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:

# Our dataset encompasses criminal incidents that transpired in Los Angeles from 2020 to 2023. Los Angeles, being the most central and populous city in California, assumes a crucial role in this analysis. Following a thorough examination of the data, our objective is to predict crime based on historical patterns.

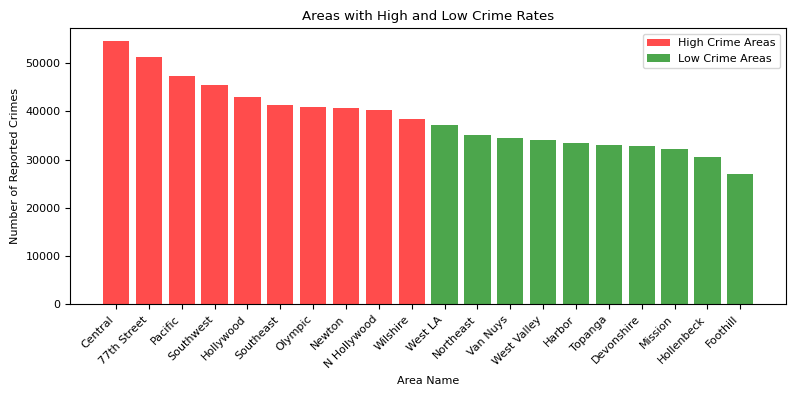
# EXPLORATORY DATA ANALYSIS (EDA)

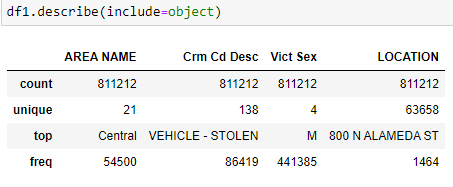
The original data set has 28 columns, and after defining our goal, we decided to drop some of them and keep only 'TIME OCC', 'AREA', 'AREA NAME', 'Crm Cd', 'Crm Cd Desc', 'Vict Age', 'Vict Sex', 'LOCATION', 'LAT', 'LON', 'Year', 'Month', 'Day'.

The first step is analysing the dataset using standard functions in order to identify null values, duplicates, and types to gain some information.

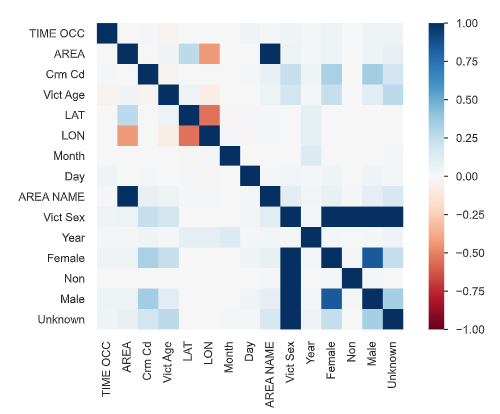
The null values we tried to deal using KNNImputer, but since it didn’t bring a good result, and also as we are dealing with a large dataset, it takes a long time to process, we decided to use the mode. The null values are in the columns age and victim sex. Using the mode makes sense for us because it would keep a similar distribution of our data. This approach is common when dealing with categorical data.

Applying EDA we gained some crucial insights. We noticed that since 2020 crime has increased and males are more likely to be a victim than women, This is curious because analysing the population of Los Angeles there are more women than men (97 men for 100 women (www.states101.com, n.d.)). The central area is the main spot to crime occurs and the top crime is vehicle stolen.

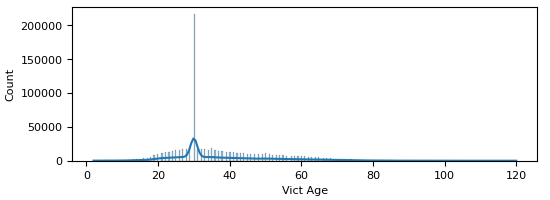




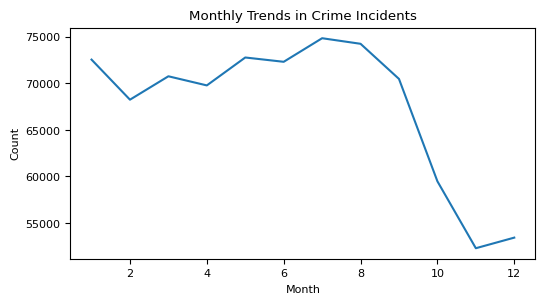
We couldn’t find relevant correlation between the features:



Crimes are more likely to occur with people aged between 25 to 40 years old:



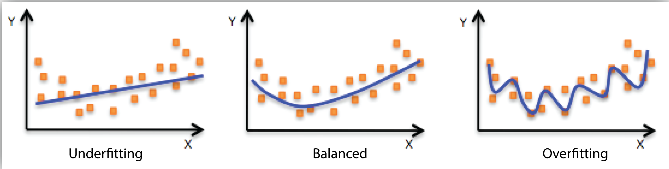
At the end of the year, we observed a decrease in crimes:



# MACHINE LEARNING

Once we want to predict likelihood of crimes, we choose the ‘crm cd’ (crime code) as y and the features 'TIME OCC', 'AREA', 'Vict Age', 'LAT', 'LON', 'Year', 'Month', 'Day', 'Female', 'Non', 'Male', 'Unknown' as X.

Even using some hyperparameters it may have a low accuracy due to its dimensionality. When DecisionTreeClassifier is applied, it hits a low accuracy of 0.24 which is called underfitting which is when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples (often called X) and the target values (often called Y) (docs.aws.amazon.com, n.d.):



In order to solve this issue, we applied dimensionality reduction and cross validation

Aqui explicar sobre isso. Explicar overfitting como ele pede. Falar sobre a diferença de accuracy entre split 10,20 e 30. O que está abaixo em amarelo, acho que da para apagar quando fazer essa parte.

* **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.

**DATA SOURCE:** The data set chosen for this project was obtained from the Kaggle library and is titled "Crime Data from 2020 to Present." (www.kaggle.com, n.d.)

# REFERENCE LIST:

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.

Brownlee, J. (2018). *A Gentle Introduction to k-fold Cross-Validation*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/k-fold-cross-validation/.

Brownlee, J. (2020). *6 Dimensionality Reduction Algorithms With Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/dimensionality-reduction-algorithms-with-python/.

www.states101.com. (n.d.). *Los Angeles County, California Gender Ratios*. [online] Available at: https://www.states101.com/gender-ratios/california/los-angeles-county.

www.kaggle.com. (n.d.). *Los Angeles Crime Dataset (2020 -- Present)*. [online] Available at: https://www.kaggle.com/datasets/nathaniellybrand/los-angeles-crime-dataset-2020-present.