|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| IS328 Assignment 2 |
| |  |  |  | | --- | --- | --- | | The Mining Minds | 10/21/24 | IS328 | |

# **Team Information**

**Team Name: The Mining Minds**

|  |  |  |  |
| --- | --- | --- | --- |
| **Member Name** | **Member ID** | **Contribution (%)** | **Section** |
| **Jayshil Singh** | **S11208397** | **100** | **Visualization** |
| **Avichal Nath** | **S11209066** | **100** | **Classification** |
| **Yash Prasad** | **S11208719** | **100** | **Preprocessing** |
| **Ryan Kumar** | **S11209231** | **100** | **Clustering** |

# **Abstract**

This project applies data mining to job postings in Fiji to determine labor market trends. We incorporated classification and clustering to develop an understanding of the most in-demand jobs, required skills, and educational qualifications. The project faced several challenges with regard to data quality such as missing values or unstructured format; it therefore used preprocessing techniques in the process in order to ensure reliable results. With Weka, we developed models interpreting data that can help inform recruitment strategies, shape educational curriculum planning, and support policymakers in their work of improving the labor market. The findings were presented through visualizations to show an overview of Fiji's job market in terms of demand.

**Contents**

[**Team Information** 1](#_Toc180422051)

[**Abstract** 2](#_Toc180422052)

[**List of Figures** 4](#_Toc180422053)

[**List of Tables** 5](#_Toc180422054)

[**Introduction** 6](#_Toc180422055)

[**Motivation:** 7](#_Toc180422056)

[**Problem Domain** 7](#_Toc180422057)

[**Aim and Objective** 7](#_Toc180422058)

[**Data Mining Techniques** 8](#_Toc180422059)

[**Dataset Used** 8](#_Toc180422060)

[**Data Mining Tasks** 8](#_Toc180422061)

[**Data Preprocessing** 8](#_Toc180422062)

[**Models and Methods** 9](#_Toc180422063)

[**Assessments** 11](#_Toc180422064)

[**Results Analysis and Interpretation** 12](#_Toc180422065)

[**Conclusion** 13](#_Toc180422066)

[**Lesson Learnt** 14](#_Toc180422067)

[**Future Work** 15](#_Toc180422068)

[**References** 16](#_Toc180422069)

# **List of Figures**

Tokenization of description:

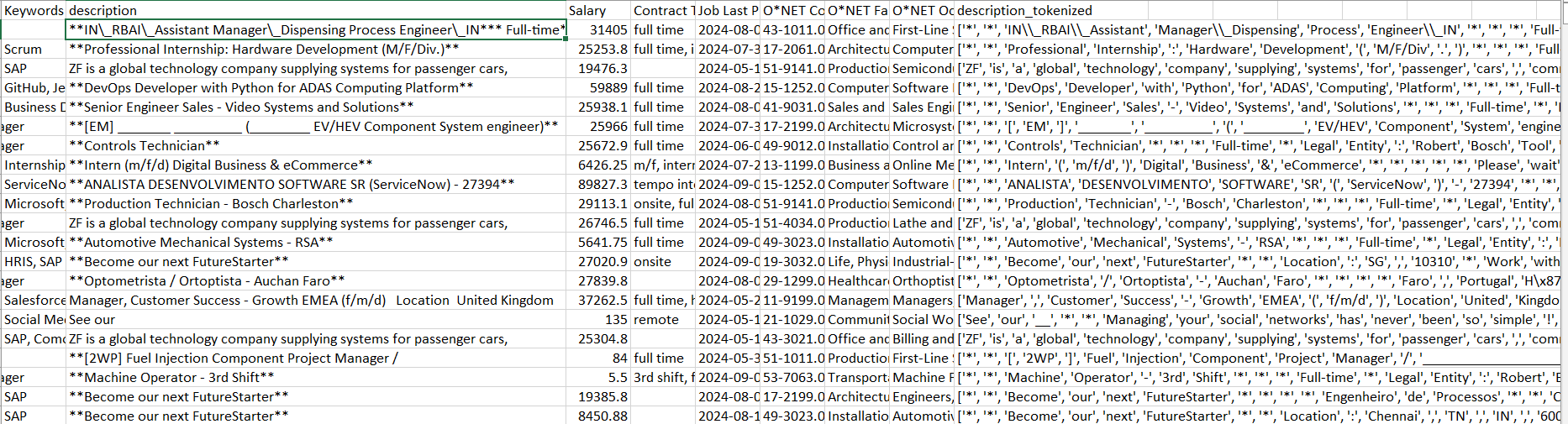


Figure Tokenization

Classification: Clustering Results:

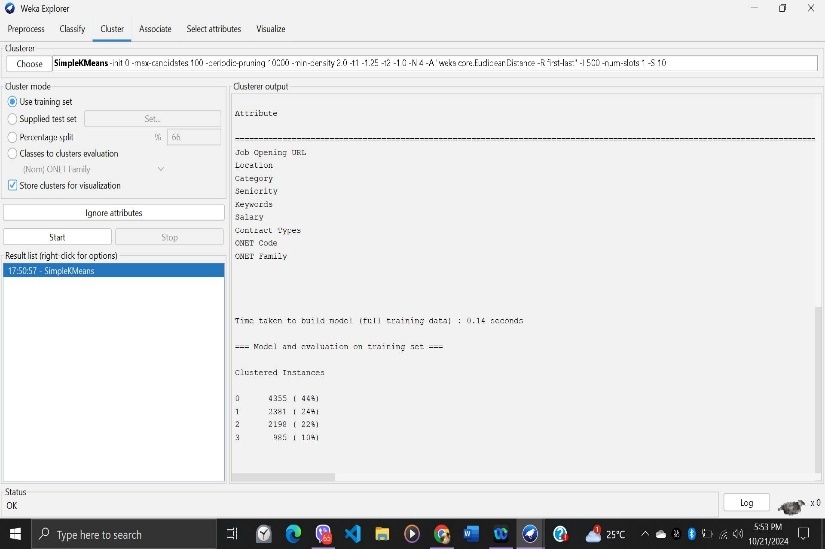
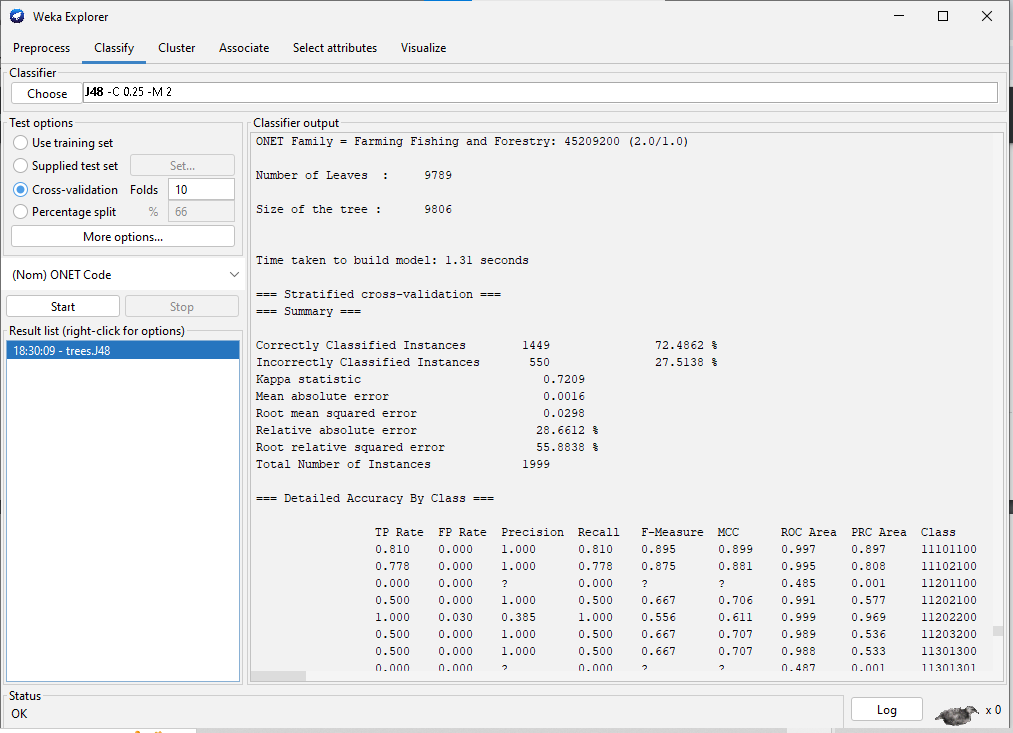
****

Figure Classification

Figure Clustering

Visualization:

# **List of Tables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Action** | **Steps** | **Data** | **Results** |
| Description | Tokenization | Break down sentences into words | Supplier Visits to keep up to date as per new technology | {“Supplier”, “Visit”, “up-to-date”, “new”, “technology”} |
| Seniority | Term Frequency – Inverse Document Frequency | Assigning numerical value to texts | manager | 1 |
| Description | Dimensionality reduction | Remove words such as prepositions and articles | Please be aware of scammers who may fraudulently allege to be from Contentful | Be aware scammers who fraudulently allege being from contentful |

Figure 1 Preprocessing

# **Introduction**

Data Mining is the process of identifying patterns, trends, and relationships using classification, clustering, association rule learning, and regression in huge volumes of data. Data mining encompasses meaningful information extracted from raw data, usually actionable. The major concept of data mining is to transform data into a comprehensible structure for further use. Normally, it is performed with the use of machine learning algorithms. This technology is widely used for market analysis, fraud detection, enhancement of customer relationship management, and predictive modeling.

The proposed job posting dataset will be mined using Weka to accomplish data mining tasks. We will utilize the built-in classification and clustering algorithms in Weka for the analysis of the trends in the job market. Any implementation of new algorithms or modification of the core Weka system is out of the scope. The focus, instead, would go towards an effective implementation of the existing data mining techniques, the preprocessing of the dataset for high quality, and presenting the results in a form that may make more sense to stakeholders and hence actionable.

## **Motivation:**

This research is thus motivated by the need to increase awareness of the labor market situation in Fiji, where availability or lack of certain jobs and their requirements in terms of skills can have huge social and economic ramifications. Identification of the most common job listings and required skills/qualifications will thus be helpful in narrowing the gap between job seekers and employers. It can also serve as an insight for educational institutions with regard to which areas of curriculum development need to be better aligned with market demand to enhance employability. Such an analysis may also be useful to policymakers in their strategy formulation that specifically deals with workforce development, reduction of unemployment, and thereby stimulating economic growth. It would mean that, by understanding these trends, efficient hiring can be done, and job seekers fitted with the appropriate skills in a highly relevant project to today's fast-changing labor landscape.

## **Problem Domain:**

The problem domain of this project relates to an analysis of job postings in Fiji's labor market, regarding trends and patterns. This project focuses specifically on discovering insights into which jobs are most frequently listed, which skills and qualifications are most sought after, and what challenges arise from unstructured job posting data. The project addresses a problem that is related to how one could extract and process large volumes of job posting data efficiently enough to provide meaningful information that could be valuable for job seekers, employers, educational institutions, and policy makers. Issues related to data quality, labor market mismatches, evidence-based decision making in recruitment, and workforce development come into view in this respect.

## **Aim and Objective**

**Aim:** This therefore means that data mining shall be used to analyze job postings within Fiji for most listed jobs and required skills/qualifications for proper analysis that could guide properly the job seeker, employer, educator, and policymaker accordingly.

**Objectives:**

1. To preprocess job posting data by addressing issues such as missing values and unstructured formats, ensuring data quality for analysis.
2. To apply classification techniques to categorize job postings based on factors such as job title, required skills, and qualifications.
3. To use clustering methods to group similar job postings and uncover patterns in job demand across different sectors.
4. To visualize the results, presenting key insights on job trends and demand to make the data accessible and actionable for stakeholders.

# **Data Mining Techniques**

## **Dataset Used**

The dataset used for this project consists of **9,919 rows** of job postings and contains **13 columns** representing various attributes related to job listings. These columns capture essential information about each job posting, such as:

* Job Opening URL
* Location
* Location Data
* Seniority
* Keywords
* Job Last Processed At
* O\*NET Family
* O\*NET Occupation Name

**Data Source:**This dataset is downloaded from Kaggle, representing actual job posting data in Fiji. This needed extensive preprocessing, addressing issues concerning missing values, unstructured text, and inconsistency of data for quality analysis.

## **Data Mining Tasks**

In this project, we will work on the "Job Posting Data in Fiji." The tasks include:

• **Classification:** Predict the type of job (e.g., full-time, part-time, contract) using attributes like job title, required qualifications, and job description.

• **Clustering:** Group job postings into clusters based on similarities in job descriptions, qualifications, and required skills.

• **Exploratory Data Analysis:** Analyze the data to identify trends in job types, key skills, and educational qualifications across job postings.

• **Data Preprocessing:** Handle missing data, unstructured text fields, and inconsistencies in the dataset through necessary transformations.

These tasks aim to provide deep insights into Fiji's job market, enabling informed decision-making by stakeholders. All tasks will be performed in Weka to ensure consistent execution and reliable results.

## **Data Preprocessing**

To account for the aforementioned data quality issues, the following pre-processing techniques will be utilized:

**Treatment of Missing Values:** The Missing Values were handled whereby all the empty fields for the standard data were filled with “None” and the empty fields for the HashMap were replaced with “{NULL}”. For the empty salary field, Global Average Method is used as there were not enough rows with the same occupation and filled salaries to calculate averages for all missing values.

We will perform text processing for the unstructured text fields (job\_description):

* **Tokenization** is the breaking down of a sentence into words to harvest the main points
* **TF-IDF** is used to create numerical features for the text data.
* **Dimensionality** is reduced by removing stop words and stemming in the data.

**Categorical:** location and job title were normalized to account for spelling mistakes and variability between records.

## **Models and Methods**

**Analysis of J48 for Job Posting Classification**

The tree primarily uses the **ONET Code** attribute to classify the data. Let me break down the interpretation:

1. **Root Node**:
   * The root node is **ONET Code <= 15209901**.
   * This means that **ONET Code** is the primary attribute used for the first major decision, effectively splitting job postings based on this value.
   * The root node suggests that **ONET Code** is crucial for distinguishing between job categories.
2. **Subdivisions**:
   * **ONET Code** branches into multiple sub-branches, using specific conditions such as:
     + **ONET Code <= 13209900**
     + **ONET Code <= 25904400**
     + **ONET Code <= 39909900**
   * These subdivisions help refine the classification process based on the different ONET Code ranges.
3. **Leaf Nodes**:
   * The **leaf nodes** represent the final classifications, corresponding to specific job categories. Examples include:
     + **Computer and Mathematical (0)**
     + **Community and Social Legal (20.0)**
     + **Office and Admin Installation (29.0)**
   * These leaf nodes indicate the predicted job category based on the chosen path through the tree.

**Insights from the Tree**

The objective is to understand factors contributing to **job category** or **seniority** and derive insights that can help create **business rules**. Here are the key insights:

1. **ONET Code as a Major Factor**:
   * It is evident that **ONET Code** is the most significant attribute for job classification.
   * The frequent appearance of **ONET Code** in the decision levels implies that job categorization is highly dependent on this attribute, which makes sense since ONET Codes are designed to classify occupational information.
2. **Granular Categorization**:
   * The tree reveals a hierarchical structure where **ONET Code** divides roles into more specific branches using numerical thresholds.
   * For instance, categories like **Business Architecture and Engineering** are split into finer subcategories, indicating that specific ONET Codes are important for detailed job role distinctions.
3. **Highly Specific ONET Code-based Classifications**:
   * The **leaf nodes** provide detailed classifications, such as **Construction and Extraction**, **Food Preparation**, and **Personal Care and Service**.
   * This shows that **ONET Code** is an effective primary indicator for creating highly specific job classifications, which can be used directly for job market analysis.

**The ONET Code serves as a key attribute for systematically classifying occupations, helping to differentiate job roles and provide highly specific job categorizations; in data mining, it acts as an effective predictor for job classification, enabling industry-specific insights and automation in occupational mapping.**

**The ONET Code serves as the primary factor in the decision tree, with the root node starting at ONET Code <= 15209901. For instance, roles within Business Architecture and Engineering are divided into finer subcategories based on more specific ONET Code splits, effectively differentiating between various detailed job roles within the broader category. This demonstrates how ONET Code provides a systematic, hierarchical classification that enables precise categorization within the industry.**

**Clustering Analysis**

* Number of Clusters: 4
* Total Instances: 9919
* Distribution of Instances Across:
  + Cluster 0: 4355 (44%)
  + Cluster 1: 2381 (24%)
  + Cluster 2: 2198 (22%)
  + Cluster 3: 985 (10%)

The dataset is predominantly dominated by Cluster 0 at 44%, while the remaining clusters account equally for the rest. The smallest cluster is Cluster 3, with only about 10% of the data.

**Centroids and Key Attributes**:

1. **Cluster 0 (44% of the data)**
   * Job Opening URL: Most engineering-related posts were from Bosch job listings.
   * Location: Jobs were found to be placed in Bangalore, India. This suggests a region with a high number of engineering jobs.
   * Category: Mainly engineering roles.
   * Seniority: Majority are non-managerial roles.
   * Keywords: SAP appears in most of the job descriptions.
   * Contract Type: Predominantly full-time jobs.
   * ONET Family: Mostly Computer and Mathematical, or Architecture and Engineering jobs.

*Interpretation*: Cluster 0 consists of non-managerial, full-time engineering roles from the Computer and Mathematical or Architecture and Engineering categories, primarily located in India.

1. **Cluster 1 (24% of the data)**
   * Job Opening URL: Listings for human resources, IT, and management jobs.
   * Location: Jobs are also located in Bangalore, India, though the nature of the roles is different.
   * Category: Mainly management roles, implying this cluster could cover mid-to-senior level management.
   * Seniority: Majority are managerial roles.
   * Keywords: SAP features prominently, indicating technical or business-related software knowledge is essential.
   * Contract Type: All roles are full-time.
   * ONET Family: Mostly Computer and Mathematical occupations.

*Interpretation*: Cluster 1 groups managerial, full-time roles in management, HR, and IT with a strong emphasis on SAP and software-related management positions.

1. **Cluster 2 (22% of the data)**
   * Job Opening URL: Product management positions.
   * Location: Includes jobs in Madrid, Spain, and other non-Indian locations.
   * Category: Predominantly engineering roles with a focus on production.
   * Seniority: Non-managerial roles dominate.
   * Keywords: Google is a prominent keyword.
   * Contract Type: Mostly full-time jobs.
   * ONET Family: Primarily from the Production family.

*Interpretation*: Cluster 2 is made up of full-time, non-managerial engineering and product management roles in international locations, with a focus on production and product management.

1. **Cluster 3 (10% of the data)**
   * Job Opening URL: Primarily frontend development and technical roles.
   * Location: Aveiro, Portugal, and other locations, indicating a mix of locations.
   * Category: Mainly internship roles.
   * Seniority: Non-managerial.
   * Keywords: Broad technical skills, including Jenkins, Java, Vue.js, and more.
   * Contract Type: Many hybrid positions (partly onsite and partly remote).
   * ONET Family: Mostly Production or Office and Administrative Support jobs.

*Interpretation*: Cluster 3 represents non-managerial internships and entry-level roles in software and frontend development, with hybrid work arrangements and diverse technical skill requirements.

## **Assessments**

# **Results Analysis and Interpretation**

# **Conclusion**

# **Lesson Learnt**

# **Future Work**

# **References**