A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name:* | Kevin Scully |
| *Student Number* | 24275 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 24th November 2024 |
| *Date of Submission* | 18th November 2024 |
|  |  |

**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.

Contents

[Introduction to Problem Description, Motivation, and Objectives 3](#_Toc182663127)

[Characterization and Cleaning of Dataset, Training and Testing of ML Models 3](#_Toc182663128)

[Purpose of Hyperparameter Tuning and Application of Hyperparameter Tuning Technique 3](#_Toc182663129)

[Interpretation of Results, Code Description, Comments, Conclusions, Citations, and References 4](#_Toc182663130)

[References 5](#_Toc182663131)

# Introduction

A current major issue for the Irish government is the shortfall in supply of social housing to meet the ever-increasing demand. The lack of social housing has a direct impact on homelessness figures, increasing the number of families living on the street. (Clarke, et al., 2024) Therefore, social housing is a key pillar of current economic policy.

This report will aim to help the government departments make decisions on the construction of social housing by using machine learning models to analyse the following objective:

1. Predict the number of social houses constructed based on the values in the other columns of the dataset.
2. Compare the results of objective 1 from the results of objective 1 to develop two machine learning models that can predict the outcome of one feature based on the information in the other. (Decision Tree Classifier, Random Forest Classifier)

The dataset used for this analysis was the ‘Social Housing Construction Status Report Q2 2022’. This provided data on the construction of social homes within Q2 2022. The information within the data includes the funding programme, the local authority, the scheme/project name, the number of units constructed, the approved housing body, and whether the construction was on-site or completed. (Department of Housing, Local Government and Heritage, 2022)

# Characterization and Cleaning of Dataset, Training and Testing of ML Models (250)

## Data Cleansing

The dataset was analysed to understand the columns and values. There was 1,566 rows and 12 columns within the dataset. After analysing the missing values within each column, I found that the columns “Stage 1 Capital Appraisal”, “Stage 2 Pre-Planning”, “Stage 3 Pre-Tender design” and “Stage 4 Tender Report or Final Turnkey/CALF approval” had over 78% of the rows as null values in each column. These columns did not provide enough information to be analysed and were removed.

The columns “On Site” and “Completed” also had a significant number of missing values. A missing value here was understood to mean that the construction was not on site or completed. Therefore, these null values were replaced with “No”. Finally, the column “Approved Housing Body” had 42 missing values, which equated to 2.7% of all the rows. As this was an insignificant number of rows, these were dropped from the dataset.

## Data Pre-Processing

Now that the data was clean, it was ready for pre-processing. The data contained a mix of categorical and numerical data types. The label encoder was used to convert the categorical data types into numerical data types. When this was complete, a target variable was defined. As the objective of this analysis was to predict the features that would impact the number of social houses that got constructed, this was picked as the target variable, with all other columns therefore being defined as the feature variables.

The target variable was converted to a categorical data type. As there was varied data as values within the target variable, these were classified into ranges of the number of houses built. The ranges used were 0-25, 26-50, 51-75, 76-100 and 101+.

## Model Building and Evaluation

The target variable was tested across five different machine models using a test to training split of 10%, 15% and 25%. The five models were k-nearest neighbour, decision tree, random forest, logistic regression and support vector machine. The results were as follows:

A screenshot of a computer

Description automatically generated

An assessment of the classes showed an imbalanced dataset, with 79% of data falling into the 0-25 class. Therefore, accuracy would not be a good indicator on performance as it is based on all predictions, it would achieve a high percentage from accurately predicting the larger class, without predicting accurately on the smaller classes. As the F1 score considers both the precision and re-call, this was the parameter used for assessing model performance and cross-validation.

# Machine Learning Models and training

*“Characterization of data, pre-processing, explanation and description of techniques used for the variation in the accuracy across three training splits (10%, 15% and 25%) using cross validation techniques.*

*An excellent characterization and cleaning of the dataset that summarizes all details from source to fields. An excellent accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is used to test the generalizability of the model and it should justify the results in an excellent way.”*

# Purpose of Hyperparameter Tuning and Application of Hyperparameter Tuning Technique (250)

*“What is the primary purpose of hyperparameter tuning in machine learning? Could you elaborate on specific hyperparameter tuning techniques (e.g., GridSearchCV) applied to machine learning models to find optimal parameters?*

*Describes the purpose of hyperparameter tuning with clarity and describes hyperparameter tuning techniques with clarity, though may have some minor gaps or less detailed explanation.”*

# Interpretation of Results, Code Description, Comments, Conclusions, Citations, and References (250)

*“Interpret and explain the results obtained, discuss overfitting / underfitting / generalisation, provide a rationale for the chosen models and use visualisations to support your findings. Comments in Python code, conclusions of the assignment should be specified at the end of the report. Harvard Style must be used for citations and references.*

*An excellent interpretation and explanation of the results, code description, comments, conclusions, citations, and references based on problem specification and objectives. The results clearly state that the models are neither overfitted nor underfitted. An excellent defence is provided.”*

# References

**There are no sources in the current document.**