# CCT College Dublin

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *CA1* |
| **Lecturer Name:** | *Marina Lantorno*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | *Marcelo Alves Queiroz* |
| **Student Number:** | *2022010* |
| **Assessment Due Date:** | *11th April 2022* |
| **Date of Submission:** | *11th April 2022* |

## Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

[CCT College Dublin 1](#_Toc100590793)

[1](#_Toc100590794)

[Declaration 1](#_Toc100590795)

[Introduction 3](#_Toc100590796)

[1. Overview of the Problem 3](#_Toc100590797)

[2. Data and Model Approach 3](#_Toc100590798)

[2.1 Street Class 3](#_Toc100590799)

[2.2 New Area 4](#_Toc100590800)

[2.3 Surface types 4](#_Toc100590801)

[3. EDA Process 4](#_Toc100590802)

[3.1 Inner Merge 4](#_Toc100590803)

[3.2 Dealing with missing data 4](#_Toc100590804)

[3.3 Duplicated Method 5](#_Toc100590805)

[3.4 Categorical Features 5](#_Toc100590806)

[4. Correlation Between Features 5](#_Toc100590807)

[5. Roads and Types Description 6](#_Toc100590808)

[5.1 Road Classes 6](#_Toc100590809)

[5.2 Area Code 6](#_Toc100590810)

[7. Machine Learning Algorithms 8](#_Toc100590811)

[7.1 Naïve Bayes Classification & Decision Tree Classifier 8](#_Toc100590812)

[7.2 Naïve Bayes Classification 8](#_Toc100590813)

[8 Confusion Matrix 9](#_Toc100590814)

[9 Cross Validation 9](#_Toc100590815)

[10 Decision Tree Model 9](#_Toc100590816)

[11 Support Vector Machine 10](#_Toc100590817)

[11.1 What is the result of SVM? 10](#_Toc100590818)

[12 Conclusion 10](#_Toc100590819)

A case study of Machine-Leaning Models to Make Predictions and find issues in Roads and Streets in Dublin City

# Introduction

This case study has as objective; collect, process, analyse, apply Exploratory Data Analysis (EDA) on the following dataset: Roads and Streets in DCC (Dublin City CONSOL), from Dublin City in order to make predictions, identify possible issues using scientific statistics methods and apply different machine learning algorithms to make predictions and comparations of which machine learning algorithm has better accuracy on the dataset.

Roads and streets available at:

<https://data.gov.ie/dataset/dublin-city-roads-and-streets> (data from 2016. Accessed: 20/03/2022).

“The methodology behind these words is CRISP-DM. From Daimler Chrysler (formerly Daimler-Benz), SPSS (then ISL), and NCR they introduced the Cross-Industry Standard Process for Data Mining (CRISP-DM) in late 1996. This model illustrates the evolution of refinements throughout time. It is divided into six stages or phases.”

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

# Overview of the Problem

When data is effectively utilized, it may be transformed into insights, which can then be transformed into an automated process. After then, the automated process will question how others perceive problems and generate solutions based on data driven decisions.

This study case will enquire and represent which roads has not been repaired over the last years, when was last time repaired and why, what type of street has been more and less repaired, predicting when it is going to be repaired, predicting which type is more durable, which kind of roads and streets Dublin have on west side and, which has on north side so that predictions can be made in regarding to the future.

# Data and Model Approach

Our study case is going to be based on CRISP-DM and data-driven decision making, we are going to let the data lead de decisions with the insights getting after each step.

Roads and streets dataset has in total 3 categorical features, which are classified as categorical, defined by the features below:

2.1 Street ClassNational Primary (It is the majority roads among urban centres), National Secondary (this category form an important part of the national route network, the second most important in the main arterial routes which are the Primary Roads), Local Road (Not a primary or secondary road nonetheless, Public road form a link in the national network of roads) and Regional Road (There is no strategic on this categoric, it’s literally local roads with no strategy behind the scenes).

* 1. New Area Is the five administrative areas within Dublin City Council, 1-Central, 2-North West, 3-North West, 4-South Central and 5-South East.

2.3 Surface typesof each road and streets are defined by 1 – Asphalt, 2 – Concrete, 3 – Macadam, 4 – Setts, 5 – Flags.

Because the problems are so straightforward, there’s not much of a need for an outline of the kind described above. Notice the clear exposition, the labelled figures and tables that are referred to in the text, and the careful integration of visual and numerical evidence.

# EDA Process

Dublin city council has two different datasets in relate to roads and streets, “in charge” and “Roads and streets”, both datasets with a correlated feature “roads and streets”. After get information about these dataset, Roads and streets is our main dataset, as it has more features and observations.

# Inner Merge

In order to get more features and complete observations inner merge was applied as fits better to what the dataset is asking which is get features from In charge dataset and create one unique dataset.

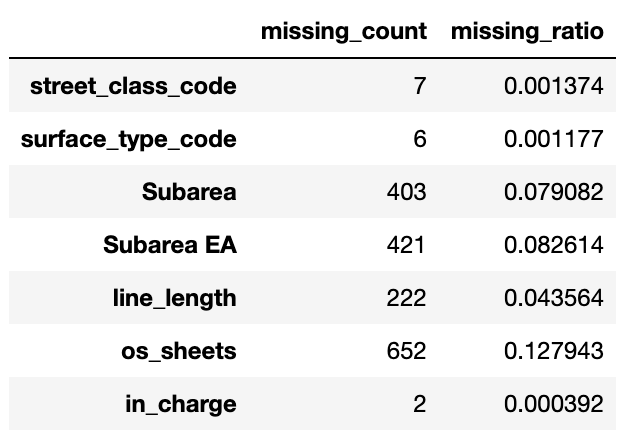
“By default, the merge function performs an inner join. It takes both the data frames as arguments and the name of the column on which the join has to be performed”. (Géron, 2019)

# Dealing with missing data

Now dealing with a dataset of 4767 observations and 8 columns, some features such as route number, impact, Irish street names, road start, road finish and year built has been discarded using drop method as were missing so many information.

A function called missing\_data has been created in order to check for missing data every time a procedure for missing data was applied.

Executed the function for the first time, we gained how many values were missing in each features and observations as below table:



**Table 1:** Representing the counts of missing data.

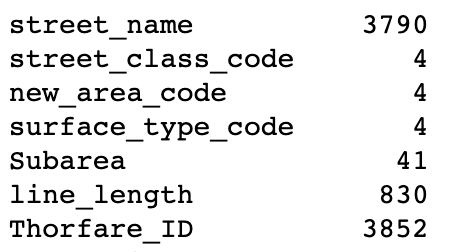
Upon insight gain above (Table 1) we can see that just a small amount of data is missing in these features, so these observations can be discarded.

# Duplicated Method

Dealing with duplicates using sum function from pandas library. Applied in a very handy way .drop\_duplicates () at the moment of creating the dataset, this method helped to avoid any future problems regarding duplicate data.

# Categorical Features

Appling unique function, the dataset is telling us that the features class code, new area code, surface type, in charge and subarea, are categorical features, as we can see below table from Jupiter Notebook representing categorical features.

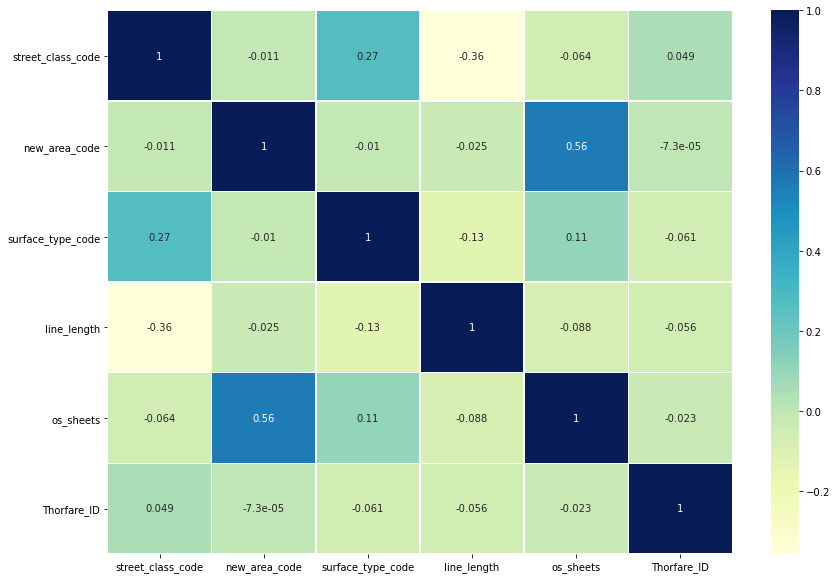


**Table 2:** Representing categorical data.

# Correlation Between Features

To better understand the connections between variables, it might be valuable in data analysis and modelling. The correlation of two variables is the statistical link between them. A positive correlation means that both variables move in the same direction, whereas a negative correlation means that when one variable's value rises, the other variable's value falls. Correlation can also be 0 or neutral, indicating that the variables are unconnected.

Below heatmap shows the correlation between the features in the dataset:

****

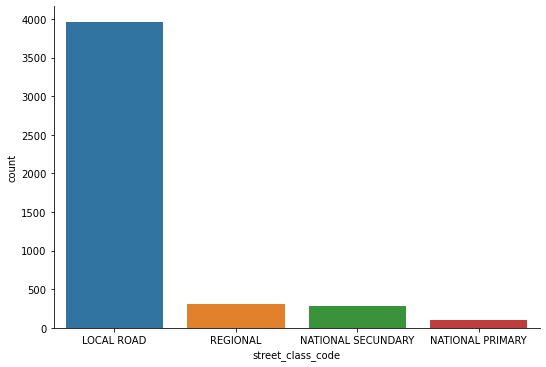
**Graphic 1:** Heatmap representing the correlation between the features.

# Roads and Types Description

# Road Classes

The road classes in Dublin are represented by Local Road, Regional, National Primary and Secondary. On the graphic below we can see clearly that in Dublin city we have more local roads, meaning that 4000 roads, we should have lots of apartments and houses in local roads.

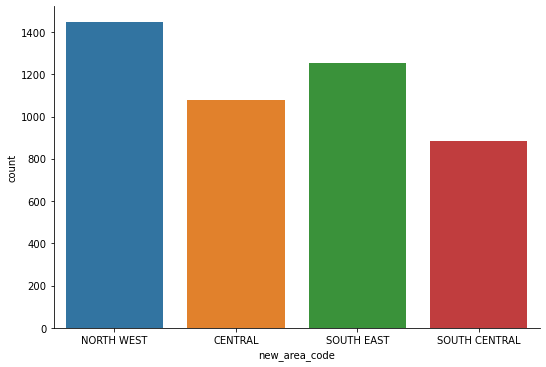
***Graphic 6:*** *Road Classes in Dublin*

****

# 5.2 Area Code

Area code are defined by; North West, Central, South East and South Central.

Below plot we can conclude that the roads and streets are sparse through Dublin condignly, there are no relevant differences between south side and north side.

****

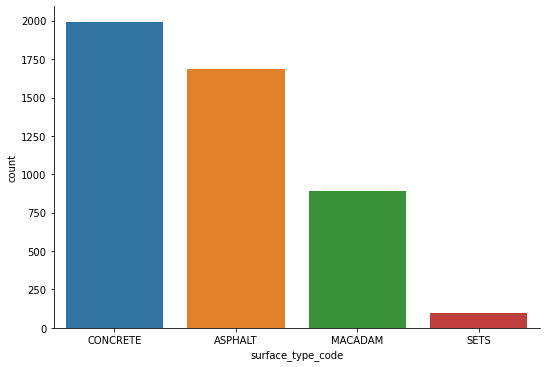
***Graphic 7:*** *Code Area in Dublin*

* 1. **Surface Type**

What would you expect to have more as surface type?

According with road guidelines in Ireland the most durable surface type is the concrete, during 25 years, just behind concrete, there is asphalt its time life is approximately 18 years, then macadam and sets.

(‘Road Guidelines’, no date; ‘https://data.gov.ie/dataset/dublin-city-roads-and-streets’, no date)

****

***Graphic 7:*** *Surface road types*

*in Dublin Roads*

1. **Normalization or Standardization**

Standardization comes into picture when features of input data set have large differences between their ranges, or simply when they are measured in different measurement units.

We try to bring all the variables or features to a similar scale. Standardization means centralize the variable at zero.

Normalization and standardization of features is done to bring all features to a similar scale.

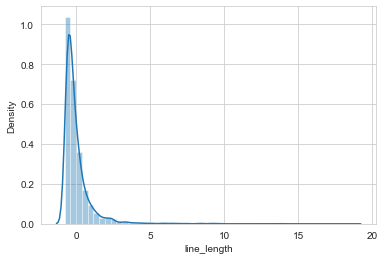
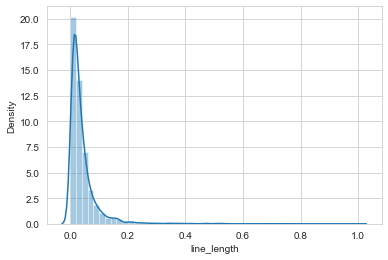
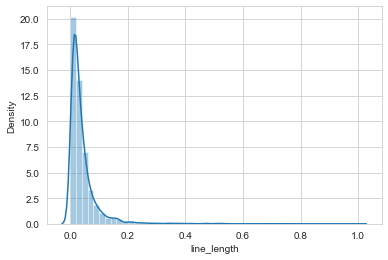
As results of EDA process and correlation, we can see that the correlation is super low and only 4 features have correlations.

Normalization and standardization must be applied in continues data which only is in line\_length, below I will perform normalization on line\_length feature, as it is the only continues feature we have in the dataset:

***Graphic 3:*** *Z-Score*

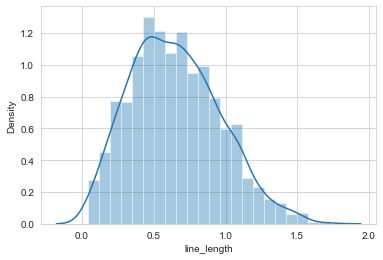
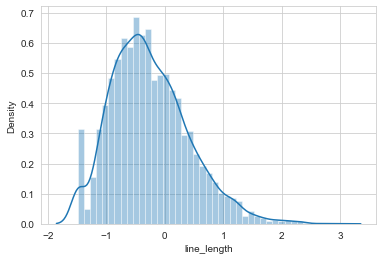
***Graphic 3:*** *Min & Max*

***Graphic 2:*** *Feature Scaling*



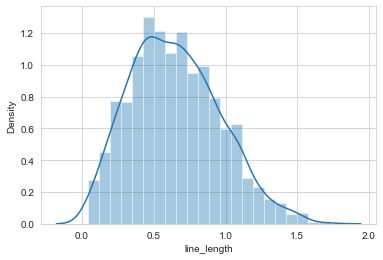
***Graphic 5:*** *Square Root*

***Graphic 4:*** *Log Transformation*



As per graphic above, Square Root worked better when it comes to normalize our data.

(Graphic 5: Square Root)



***Graphic 5:*** *Square Root*

# 7. Machine Learning Algorithms

The dataset has two different types of variables, labelled and unlabelled, which means we could use machine learning algorithms for semi supervised learning as we have two different types. We could also use algorithms for independent variables as our variable street class code is independent but there is no correlation between our labelled and unlabelled features, after long study of the data, the insight gain is that, apart from our independent variable, we can perform supervised machine learnings algorithms to predict what is the type of road using labelled variables new area code and surface type code.

# Naïve Bayes Classification & Decision Tree Classifier

In this study case, we propose two different learning algorithms for categorical data, Naïve Bayes Classifier and Decision Tree.

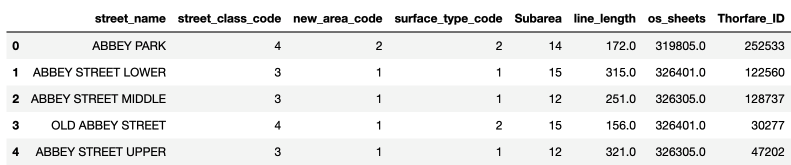
“Decision Trees are the fundamental components of Random Forests, which are among the most powerful Machine Learning algorithms available today.”

(Géron, 2019)

The overall structure of a Nave Bayes Tree is a decision tree, with nave Bayesian classifiers deployed at the leaves. When the sample data set is tiny, naive Bayesian classifiers are thought to perform better than decision trees. As we are handling big dataset, we expect to have better accuracy using Decision Tree Algorithm.

# Naïve Bayes Classification

To get our dataset through the algorithms, we need to change the categorical features back to numbers as the algorithms do not recognize strings or float, I performed replace function in the features: street class code, new area code and surface type code see below our current dataset which has been used to perform the Naïve Bayes algorithm and re-setting the index using reset function:



**Table 3:** Representing categorical features by numbers.

Dataset defined. To start with naïve bayes classification we need to split our data in two: Test and training set, our independent variable X is defined by feature new\_area\_code, Y is defined by our dependent variable surface\_type\_code, which is our target variable in our Jupyter notebook.



*Output after splitting the data into training and test*

After run the algorithm for the first time, we got an accuracy of 23% which is not viable. After the results, Standard Scaler has been applied in order to get better accuracy.

“Standardization of a dataset is a common requirement for many machine learning estimators: They might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).”

*Standard Scaler Formula*

The accuracy after applied standard scaler is better, we can conclude that after this procedure with an accuracy of 43% still not a good model but we have been done improvements in the model.

(Code is on Jupiter Notebook)

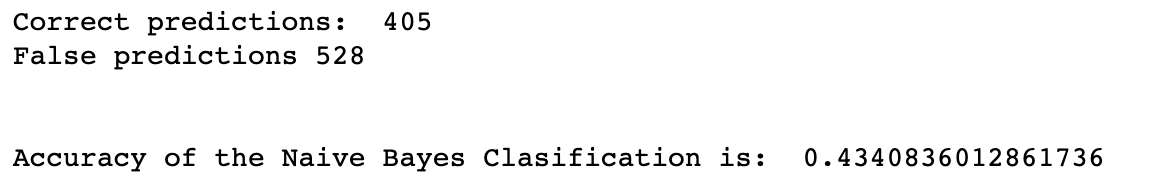
The method Cross Validation and Matrix Confuse was applied to have these insights.

## Confusion Matrix

When it comes to tackling classification difficulties, the confusion matrix is a prominent tool. It may be used to solve multiclass classification issues as well as binary classification problems.

What insights we gained after Confusion Matrix?

Well, our model can predict 405 correct predictions and 528 false predictions giving us the accuracy below:



*Accuracy results from our Jupyter Notebook –* ***Confusion Matrix***

# Cross Validation

Cross-validation is a statistical approach for estimating machine learning model performance. It's a way of determining how well the results of a statistical study will generalize to a different collection of data. In our Jupyter Notebook, we can find the results of cross validation applied to confirm our accuracy in Confusion Matrix (Cross Validation)

# Decision Tree Model

Why decision tree?

“By learning basic decision rules inferred from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data).

We start at the root of the tree when using Decision Trees to forecast a class label for a record. The values of the root attribute and the record's attribute are compared. We follow the branch that corresponds to that value and goes to the next node based on the comparison.”

(Géron, 2019)

With the Decision Tree Algorithm, We have done the same procedures, convert the target variables in arrays, split the data, display the data, created a decision tree classifier. The model accuracy performed was slightly better than the previous one, Naïve Bayes Classification. With accuracy of 50% Decision tree model has a better performance so far.

# Support Vector Machine

“Support Vector Machine (SVM) is one of the most useful supervised ML algorithms. It can be used for both classification and regression tasks.

Support Vector Machine is one of the classical machine learning techniques that can still help solve big data classification problems. Especially, it can help the multidomain applications in a big data environment. However, the support vector machine is mathematically complex and computationally expensive.” (Suthaharan, 2016)

# What is the result of SVM?

SVM choose the hyperplane that does maximum separation between classes, our data can be linear separable and controlling the errors of our models with hyper parameter C in 2, gamma as 0.9 to give more weightage to point closest to support vector we can perform SVM model with 100% accuracy!

# Conclusion

As we can conclude after the analysis of this dataset, we can predict which type of material any road has been made of it or which class it is being part of. As a matter of fact we can not make predictions when was the last time the streets was made, when was last time repaired and when it will need to be repaired again, the reason is that, we do not have relevant information, lots of missing data were missing in these features, we could perform KNN to deal with missing data but lots of relevant observations were missing and would take some time to request data to Dublin City Council, further work is required to establish this.