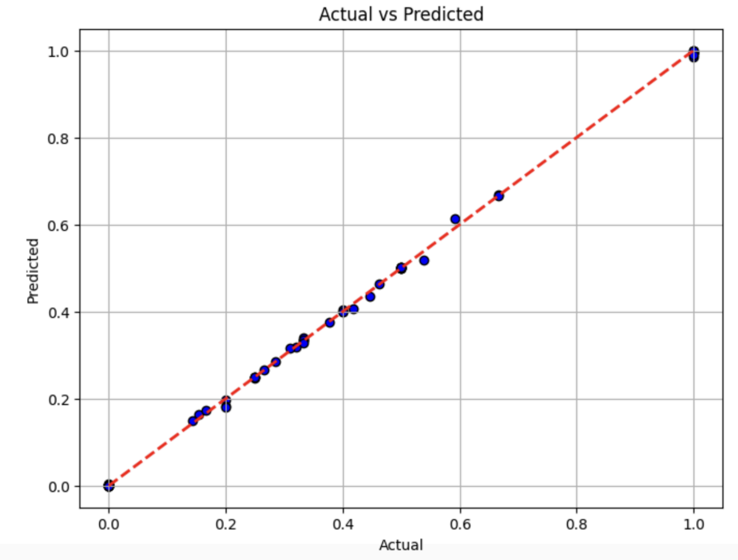
Our initial approach to handling real-world data begins with data extraction using large language models (LLMs), which enables us to efficiently process and prepare large datasets. This step ensures that the data is well-structured for subsequent analysis and contributes to improving the overall accuracy of our predictive models.

For both labeled and unlabeled datasets, we apply a consistent prompting strategy to generate CSV files with the same format. In the case of labeled data, we primarily use Gemini AI to generate all required columns, including gender, position, company, responsibilities, and so on. Eight extra columns are created for further research.

For the unlabeled data, since the position column is already provided, we focus on generating the remaining seven columns.

Subsequently, we split the dataset into training and test sets to evaluate the model's performance on unseen data. We tokenized all relevant columns, with special focus on the job description, converting the text into numerical features using TF-IDF vectorization. Categorical columns such as company name, position, language, education, experience, and personality were transformed using one-hot encoding to represent them as binary features. The numeric column, women proportion, was included as is. We then combined all these features into a single dataset. Using Scikit-learn's machine learning tools, we trained a model and evaluated its accuracy. The initial results are promising, demonstrating the effectiveness of our data preprocessing and feature engineering approach.

The model and analysis are implemented in the notebook **ML_HR.ipynb**.



The model code is ML_HR .ipynb