**Report**

1. **Introduction:**
   * Field of Research: This research focuses on analyzing the job market in Egypt using data extracted from Wuzzuf.net, one of the largest job portals in Egypt.
   * You can find it here : <https://wuzzuf.net/jobs/egypt>
2. **Data Description:**
   * Types of Data:
     + Text (job title, required skills, company names, country, city, area, posted)
     + Numbers (number of experience year)
     + Classifications (job category, job state , exp level posted , job type)
   * **Data Size:**
     + 5193 rows
3. **Methodology:**
   * Data Cleaning:

A screenshot of a computer

Description automatically generated

- we noticed that there are data types need to be converted and skills, exp years have null values

**1-** **Unnamed column:** unimportant column so we dropped it.

**2-** **job title:** we noticed that the unique values of job title are large about 3206 so we tried to reduce it by adding new column called **job category** that contains 17 unique values, it groups the similar jobs in specific category, then we dropped the job title column.

**3-** **company name:** it is cleaned but it has too much unique values, we can check how many companies that is not popular or post only one job, it seems a lot (1037).

**4-** (**area column – city – country – job type – job state – exp level):** a cleaned is well

**5-** **Skills:** we noticed that the unique values of skills which is too large about (3920 Row) so we tried to convert skills of each job into list and named the column **new skills**, then we extract from new skills **frame works** and **programming language** Columns.

**6-** **exp years:** the data type needs to be handled also there are null values we used a function to convert it to int first (by calculating avg of each exp year) and we used a KNN imputer to fil the null values, then we stored it in (Number of Experience Year) and dropped the original column.

**7- posted:** we noticed that the units of posted vary from minutes, hours and months so we decide to convert all the units to days then convert the days to a date column by calculating the difference between the date of today and the number of days we calculated before storing it in a new column called **posting date** and dropping the original one.

**In the end** we found that **(frameworks -Programming language)** contain null value, so we fill null value with **(not found - others)**

1. Cleaning and Analysis Methods:
   1. We used pandas library for manipulating data.
   2. KNN Imputer for null values.
   3. Mean, Median and STD with EXP Years.
   4. Group by method with Categorical column.
   5. Time series with posted date.
2. **Analysis Results:**

* **Visualizations:**

A screenshot of a computer

Description automatically generated

In addition to using Tableau for creating the dashboard, we also used the following plots: **(bar plot – count plot – hist plot – Facet Grid – Line plot)** to answer our question we wrote in notebook.

* **Interpretation of Results:** 
  + 1. the most popular country in the dataset is Egypt.
    2. the most common job categories in the data are management and leadership and it comes in second place in the dataset and Egypt, United Arab Emirates.
    3. the most common job categories in the data are Data entry in the Saudi Arabia.
    4. The most company hiring is Confidential in the Dataset, Egypt and Saudi Arabia.
    5. The most company hiring is Envirogent in the United Arab Emirates.
    6. most of jobs available in each city in Egypt is Cairo and Giza, it is normal.
    7. most of jobs available in each city in Saudi Arabia is Riyadh.
    8. most of jobs available in each city in United Arab Emirates is Dubai.
    9. most of job types in Egypt, Saudi Arabia and United Arab Emirates is Full time Job.
    10. most job states in (Dataset and Egypt) is On-Site, and in (Saudi Arabia and United Arab Emirates) not mentioned.
    11. most of experience level in Dataset, Egypt and United Arab Emirates is Experienced, and in Saudi Arabia Not specified.
    12. The most frequent Number of Experience Year is 5 in Dataset.
    13. Mean of years of experience in Dataset and (Egypt is 6 years for management and leadership job) and in (United Arab Emirates needs 6 years for Hr professional).
    14. As we increase the number of experience years the experience level increase and it is logically true.
    15. R is the most programming language worked with in the Dataset.
    16. React is the most framework worked with in the Dataset.
    17. In the June month there is more than 700 jobs were posted, The sudden drop around July 1 could be due to a specific event or change in the job market, such as the end of a hiring season.

1. **Conclusions:**
   * Based on the analysis of job market data from Wuzzuf.net, several key insights have been identified, each carrying significant implications for job seekers, employers, and policymakers.
   * **Geographical Distribution:**
     + 1. **Insight:** Egypt leads in job postings, with Cairo and Giza being top cities; Riyadh and Dubai are key cities in Saudi Arabia and UAE.
       2. **Implication:** Job seekers should focus on these cities; employers and policymakers should enhance infrastructure and workforce development.

* **Job Categories:**
  + 1. **Insight:** Management and leadership roles are most common in Egypt and UAE; data entry jobs dominate in Saudi Arabia.
    2. **Implication:** Job seekers should align skills with high-demand categories; training programs should match market needs.
* **Company Hiring Trends:**

1. **Insight:** Confidential companies are top employers in Egypt and Saudi Arabia; Envirogent leads in UAE.
2. **Implication:** Job seekers should target these companies; employers should monitor competitors.

* **Job Types and States:**
  1. **Insight:** Full-time jobs are most common; on-site roles prevail in Egypt.
  2. **Implication:** Job seekers preferring full-time and on-site roles have many opportunities; employers may consider remote work options.
* **Experience Levels and Requirements:**
  1. **Insight:** Experienced roles dominate, typically requiring 5-6 years of experience.
  2. **Implication:** Job seekers should gain relevant experience; employers should set experience requirements in line with industry standards.
* **Skills Demand:**
  1. **Insight:** R and React are the most in-demand skills.
  2. **Implication:** Job seekers should develop these skills; training institutions should focus on them.
* **Posting Trends:**
  + 1. **Insight:** Job postings peaked in June, dropped sharply in July.
    2. **Implication:** Job seekers should be active during peak seasons; employers should align recruitment cycles accordingly.
* **Recommendations**
  + 1. **For Job Seekers:**

1. Develop high-demand skills.
2. Tailor applications to job categories.
3. Consider relocating to high-opportunity cities.

* **For Employers:**

1. Understand hiring trends.
2. Offer remote work options.
3. Align recruitment with market trends.

* **For Policymakers:**

1. Enhance workforce programs.
2. Improve infrastructure in key cities.
3. Support training in essential skills.

**6.** **Job Recommendation System:**

* **Introduction:**

1. This report documents the process for a Job Recommendation System implemented using Streamlit. The system provides a user-friendly interface where individuals can input their skills, experience level, desired job category, and preferred country. Based on these inputs, the system generates personalized job recommendations using a pre-trained machine learning model.

* **libraries and frameworks are used in the model:**

1. **Streamlit**: For building the web application interface.
2. **Pandas**: For handling and manipulating the dataset.
3. **Scikit-learn**: Used for machine learning, specifically for TF-IDF vectorization and classification.
4. **PIL (Python Imaging Library)**: For image processing.
5. **NumPy**: For numerical operations, particularly in calculating similarities.
6. **Joblib**: For saving and loading the machine learning model.

* **Data Preprocessing:**

1. **Feature Engineering:** TF-IDF Vectorization: The skills and job descriptions are vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This transforms the text data into numerical vectors that represent the importance of each term in the context of all job postings.
2. **Label Encoding:** Categorical variables, such as job category and country, are encoded into numerical labels using Label Encoding. This process is crucial for machine learning algorithms that require numerical input.

* **Model Selection:**

The system uses a machine learning model to predict the most relevant job categories based on user input. The choice of model and its training process are detailed below:

1. **Model Choice:** A **Random Forest** classifier is selected for this task due to its robustness, accuracy, and ability to handle high-dimensional data. Random Forest is an ensemble method that combines multiple decision trees to improve predictive performance and reduce overfitting.
2. **Training the Model:** The model is trained on the preprocessed dataset:

**Input Features**: The TF-IDF vectors representing job descriptions and skills.

**Target Labels**: The job categories encoded as numerical labels.

**Model Training**: The Random Forest classifier is trained using the labeled data, learning to map the input features to the corresponding job categories. The model benefits from the ensemble approach, where multiple trees are built during training, and the final prediction is made by averaging the predictions of all the trees.

* **Get Recommendations Function:**

1. This function integrates the model and data preprocessing steps to generate job recommendations based on user input. It:
2. Converts user input into a TF-IDF vector.
3. Predicts the probability of each job category.
4. Filters job postings by category and country.
5. Calculates the similarity between the user's profile and each job's features.
6. Returns the most relevant job postings.

* **Streamlit App**

The final step is integrating the model into a user-friendly web application using Streamlit. The app provides an interactive interface for users to input their profile and receive job recommendations.

* + **How to use:**
  1. **Run the Streamlit App**: Use the command streamlit run model.py to launch the application.
  2. **Input Your Information**: Fill in the skills, experience level, job category, and country in the input fields.
  3. **Get Recommendations**: Click on the "Get Recommendations" button to see a list of jobs tailored to your profile.
* **Conclusion**

The Job Recommendation System leverages natural language processing and machine learning to match users with jobs that fit their profile. By integrating a logistic regression model into a Streamlit app, the system provides an intuitive platform for users to explore job opportunities tailored to their skills and experience.