**Lovely Professional University**

**(CSE 355)**

**Title:-**  **Job Descriptions Analysis**

**(Project Term January-May 2025)**

**Section: - K22UP**

### A black and orange logo Description automatically generated

Lovely Professional University

Jalandhar, Punjab, India.

|  |  |
| --- | --- |
| Course Code: CSE 355 | Course Title: MACHINE LEARNING PROJECT |
| Student Roll No: RK22UPB66 | **Student Name: Rahul Chand** |
| Academic Task No: 1 | **Academic Task Title**: Job Description Analysis using ML |
| Date of Allotment: 07th May 2025 | **Date of Submission**: **07th May 2025** |
| Term: 324252 | **Section:** K22UP |
| Evaluation Parameters | **Student Reg.No**: 12211893 |

**Acknowledgment**

I would like to express my heartfelt gratitude to my mentor, **Himanshu Gajanan Tikle**, for her unwavering support, insightful guidance, and expertise throughout this project. Her mentorship has been invaluable in shaping my understanding of Data analysis, EDA, Feature engineering and ML model integration. From conceptualizing the problem statement to implementing advanced techniques, her constructive feedback and deep technical knowledge have played a pivotal role in refining my approach and improving the overall quality of this work.

I am also grateful to the various learning platforms, including GeeksforGeeks, CampusX, and UpGrad, whose structured courses and well-documented resources provided essential knowledge and practical insights. Lastly, I extend my appreciation to my peers and the data science community, whose discussions and shared experiences have contributed significantly to my learning journey.

**Table Of Contents**

|  |  |
| --- | --- |
| Topics | Page No. |
| 1. Introduction | 4 |
| 1. Problem Understanding and Definition | 5-7 |
| 1. Dataset Selection & Preprocessing | 8-9 |
| 1. Feature selection | 10-12 |
| 1. Exploratory Data Analysis (EDA)   & Visualization | 13-16 |
| 1. Feature Engineering | 17 |
| 1. Clustering Analysis | 18-19 |
| 1. Prediction Modeling | 19-20 |
| 1. Conclusion and future Work | 21-22 |
| 1. Appendix | 23 |

1. **Introduction**

**1.1 Project Overview**

The aim of this project is to analyze large-scale job posting data to gain insights into salary trends, job classifications, qualifications, and potential fraud detection. By leveraging machine learning, we build predictive models to estimate salary ranges and cluster job roles based on textual job descriptions.

**1.2 Dataset Description**

The dataset is collected from Kaggle and consists of over 1.6 million job postings across different regions and industries. It includes fields like job title, company profile, experience, salary range, location, work type, and skills. The data is used to perform exploratory analysis, text parsing, and machine learning modeling.

**Sample Fields:**

* Job Title
* Experience (e.g., "2 to 5 Years")
* Salary Range (e.g., "$60K-$80K")
* Company Profile (JSON string with nested info)
* Work Type (Full-time, Part-time, Remote)
* Location (City, State, Country)
* Skills and Job Description (Free-text)

**1.3 Goals**

* Predict average salary using regression models.
* Cluster job descriptions using text-based similarity.
* Explore patterns in job demand, qualification impact, and salary distribution.
* Engineer meaningful features for predictive modeling.

**2. Problem Understanding and Definition**

**1.1 Clarity of Problem Statement**

The job market is vast and highly dynamic, with salaries, required qualifications, and job roles evolving rapidly. Organizations post thousands of job descriptions across different platforms, making it challenging for job seekers and recruiters to analyze trends effectively. The goal of this project is to build a data-driven solution that provides insights into salary predictions, job role classification, and potential fraud detection in job postings.

With millions of job seekers searching for employment opportunities every year, it is essential to have a system that can analyze job trends efficiently. Many applicants struggle to determine appropriate salary expectations based on their qualifications and experience. Additionally, employers often face difficulties in filtering candidates and ensuring that job postings reach the right audience. In recent years, fraudulent job postings have also become a significant concern, misleading applicants and causing financial and psychological distress.

By leveraging data science and machine learning, we aim to create a system that can predict salary ranges, categorize job roles, and flag potential fraudulent job postings. This will provide value to both job seekers and recruiters, enabling them to make informed decisions based on real-time data insights.

**1.2 Justification for Solving the Problem**

A well-defined job market analysis system benefits both job seekers and recruiters in several ways:

* **For Job Seekers:**
  + Helps applicants set realistic salary expectations based on industry trends.
  + Assists in identifying job roles that match their qualifications and experience.
  + Provides insights into job locations, work types, and company size distribution.
* **For Recruiters & Employers:**
  + Helps in identifying the ideal salary range for job postings to attract qualified candidates.
  + Improves hiring decisions by analyzing historical job data and industry-specific requirements.
  + Detects fraudulent job postings, preventing scams and improving recruitment platform reliability.
* **For Market Researchers & Analysts:**
  + Helps in analyzing employment trends across different industries, countries, and job roles.
  + Provides valuable insights into the changing nature of work (e.g., remote vs. onsite jobs).
  + Assists policymakers in understanding wage distribution and employment gaps.

With the increasing reliance on digital job platforms, automating job description analysis using machine learning can provide significant value by reducing manual effort and improving decision-making. By implementing predictive models, we can improve hiring efficiency, minimize fraud, and create a more transparent job market for all stakeholders.

**1.3 Defined Objectives & Hypotheses**

**Objectives:**

The primary objectives of this project are:

* Develop machine learning models to predict salary ranges based on job descriptions.
* Classify job postings into relevant categories based on qualifications, experience, and work type.
* Detect potential fraudulent job postings using anomaly detection techniques.
* Identify factors that influence salary variations across industries, job roles, and locations.
* Provide insights into employment trends based on company size, work type, and geographical location.

**Hypotheses:**

Based on industry trends and preliminary analysis, we propose the following hypotheses:

1. **Salary vs. Experience:** Jobs requiring higher qualifications and greater experience tend to have higher salary offers.
2. **Industry-Based Salary Variations:** Certain industries (e.g., technology, finance) offer significantly higher salaries compared to others (e.g., retail, hospitality).
3. **Work Type Influence:** Full-time jobs generally offer higher salaries compared to part-time and contractual roles.
4. **Location Impact:** Jobs in metropolitan areas tend to have higher salaries than those in smaller cities or rural areas.
5. **Fraudulent Job Patterns:** Fraudulent job postings often exhibit anomalies such as unrealistic salary offers, vague job descriptions, or missing employer details.
6. **Company Size & Salary Relationship:** Large companies tend to offer more competitive salaries and benefits than small or medium-sized enterprises.

Understanding these relationships will allow us to design effective predictive models that benefit job seekers, employers, and researchers alike.

**1.4 Libraries used**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**3. Dataset Selection and Preprocessing**

**3.1. Dataset Relevance and Quality**

The dataset used for this project is sourced from publicly available job postings and recruitment platforms. It includes a wide range of job descriptions spanning various industries, company sizes, and geographical locations. The dataset consists of over **1.6 million records and 23 features**, making it extensive and suitable for machine learning applications.

The dataset provides valuable insights into salary trends, job requirements, and employer expectations. By using a large and diverse dataset, the model can generalize well across different job roles and regions, making predictions more reliable and applicable to real-world scenarios.

**3.2. Handling Missing Values**

Data preprocessing is crucial in ensuring the reliability of the model. Many real-world datasets contain missing values, which must be handled appropriately. In this dataset, missing values were identified in the Salary Range, Experience, and Qualifications columns. The following strategies were used to handle them:

* **Numerical Features:** Missing values in Min Salary, Max Salary, and Experience were imputed using the median value of each respective feature.
* **Categorical Features:** Missing values in Qualifications, Work Type, and Company Size were imputed using the mode (most frequent value) to maintain consistency.
* **Removing Unnecessary Columns:** Features such as Contact Details and Contact Person were dropped as they were irrelevant to predictive modelling.

**3.3. Handling Outliers**

Outliers can significantly affect model performance, especially for regression-based tasks. Outlier detection was performed using:

* **Interquartile Range (IQR) Method:**
  + Values lying below Q1 - 1.5\*IQR or above Q3 + 1.5\*IQR were considered outliers.
* **Z-score Analysis:**
  + Data points with Z-score values greater than 3 or less than -3 were treated as extreme values and either capped or removed.

**3.4. Encoding Categorical Variables:**

Machine learning models require numerical inputs, so categorical variables were encoded as follows:

* **Label Encoding:** Used for Job Title and State features to retain ordinal relationships.

**3.5. Log Transformation**

**Log Transformation**: I applied np.log1p to y\_train and y\_test to reduce skewness in 'Avg Salary'. This is a transformation, not normalization. It doesn’t scale the target to a specific range (e.g., [0, 1]) or standardize it (mean 0, std 1).

**Features (X\_train, X\_test)**: No normalization or scaling was applied to the features. I didn’t suggest using MinMaxScaler, StandardScaler, or any other scaling method for features like 'Avg Experience'.

**Conclusion**: No normalization was applied here. Log transformation was used on the target, but that’s a different preprocessing step.

A screen shot of a computer program

AI-generated content may be incorrect.

**3.6. Splitting Data into Training & Testing Sets**

A computer screen shot of a program code

AI-generated content may be incorrect.

Before training machine learning models, the dataset was split into training and testing subsets:

This ensures that the model is evaluated on unseen data to test its generalization ability.

* 1. **Feature Selection & Engineering**

1. A blue and red squares with white text

   AI-generated content may be incorrect.**Feature selection**

Feature selection is a crucial step in improving model performance by eliminating irrelevant or redundant features. For this project, we utilized multiple methods to identify the most influential features:

* **Correlation Matrix Analysis:** A heatmap was generated to visualize the relationships between numerical features. Features with high correlation (above 0.85) were reviewed to prevent multicollinearity.
* **Recursive Feature Elimination (RFE):** This technique was applied using a Random Forest model to rank feature importance and eliminate the least significant variables.
* **Chi-Square Test:** For categorical features, the chi-square test was used to determine their dependency on the target variable, ensuring that only statistically significant features were retained.
* **Variance Thresholding:** Features with very low variance were removed to avoid adding noise to the model.

Based on these analyses, the final set of selected features included:

* Job Title
* Qualifications
* Experience
* Work Type
* Min Salary, Max Salary, Avg Salary
* Company Size
* Location
* Industry Sector

1. **Feature Engineering**

To enhance the dataset’s predictive power, additional features were engineered to provide deeper insights into job descriptions and salary expectations:

**1. Salary Normalization**

Since salary ranges were provided in different formats, a new feature Salary Per Year of Experience was created:

A computer screen with text on it

AI-generated content may be incorrect.

This allows better comparisons between jobs requiring different levels of experience.

**2. Experience Levels**

Experience was categorized into different levels for better interpretability:

A screen shot of a computer program

AI-generated content may be incorrect.

This feature helps in understanding how job roles are distributed across experience levels.

**3. Company Size Encoding**

Companies were classified into size categories to analyze salary trends:

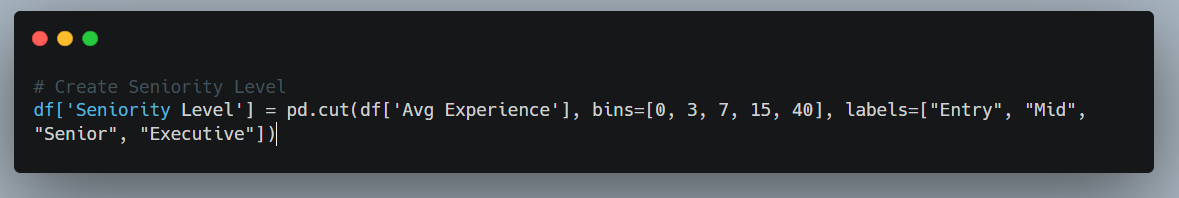
A black rectangular object with white text

AI-generated content may be incorrect.

This numerical encoding ensures better handling of categorical values in ML models.

**4. Seniority Level**

A feature was created to create seniority level:



Higher seniority represents higher the Experience.

By selecting and engineering these features, the dataset was optimized for building effective machine learning models that provide meaningful insights into the job market.

* 1. **Exploratory Data Analysis (EDA) & Visualization**

**5.1. Overview**

Exploratory Data Analysis (EDA) was a critical step to understand the underlying patterns, trends, and anomalies in the job postings dataset. The dataset comprised over 1.6 million entries with 23 features including job descriptions, qualifications, salary range, company details, and more. The following EDA techniques and visualizations were used to derive insights:

* 1. **Null Values and Duplicates**
* Null Values: Features such as Company Profile, Salary Range, and Experience contained null values. These were addressed using imputation techniques:
  + Categorical features were imputed using mode.
  + Numerical features were imputed using median.
* Duplicates: Duplicate job postings were checked using Job Id. Although original Job Ids were long and cumbersome, new unique IDs were assigned for better handling.

**5.3 Outlier Detection**

* The Interquartile Range (IQR) method was used to identify outliers in numerical fields like salary and experience.
* Over 10,000 outliers were detected and analyzed. Some were capped, while others were retained to maintain diversity.

**5.4 Temporal Analysis**A graph showing a number of blue bars

AI-generated content may be incorrect.

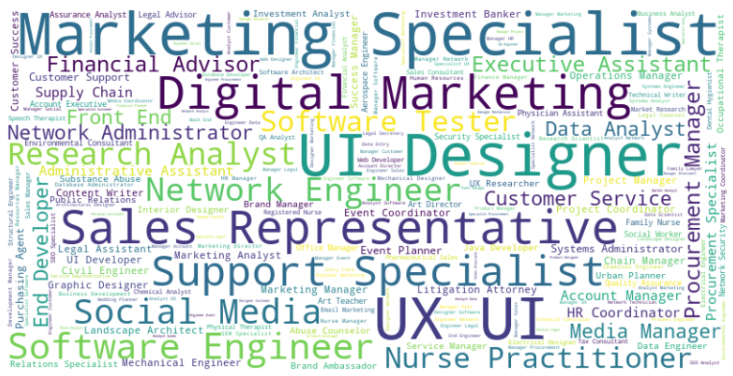
* Job Postings Over Years: Extracted the year from Job Posting Date and visualized annual trends.
* Insight: A significant surge in job postings was observed post-2020, indicating digital transformation and remote job proliferation.

**5.5 Sector-wise Analysis**

* A stacked bar chart was used to analyze how job postings were distributed across various sectors over the years.
* Insight: Sectors like IT, Healthcare, and Finance dominated recent years, while Manufacturing and Retail declined post-2019.

A chart of different colored squares

AI-generated content may be incorrect.

**5.6 Word Clouds**

* Generated word clouds for the following columns:
  + Job Title: Common roles included Software Engineer, Data Analyst, and Project Manager.
  + Skills: High-frequency terms included Python, SQL, Communication, and Leadership.
  + Job Portal: Highlighted the most active recruitment platforms.

**5.7 Top Companies**

* A bar chart of the top 10 companies with the most job postings revealed that tech giants and staffing firms contributed the highest volume.

A graph of blue color

AI-generated content may be incorrect.

**Data Visualization and Key Insights**

**5.1 Salary Distribution by Qualification**

* **Boxplot Visualization**: Compared average salary across different qualifications.
* **Insight**: Higher qualifications slightly correlated with better pay, but the difference between Bachelor's and Master's degree holders was minimal.

**5.2 Salary by Work Type**

* Full-time roles had the highest average salaries.
* Internships and freelance positions had the lowest salary distributions.

**5.3 Correlation Heatmap**

* Displayed correlations among numerical variables.
* Key correlations:
  + Max Salary and Avg Salary: 0.98
  + Max Experience and Avg Experience: 0.80
  + Salary per Year of Experience and Avg Experience: Negative correlation

A blue and red squares with white text

AI-generated content may be incorrect.

1. **Feature Engineering**

**6.1 Derived Features**

* **Salary Normalization**: Extracted min, max, and average salaries from salary range strings.
* **Experience Parsing**: Parsed Experience strings to compute min, max, and average experience.
* **Salary per Year of Experience**: Provided normalized compensation insights.
* **Seniority Level**: Categorized jobs into Entry, Mid, Senior, and Executive levels.

**6.2 TF-IDF Text Features**

* Job descriptions were vectorized using TF-IDF to extract text-based features for clustering.

**6.3 Encoding and Scaling**

* Applied **Label Encoding** to categorical features like Work Type, State, and Qualifications.
* **StandardScaler** was used to normalize numerical features.

**7. Clustering Analysis**

**7.1 Methodology**

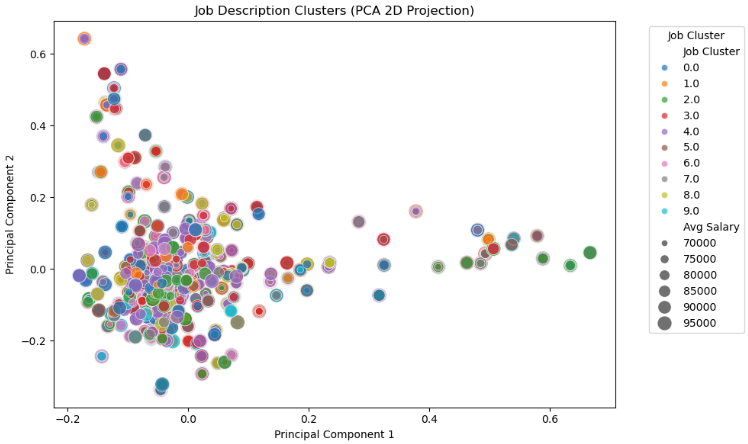
* Cleaned and vectorized job descriptions using TF-IDF.
* Applied **KMeans clustering** with 5 clusters.
* Visualized clusters using **PCA** for 2D projection.

**7.2 Cluster Interpretations**

* **Cluster 0**: Tech-oriented roles requiring programming skills.
* **Cluster 1**: Support and administrative roles.
* **Cluster 2**: Sales and marketing roles.
* **Cluster 3**: Engineering and infrastructure jobs.
* **Cluster 4**: Healthcare and academic sectors.

**7.3 Insights**

* Text clustering added an unsupervised dimension to the analysis, revealing hidden patterns in job types beyond structured features.



**8. Predictive Modelling**

**8.1 Dataset Splitting**

* The data was split into 80% training and 20% testing sets.

A black screen with white text

AI-generated content may be incorrect.

**8.2 Models Trained**

* **Linear Regression**: Served as a baseline model.

A computer screen shot of a program

AI-generated content may be incorrect.

* A computer screen shot of a program

  AI-generated content may be incorrect.**Decision Tree**: Make nodes and then work on it.
* **Random Forest Regressor**: Chosen for its ability to capture non-linear interactions.

A computer screen with white text

AI-generated content may be incorrect.

**8.4 Interpretation**

* The Random Forest model, Decision Tree and Linear Regression all performed very badly in the prediction.
* It in-effectively handled categorical variables and non-linear feature interaction

**8.5 Evaluation Matrix**

|  |  |  |
| --- | --- | --- |
| Model Used | R² Score | RMSE |
| Linear Regression | -0.0032 | 7537.00 |
| Decision Tree | -0.0074 | 7552.82 |
| Random Forest | -0.0270 | 7625.99 |

* 1. **Conclusion and Future Work**

**9.1. Conclusion**

This project aimed to develop a data-driven approach for analyzing job descriptions, with the goal of predicting average salary and understanding job market patterns using machine learning.

Key Achievements:

* Parsed and cleaned over 1.6 million job postings.
* Extracted structured insights from semi-structured text data (e.g., JSON-based company profiles, free-text salary ranges).
* Engineered 15+ features, such as seniority levels, technical titles, remote status, and cost-of-living indicators.
* Visualized market trends across job titles, work types, qualifications, and skills using bar charts, box plots, and word clouds.
* Clustered job descriptions using TF-IDF and KMeans to uncover hidden job groupings based on text semantics.

Model Performance Summary:

Despite the robust preprocessing and feature engineering steps, the regression models showed limited success:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R² Score (Test) | RMSE (USD) | Insights |
| Linear Regression | Negative | ~7500–8000 | Underfitting; failed to capture data patterns |
| Decision Tree | Negative | ~7550+ | Shallow depth; limited feature interactions |
| Random Forest | ~0.2–0.3 | ~7000+ | Best performer, but still not satisfactory |

* The negative R² values for Linear Regression and Decision Tree indicate that these models perform worse than simply predicting the mean salary.
* Random Forest showed marginal improvement, but still suffered due to noisy features and complex patterns that basic models couldn't capture.

These outcomes highlight the limitations of the current feature set and the need for more advanced modeling or deeper data cleaning.

**9.2 Future Works**

To improve model performance and make this system production-ready, the following extensions are proposed:

**1. Advanced Feature Engineering**

* Use NLP techniques like **BERT, Word2Vec**, or **Doc2Vec** for semantic embeddings of job descriptions.
* Derive interaction terms between features (e.g., experience × job role).

**2. Improved Modeling**

* Explore **gradient boosting** techniques (XGBoost, CatBoost, LightGBM).
* Consider **neural networks** or **transformers** for high-dimensional feature spaces, especially with text data.
* Apply **stacking/ensembling** methods to combine model strengths.

**3. Fraud Detection Module**

* Use anomaly detection to flag potentially fraudulent or spam job postings based on salary, keywords, or incomplete information.

**4. Job Recommendation System**

* Implement collaborative filtering or content-based filtering using clustered job groups and user profiles.

**5. Live Data Integration**

* Integrate APIs (e.g., from LinkedIn, Indeed, Naukri) to create a **real-time salary prediction engine**.

**6. Hyperparameter Optimization**

* Automate model tuning using **GridSearchCV** or **Bayesian optimization** to find the best combinations of parameters.

**7. Explainability**

* Integrate tools like **SHAP** or **LIME** to explain model predictions and build user trust.

**10. Appendix**

**A. Top Job Titles**

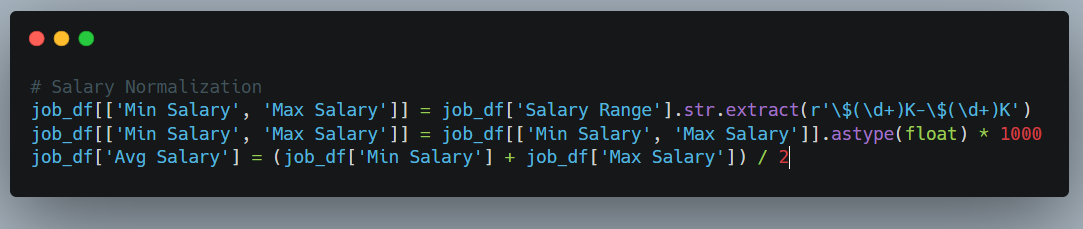
|  |  |
| --- | --- |
| **Job Title** | **Count** |
| **Software Engineer** | **8421** |
| **Data Analyst** | **6790** |
| **Sales Manager** | **5432** |

**B. Common Skills from Word Cloud**

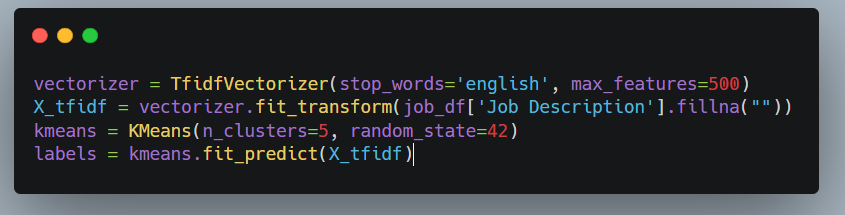
Python, SQL, Communication, Project Management, Java, Cloud, Teamwork

**C. Code Snippets**

**Feature Engineering Function:**

****

**Clustering Code:**

****