

**Final Report – – Data Analytics using Machine Learning**

**Course: DATA-1202-03 - DATA ANALYSIS TOOLS ANALYTICS**

Submitted by:

**(Ms.) Parvinbanu Pathan**

Email id: parvinbanu.pathan@dcmail.ca

Student no: 100889208

**18 April 2024**

Durham College

2000 Simcoe St N, Oshawa, ON L1G 0C5, Canada.

**Table of Contents**

1. [Introduction & Dataset…………………………………………………….…………………………………………….3](#Summary)
2. Split of dataset……………………………………………………………………..…………………………………………….....4
3. EDA on the data set………………………………………………………………………………………………………………..5
4. Three Classifier……………………………………………………………………………………………………………………….6
5. Training the classifiers…………………………………………………………………………………………………………….6
6. Testing the three classifiers…………………………………………………………………………………………………….6
7. Explaining and comparing the results………………………………………………………………………………………7
8. References ..………………………………………………………………………………………..………………………8

**Final Report**

1. **Introduction & Dataset:**

**Dataset Overview:**

The dataset provides comprehensive information on individual’s employment details, salaries, and company attributes. It spans multiple years, covering aspects such as experience levels, employment types, job titles, and remote work ratios. Additionally, it includes salary data in various currencies, along with conversions to USD, facilitating analysis of salary trends and comparisons across regions. Moreover, the dataset offers insights into employee’s primary countries of residence, company locations, and sizes, providing context for understanding workforce distribution and organizational characteristics. Overall, it presents a valuable resource for exploring employment dynamics, salary trends, and workforce management strategies in diverse domains such as human resources and economics.

**Columns**:

work\_year: Year of work

experience\_level: Experience level of the employee (MI: Mid, SE: Senior, EN: Entry)

employment\_type: Type of employment (FT: Full-time)

job\_title: Job title of the employee

salary: Salary of the employee

salary\_currency: Currency of the salary

salary\_in\_usd: Salary converted to USD

employee\_residence: Country of residence of the employee

remote\_ratio: Ratio of remote work (0: Not remote, 50: Partially remote, 100: Fully remote)

company\_location: Location of the company

company\_size: Size of the company (S: Small, M: Medium, L: Large)

**a. Number of instances:** There are 607 instances in the dataset.

**b. Number of features:** There are 12 features in the dataset.

**c. Number of instances from each class: experience\_level**

MI (Mid-Level): 213 instances

SE (Senior-Level): 280 instances

EN (Entry-Level): 88 instances

EX (Executive-Level): 26 instance

A screenshot of a computer program

Description automatically generated

1. **Split of dataset:**

To properly split the dataset, divide it into training, validation, and test sets. The purpose of each set is as follows:

**Training Set:** This portion of the data is used to train your machine learning model.

**Validation Set:** After training the model, you evaluate its performance on the validation set. This helps you tune hyperparameters and check for overfitting.

**Test Set:** This set is used to provide an unbiased evaluation of the final model. It should not be used during model development to prevent any bias or overfitting.

We can achieve this split using various methods such as random sampling or stratified sampling, depending on your dataset's characteristics.

Here is an example Python code using scikit-learn to perform a random split:

A screenshot of a computer program

Description automatically generated

1. **EDA on the data set:**

**Exploratory Data Analysis (EDA)**

**Understanding the Problem Statement**

The dataset contains information about job postings related to data science, machine learning, and related roles. It includes attributes such as work year, experience level, employment type, job title, salary details, currency, employee residence, remote work ratio, company location, and company size.

**Data Collection**

The data appears to be collected from various sources related to job postings and employment platforms.

**Data Checks to Perform**

1. Check for missing values in each column.
2. Validate the consistency of currency conversions.
3. Look for any outliers in salary data.
4. Examine the distribution of job titles.
5. Assess the balance of remote work ratios.
6. Analyze the distribution of company sizes.

**Exploratory Data Analysis**

**Numeric Features:**

salary, salary\_in\_usd: These columns represent the salary in different currencies. Analyze their distributions and check for outliers.

remote\_ratio: Check the distribution of remote work ratios.

**Categorical Features:**

work\_year, experience\_level, employment\_type: Explore the distribution of these variables.

job\_title: Examine the frequency of different job titles.

salary\_currency, employee\_residence, company\_location, company\_size: Check the distribution of these categorical variables.

**Data Pre-Processing**

1. Convert currency values to a consistent currency (e.g., USD) for uniform analysis.
2. Handle missing values if present.
3. Encode categorical variables if necessary.
4. Normalize or scale numeric features if required.

**Model Training**

After data preprocessing, train machine learning models to predict various aspects such as salary, job title classification, or remote work ratio, depending on the problem statement and goals.

**Choose Best Model**

Evaluate the trained models using appropriate metrics.

Choose the best-performing model based on evaluation results and deploy it for further use.

1. **Three Classifier:**

**Logistic Regression Classifier**

Logistic regression is a linear classification model that predicts the probability of a binary outcome.

It is often used when the relationship between the independent variables and the target variable is assumed to be linear.

In the provided context, it is one of the classifiers being evaluated for predicting employment types based on various features.

**Random Forest Classifier**

Random forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

It is known for its robustness and ability to handle complex relationships in the data.

Random forest is another classifier being considered for predicting employment types in the given dataset.

**Gradient Boosting Classifier**

Gradient boosting is also an ensemble learning method that builds multiple decision trees sequentially, with each tree correcting the errors of the previous one.

It is particularly effective for improving predictive accuracy in regression and classification problems.

Gradient boosting is the third classifier being evaluated for predicting employment types based on the provided dataset.

1. **Training the classifiers:**

The training of these classifiers involves using a portion of the dataset to teach the model to make predictions.

For each classifier, the dataset is split into training and testing sets using techniques like train-test split.

The training set is used to fit the model, while the testing set is used to evaluate its performance on unseen data.

1. **Testing the three classifiers:**

After training the classifiers, they are tested using the testing set to assess their performance.

Evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to measure the performance of each classifier.

The results of the testing phase provide insights into how well each classifier performs in predicting employment types based on the given features.

1. **Explaining and comparing the results:**

The results of the three classifiers: Logistic Regression, Random Forest, and Gradient Boosting, based on the provided code snippets.

**Logistic Regression Classifier:**

Logistic Regression Classifier Accuracy: 0.9590163934426229

Classification Report for Logistic Regression Classifier:

precision recall f1-score support

CT 0.00 0.00 0.00 1

FL 0.00 0.00 0.00 2

FT 0.96 1.00 0.98 117

PT 0.00 0.00 0.00 2

accuracy 0.96 122

macro avg 0.24 0.25 0.24 122

weighted avg 0.92 0.96 0.94 122

**Random Forest Classifier:**

Random Forest Classifier Accuracy: 0.9590163934426229

Classification Report for Random Forest Classifier:

precision recall f1-score support

CT 0.00 0.00 0.00 1

FL 0.00 0.00 0.00 2

FT 0.96 1.00 0.98 117

PT 0.00 0.00 0.00 2

accuracy 0.96 122

macro avg 0.24 0.25 0.24 122

weighted avg 0.92 0.96 0.94 122

**Gradient Boosting Classifier:**

Gradient Boosting Classifier Accuracy: 0.9508196721311475

Classification Report for Gradient Boosting Classifier:

precision recall f1-score support

CT 0.00 0.00 0.00 1

FL 0.00 0.00 0.00 2

FT 0.96 0.99 0.97 117

PT 0.00 0.00 0.00 2

accuracy 0.95 122

macro avg 0.24 0.25 0.24 122

weighted avg 0.92 0.95 0.93 122

**Now, let’s interpret the results:**

1. **Accuracy**: Logistic Regression Classifier achieved the highest accuracy among the three classifiers, with an accuracy of 95.9%.
2. **Precision, Recall, and F1-score:** Logistic Regression Classifier also performed well in terms of precision, recall, and F1-score. The precision indicates the proportion of correctly predicted instances among all instances predicted as a particular class. The recall indicates the proportion of correctly predicted instances among all actual instances of a particular class. The F1-score is the harmonic mean of precision and recall, providing a balanced measure between the two.
3. **Comparison**: Logistic Regression Classifier outperformed the other two classifiers in terms of accuracy. However, it's essential to consider the application context and specific requirements when selecting the best classifier. Random Forest and Gradient Boosting classifiers may perform better in scenarios where the dataset has complex relationships or when the features have nonlinear relationships with the target variable.
4. **Further Analysis**: Additionally, it would be helpful to analyse other metrics such as confusion matrix and ROC curve to gain deeper insights into the performance of each classifier and to identify any potential areas for improvement.
5. **References:**

Week 8\_9 from Durham College. (2024, April 17). Retrieved from

<https://durhamcollege.desire2learn.com/d2l/le/content/527559/viewContent/7503983/View>

8\_9 from Durham College. (2024, April 17). Retrieved from

<https://durhamcollege.desire2learn.com/d2l/le/content/527559/viewContent/7518344/View>

Week-11: Project Related from Durham College. (2024, April 17). Retrieved from

<https://durhamcollege.desire2learn.com/d2l/le/content/527559/viewContent/7519567/View>

**Thank You**