**MSc in Data Analytics – “Comparative Analysis of Construction Sectors in Ireland, Portugal and UK: A Data Mining Approach.”**

Omprakash Vootla

7 Parklands Row, Parklands

Saggart, Dublin 24 Ireland.

Email: [sbs23020@student.cct.ie](mailto:sbs23020@student.cct.ie)

Student Number: Sbs23020

***Abstract***

*Objective of this Data Mining project is to compare the Construction sector in Ireland and other countries. This study conducted on multiple datasets based on the topics, first Ireland dataset contains the information of average earnings per week in construction industry of various employment types, Second Ireland dataset contains the information of incidents in construction industry.*

*Third dataset contains the information of Number of employees per quarter in Portugal construction industry.*

*Fourth dataset contains the information of UK incidents in construction industry.*

*Approach is to apply EDA (Exploratory Data Analysis), data processing, data preparation and data visualisation to understand the earnings, employment, and incidents trends in construction industry over the time.*

*In EDA we renamed columns, removed unwanted columns, and fixed missing values, we explored Supervised and Unsupervised Machine learning Models.*

***Introduction***

Fire poses a serious risk to the people, property, and environment. This study on fire locations and fire stations turnouts from county councils of Ireland.

This Study is focused on understanding the factors contributing to fire incidents and mitigate their impact for the safety of the people.

Our aim is to analyse data using EDA, Data processing, Data preparation, and feature selection and finally apply Machine learning models to predict target variable.

We merged two data sets and combined same location columns assuming this will give better outcome.

We explored data mining techniques to analyse the relationship between the several types of fire incident locations and fire stations callouts Turnouts to fire, Turnouts to Special Service Incidents or Turnouts to False Alarms.

Both datasets have a limited number of data points, and twenty-six columns has the labelled data, and only County Council column has categorical data. Regression models will be best fit for labelled data. We explore regression models like Linear Regression, Ridge Regression, Decision Tree, and Random Forest and Classification models like Logistic Regression, Decision Tree, Random Forest Classifier and Support vector machine models to compare the results and find best model.

To apply Machine learning models, we split the data set as train and test sets with 70%-30%. We evaluate the results by various methods like, MSE, R-Square, F1-Score, Accuracy and Confusion Matrix etc.

Finally, we find the best results from Ridge Regression Random Forest model in predicting the factors contributing to occurrence of fire incidents.

With these results we can help the fire safety authorities to make initiative-taking measures in planning and resource allocation at high-risk counties.

# **Scope and Methodology:**

In this section we are exploring the Methodology for this data mining project. We are implementing **CRISP-DM** framework for this project. CRISP-DM is well-established and widely used for data mining and Machine learning projects. This framework consists of six phases.

Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment.

Diagram

Description automatically generated