**Statistical Analysis and Machine Learning for Traffic Prediction in Conyngham Road Dublin**

Author: Jefferson William Teixeira

e-mail: [sba22201@student.cct.ie](mailto:sba22201@student.cct.ie)

Student ID: sba22201

**Abstract**

*Every year Dublin City Council conducts traffic counts at various locations on entry points around Dublin city centre. The locations are chosen to ensure that any person entering the City Centre from outside must pass through one of the locations where the surveys are undertaken. These traffic counts provide a reliable measurement of the modal distribution of persons travelling into, and out of, Dublin city on a year-on-year comparable basis. The data collected is divided into the transport modes allowing a better understanding of the changing usage trends in pedestrian and various vehicle types. In this work, statistics models and machine learning techniques have been applied to extract descriptive measurements and appropriate probability distributions from the data and to make predictions for the future using the historical data. Understanding and predicting the traffic behaviour is beneficial to the city planning as it indicates how the public resources can be applied more efficiently in terms of road infrastructure and public transportation, for example. More specifically, the inbound traffic data of different transportation modes that were collected in Conyngham Road, Dublin have been studied. The machine learning framework CRISP-DM was adopted to guide the analytical process from the data understanding step to the evaluation of the modelling results. Several experiments were performed using different machine learning models and the technique with the best results was KNN Regression algorithm, having a R2 score of 81% using four neighbours as detected by GridSearchCV technique. On the other hand, the models with the lowest performance were Polynomial Regression, with a score of 38% and Ridge Regression, with 35% score, both having their hyperparameters improved by GridSearchCV.*

Contents

[2 Introduction 4](#_Toc121587141)

[3 Materials and Methods 4](#_Toc121587142)

[3.1 Project Management Framework 4](#_Toc121587143)

[3.2 Data Understanding 5](#_Toc121587144)

[3.3 Solution Implementation 5](#_Toc121587145)

[3.4 Pre-processing 5](#_Toc121587146)

[3.5 Data Preparation 5](#_Toc121587147)

[3.6 Statistics Logic 5](#_Toc121587148)

[3.6.1 Technique 1 5](#_Toc121587149)

[3.6.2 Technique 2 5](#_Toc121587150)

[3.7 Machine Learning 5](#_Toc121587151)

[3.7.1 Experiment Setup 6](#_Toc121587152)

[3.7.2 Model 1 6](#_Toc121587153)

[3.7.3 Model 2 7](#_Toc121587154)

[4 Results 7](#_Toc121587155)

[5 Conclusions 7](#_Toc121587156)

[6 References 7](#_Toc121587157)

# Introduction

This report is organized as follows: Section 3 describes the dataset and the end-to-end workflow that were executed, from the data preparation to the analysis of the results. Moreover, it describes the approaches adopted to answer the proposed questions for the statistical and traffic prediction analysis. Section 4 summarize the results of the statistical and ML models, while Section 5 presents the conclusions of the project.

# Materials and Methods

This projected is divided into statistical and machine learning analysis. In this section the main questions that were addressed in each part of the project are defined as well as the approach that was used to solve them in the traffic dataset.

This section is structured as follows: The project framework adopted in this project is outlined in Section 3.1. The dataset and its characteristics are described in Section 3.2. The implementation decisions of the solution are summarized in Section 3.3. The steps in the pre-processing stage are explained in Section 3.4, while the data preparation is detailed in section 3.5. Finally, Sections 3.6 and 3.7 show the techniques that were applied in the statistical and ML analysis, respectively.

## Project Management Framework

For this project, the Cross Industry Standard Process (CRISP-DM) (Chapman *et al.*, 2000) was used as a reference framework. This way the following steps were performed, as illustrated in the Figure *1*.

* Data understanding: In this step, the dataset was studied to understand its formatting and constraints as well as to decide which data could be used in the analysis. As a result, the scope of the project was nailed down to cover the analysis of the inbound traffic on Conyngham Road, Dublin.
* Data preparation: In this step, the raw dataset was manually collected from the source and passed as input to the code that was implemented to process and prepare the data for the Machine Learning and Statistical analysis.
* Modelling: This step was divided into the statistical and ML modelling steps. On the statistical part, the appropriate probability distributions and descriptive statistics were used to answer the questions Q1, Q2 and Q3. Regarding the ML study, the supervised regression learning models were applied, which could be used to answer question Q4.
* Evaluation: In this step, the results and performance of the ML models were evaluated and discussed based on the K-Fold cross-validation technique. The modelling was also improved using a hyperparameter tunning technique called GridSearchCV.

Diagram

Description automatically generated

Figure 1- CRISP-DM Framework.

In the following sections, it is described in detail how each step of the framework was applied in this project.

## Data Understanding

## Solution Implementation

## Pre-processing

## Data Preparation

## Statistics Logic

### Technique 1

### Technique 2

## Machine Learning

Machine learning is about extracting knowledge from the data (Guido and Müller, 2016). The main goal is to enable computers to identify patters in the data in other to extract relevant information or insights that can be used in many applications.

The main algorithms in machine learning fall into two categories: Supervised and Unsupervised (Guido and Müller, 2016). In the first one, the computer is trained with labelled data, while in the second one, the data has no label and the technique applied must identify the relationships in the data. There are two types of supervised problems: Classification and Regression (Guido and Müller, 2016). While the first one is concerned with predicting labels, the second is commonly used to predict continuous numbers. The right technique depends on the problem being solved and the characteristics of the dataset.

### Experiment Setup

The following algorithms were tested against the traffic dataset: Polynomial, KNN and Ridge regression. The steps and the code executed for this part of the project can be seen in the *Machine Learning Models* section of the Jupyter notebook.

Each model was executed, and the results were compared based on a score function given by the coefficient of determination (R2). This score provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model (Scikit-learn, 2022).

Polynomial and KNN regressions were applied to all vehicles so the performance of the models can be seen across the board. The dataset with best score results were then tested against Ridge, so the two linear regression types of algorithms could be compared.

The scores were obtained using two approaches: Manual and cross-validation. In both approaches the dataset was split into train (70%) and test (30%) data, the difference is that in the first approach, the scores were obtained manually for each execution of the model with the specific parameters set, while in the second approach a K-Fold cross-validation technique was used. K-Fold basically splits the training set into smaller sets, where the final score is the average of the scores obtained with different train sub-sets given by the K-Fold splitting (Scikit-learn, 2022).

Section 3.7.2 discusses the linear regression models based on the results of manual and cross-validation scores. Section 3.7.3 presents the results of KNN model using the manual validation approach. Finally, Section 4 presents the overall results obtained using a technique called GridSearchCV.

### Model 1

Linear models are the simplest and most classic models for regression problems (Guido and Müller, 2016). They are widely adopted to make value prediction using a linear function of the input variables or features of the dataset.

Classic linear regression models find the function parameters that minimize the mean squared error between predictions and the true regression targets defined on the training set. The mean squared error is the sum of the squared differences between the predictions and the true values, divided by the number of samples. Polynomial regression is a generalization of this technique that tries to fit a N degree polynomial function to the dataset.

Ridge regression uses the same formula as Linear Regression to make predictions. However, this model tries to reduce the influence of the features in the outcome to avoid overfitting, a side-effect that occurs when the model becomes too complex with several parameters. The Ridge model makes a trade-off between the simplicity of the model (near-zero coefficients) and its performance on the training set (Guido and Müller, 2016). This mechanism is controlled by an alpha parameter.

### Model 2

KNN regression is an adapted version of the classic KNN algorithm commonly used for classification problems (Guido and Müller, 2016). In the regression version, it uses the true target values points to predict the value of the new entries.

The parameter of this model is the number of neighbours (K) it uses to make the predictions, which is calculated based on the average of the neighbours’ target values.

# 

# Results

GridSerachCV is a technique used to estimate the best parameters for a ML model. It implements an exhaustive search over the hyperparameters provided and calculates the best options based on the executions.

The selection of parameters that were obtained with GridSearchCV on the models that were applied to the car traffic dataset are summarized in Table 1. It shows the name of the model, the tested parameters, the optimal choice and the R2 score obtained when the best parameters were used. The code executed in this part can be seen in the *Hyperparameter tunning with GridSearchCV.*

# Conclusions

# References