A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Angelo Jonathan Cruz Hernandez |
| *Student Number* | 2024189 |
| *Module Title* | Machine Learning (10 ETCS) |
| *Assessment Title* | CA1 Project |
| *Assessment Due Date* | 21st April 2024 |
| *Date of Submission* | 21st April 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.

Introduction

M

Contents

[Introduction 3](#_Toc165066956)

[Motivation 3](#_Toc165066957)

[Problem Domain 3](#_Toc165066958)

[Problem Goals 3](#_Toc165066959)

[Data Management 3](#_Toc165066960)

[Characterization 3](#_Toc165066961)

[Pre-processing 4](#_Toc165066962)

[Cross Validation within training splits 20%, 25%, 30% 4](#_Toc165066963)

[Hyper Parameter Purpose 4](#_Toc165066964)

[Results 4](#_Toc165066965)

[Conclusion 5](#_Toc165066966)

[References 5](#_Toc165066967)

[Project Repository 6](#_Toc165066968)

# Introduction

With the raising of Artificial Intelligent many aspects of life seem to be improved and it appears that human task can be potentialized. Human Safety is one of the segments where this technology can help and not better enough, the employment forecasting preview increases in employment rate.

# Motivation

Human empowerment with AI technology is an exciting future, we have experienced many important events along our conscious life, and many have left a bigger footprint and now is the time of machine learning.

# Problem Domain

The current work is focused on human safety, specifically human mobility in the city of New York. “In 2022, there were 100,508 total car accidents over the course of the year, 37,848 of which involved an injury or fatality” (Ramirez, 2023)

# Problem Goals

The main goal of this project is to determine the model of a vehicle that has caused an accident. The vehicle model is the target variable in the dataset and is composed of a limited number of items therefore a classification model will perform better than a regression one.

# Data Management

### Characterization

The dataset has different kinds of null values where some of them appear as ‘unknow’, ‘Unspecified’ and ‘other’. The dataset is composed of more than 2 million rows with many columns that have more than 25% of missing values which means that they are not representing much of the content of the dataset. The trade-off of deleting them is a price that we pay in this current project to focus on improving more representative columns.

In the target variable ‘VEHICLE’ there were very low occurrences of some vehicle models of which also are not representative in comparison of the ones that have low or high occurrences. The models with less than 10 occurrences were deleted from the dataset.

### Pre-processing

The considered vehicle model categories were filled with its mode and mean for categorical and numerical columns where there were missing values.

The categorical columns must be encoded, it’s preferred to encode them using Hot-Encoding but there are two independent columns that has many different categories (ON\_STREET and FACTOR) for that reason Label Encode are best option even thought it should be used on the independent column.

The dataset has two columns representing the date and time when the accident occurs. From the date only month and day were taken because of its significance on important events and days off. The time was considered in minutes instead of the Time datatype.

### Cross Validation within training splits 20%, 25%, 30%

When using Cross Validation, there are random selected splits that are taken for the helper, and it divide in train and testing set that are used by the estimator. All the scores gotten in the loop process are then summarized in a final average. Cross Validation k-folds delete the necessity of divide the dataset in Training, Validation and Testing sets. Because on every running of the helper function the k-folds are selected at random the average score would change even in the same training split.

Applicating Cross Validation on the dataset and running it on all three testing splits, it gave the following best split for KNN, DecisionTreeClasifier and Random Forest models:

KNN test score: 0.5000

DTC test score: 0.4000

RFC test score: 0.4500

### Hyper Parameter Purpose

Hyper Parameters control the capacity of the model, how flexible the model is, how many degrees of freedom it has in fitting the data. Because every dataset is different the hyperparameters that were configured to fit a model will not work for different. Without hyperparameters tunning the model tent to be overfitted which means that doesn’t generalize results.

To find the optimal split for training and testing we use cross\_val\_score, the helper function to apply Cross Validation for every split, that is a way of hyperparameter tunning then we got all the scores for every split, we average them and conclude which split is better based on the higher average.

# Results

The models used have low accuracy on the testing split, however Random Forest performs slightly better than KNN and Decision Tree algorithms. The accuracy score decreases when the testing set increases and and both models have similar accuracy score on the training and testing splits even though we apply hyperparameter tunning using Cross Validation.

Overfitting refers to a model that performs good on training set but not that good on testing set so it does not generalize outputs. When looking at the scores both models got in training and testing as we say before, are close so our models are not suffering of overfitting.

Underfitting is the opposite of Overfitting, but the models’ scores are similar on both sets, so they are not performing as underfitted models. Having that the models are neither overfitted nor underfitted we can say that they generalize correctly even though the accuracy score is low.

Because our problem is referred to find a vehicle, which is categorical, a categorical model was chosen. There are several categorical models that could be used but **K Near Neighbourhoods**, **Decision Tree and Random Forest performed good** on the data, thus the data has more than 2 million observations.

A screenshot of a computer

Description automatically generated

A diagram of a vehicle

Description automatically generated

A screenshot of a computer

Description automatically generated

# Conclusion

Three models were encountered that perform with an accuracy score above 0.4, those models are KNN, Decision Tree and Random Forest, Random Forest has a slightly higher score which makes it the best option but there is still room for improvement. The score is not reflecting a model that possibly could be in production, there might be considerations and rechecking the data to make adjustments as to get better performance.

# References

Goyal, Chirag. “Missing Values | Treat Missing Values in Categorical Variables.” *Analytics Vidhya*, 27 Apr. 2021, www.analyticsvidhya.com/blog/2021/04/how-to-handle-missing-values-of-categorical-variables/.

Ramirez, Adam, and Christy Bieber. “NYC Car Accident Statistics in 2023 – Forbes Advisor.” *Www.forbes.com*, 30 June 2023, www.forbes.com/advisor/legal/nyc-car-accident-statistics/.

SciKit-Learn. “3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.21.3 Documentation.” *Scikit-Learn.org*, 2009, scikit-learn.org/stable/modules/cross\_validation.html.

# Project Repository

<https://github.com/crzjothda/machlearning-ca1>