CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning (10 ETCS) |
| **Assessment Title:** | CA1 Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Caroline Floss Bresolin  Ewerlayne Debora da Silva Henrique |
| **Student Number:** | 2023190 |
|  | 2023262 |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** |  |
| **Date of Submission:** |  |

**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.

**CA 1 Project**

Caroline Floss Bresolin-2023190

Ewerlayne Debora da Silva Henrique-2023262

*Higher Diploma in Science in Data Analytics for Business Strategic Thinking*

*Dr. Muhammad Iqbal*

*CCT College Dublin, Ireland*

**TABLE OF CONTENTS**

[1. INTRODUCTION: 4](#_Toc151762692)

[2. EXPLORATORY DATA ANALYSIS (EDA) 4](#_Toc151762693)

[3. MACHINE LEARNING 5](#_Toc151762694)

[4. CONCLUSION 7](#_Toc151762695)

[5. REFERENCE LIST: 8](#_Toc151762696)

# INTRODUCTION:

The data set chosen for this project was obtained from the Kaggle library and is titled "Crime Data from 2020 to Present." He covers criminal incidents that occurred in Los Angeles in the years 2020, 2021, 2022, and 2023. Los Angeles, the most central and populous city in California, has become a critical location for this analysis. The primary goal is to broaden our understanding of critical factors that can strengthen security, emphasising common elements such as age, gender, and location.

# EXPLORATORY DATA ANALYSIS (EDA)

During the Análise Exploratória de Dados (AED) phase, we employ various visual techniques to gain deep insights into the data set. The primary goal was to understand the nature of the data, identify hidden patterns, and potential anomalies.

At the start of data exploration and cleaning, we use code to convert all instances of "n.a.", "?", "NA", "n/a", "na", "--" to numeric values (NaN). Following that, we identified low-importance columns and decided to remove them, as well as duplicate values. To improve visualisation and analysis, we adjusted the column headings for date, month, and year. In this process, we also removed columns that proved to be irrelevant to the analysis's goals.

When dealing with the column "Vict Sex" and its 106,314 numeric values, we implemented specific treatments. We added new columns to categorise sex as Masculino, Feminino, and Unidentified. Concerning the "Vict Age" column, we used the KNN inference technique to fill in the missing values, taking into account the values of nearby neighbours and their proximity to other points. However, we observed an increase in the median age for 39 years old, which did not reflect the most common age in our data set. To correct this discrepancy, we chose to replace the values by fashion, representing the most common age in the data set. This approach is common when dealing with categorical data.

After finishing the cleaning phase, we prepared the data for machine learning model training by dividing it into features (X) and variables (Y).

# MACHINE LEARNING

In our data, we use some methods such as:

* **Principal Component Analysis, or PCA**

PCA (Principal Component Analysis) was selected as the method to reduce dimensionality for several reasons (Hasan and Abdulazeez, 2021, pp 25). Also might be the most popular technique for dimensionality reduction with dense data (few zero values)(Brownlee, 2020).

* **Singular Value Decomposition, or SVD**

Is one of the most popular techniques for dimensionality reduction for sparse data (data with many zero values) (Brownlee, 2020).

* **Linear Discriminant Analysis, or LDA**,

It is a multi-class classification algorithm that can be used for dimensionality reduction.

The number of dimensions for the projection is limited to 1 and C-1, where C is the number of classes. In this case, our dataset is a binary classification problem (two classes), limiting the number of dimensions to 1 (Brownlee, 2020).

* **Locally Linear Embedding, or LLE**

Creates an embedding of the dataset and attempts to preserve the relationships between neighborhoods in the dataset (Brownlee, 2020).

* **Make\_Classification**

It is a function provided by the scikit-learn library in Python that allows you to generate synthetic datasets for classification tasks. It's often used for testing and prototyping machine learning models. This function creates a random dataset with specified characteristics, including the number of samples, features, informative features, redundant features, classes, and more.

* **k-Fold Cross-Validation**

(Brownlee, 2018)

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

Note that k-fold cross-validation is to evaluate the model design, not a particular training. Because you re-trained the model of the same design with different training set

# CONCLUSION

This work assisted us in putting into practise the concepts we learned in the classroom. It also allowed us to have a broad perspective with different points of view by analysing data banks from the past and present. This analysis provided us with relevant information for decision making. We were able to determine the locations, average age, and gender of the victims of the crimes. By presenting this analysis, we may be able to increase control in some areas and define strategies for reducing violence and providing security to the public.

We had the opportunity to test several models for data reduction and some machine learning tests to determine accuracy and whether the method was appropriate for our needs.

# REFERENCE LIST:

<https://www.kaggle.com/datasets/nathaniellybrand/los-angeles-crime-dataset-2020-present>

Hasan, B.M.S. and Abdulazeez, A.M., 2021. A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), pp.20-30.