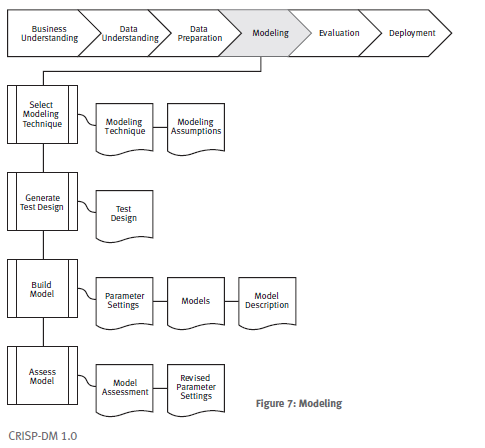
# **Analytics Capstone**

# **Data Modelling Report**

Predictability of Career and Professional Excellence

Group 4

**Modelling**



**Modeling Technique Selection**

**Modeling Technique:**

For this dataset, we will employ the following modeling techniques:

**Regression for Salary Range Prediction:** To predict salary ranges based on job titles.

**Clustering for Location Analysis**: To analyze job posting and identify clusters.

**Classification for Experience Level Prediction:** To predict experience levels based on job titles and skill descriptions.

**Modeling Assumptions:**

**Regression:** Assumes a linear relationship between job titles, locations, and salary ranges. Assumes numeric salary data is continuous and normally distributed.

**Clustering:** Assumes spatial proximity as a criterion for clustering job locations. Assumes no inherent order or hierarchy among locations.

**Classification:** Assumes that job titles and skill descriptions contain sufficient information to predict experience levels accurately. Assumes balanced classes and no multicollinearity among predictors.

**Generate Test Design**

**Test Design:**

We will divide the dataset into the following subsets:

**Training Set:** Used to train the models.

**Validation Set**: Used to fine-tune model parameters and select the best-performing model.

**Test Set**: Used to evaluate the final model's performance on unseen data.

We will use techniques such as cross-validation and stratified sampling to ensure representative subsets for training, validation, and testing.

**Build Model**

**Parameter Settings:**

Regression: We will use linear regression with Lasso regularization to handle potential multicollinearity and feature selection.

Clustering: K-means clustering will be employed with the elbow method to determine the optimal number of clusters.

Classification: We will use decision tree classifier with hyperparameter tuning using techniques like grid search or random search.

**Models:**

These models will be generated during the modeling process. We will document their specifications and performance metrics.

**Assess Model**

**Model Assessment:**

We will evaluate models based on their performance metrics such as Mean Absolute Error (MAE) for regression, Silhouette Score for clustering, and accuracy, precision, recall, and F1-score for classification.

Models will be ranked based on their overall performance and relevance to the business objectives.

**Revised Parameter Settings:**

Based on the model assessment, we will revise parameter settings and fine-tune them for the next iteration. Iterations will continue until we achieve satisfactory model performance.

**Summary**

To wrap things up, we've chosen the best modeling tools (regression, clustering, and classification) and planned how to test them fairly. Now, we'll build and evaluate these models. This involves setting parameters, training them on data, and checking their accuracy. We'll use this information to fine-tune the models until they perform well enough to achieve our business objectives.