OTAGO POLYTECHNIC AUCKLAND INTERNATIONAL CAMPUS

Project Planning

Crime Prediction Using Predictive Analytic

April Love Naviza
Vimitaben Mukeshchanadra Vaidya
Wisanu Boonrat

5/24/2020

Graduate Diploma in Information Technology

Executive Summary

This documentation will outline the process of planning and developing a machine learning model wherein different predictive modelling techniques will be applied. The project seeks to predict the crime in New Zealand by analysing the dataset and patterns that are likely to show future results.

The first part will provide an understanding of the project objectives, requirements, and scope of our work. An activity diagram will visually present the sequence of our activities during this project.

In the initial stage, the data exploration occurs. There will be 2 sections in handling the newly acquired dataset: Dataset Analysis and Process data wherein we will concentrate on feature selection and engineering features.

There are various types of algorithms that have been using for predictive policing. We have considered and evaluated other related projects to help us compare and select which algorithms will perform a more accurate prediction. The pros and cons of our chosen algorithms will be discussed in section 2.

We will perform model validation to ensure that the model is not underfitting or overfitting. And finally, execute model evaluation metrics to assess the performance of the model. We have identified some model evaluation metrics and display their advantages.

The main objective of this project is to obtain the different datasets, examine the data, identify which is the best fit, and build a model that will predict with an accuracy rate of 80-90%.

Table of Contents

Table of Tables

Table of Figures

1 Introduction

This project comprises of 4 phases that are business understanding, data acquisition, predictive modelling, and data visualization. The structure has pattered from the data science life cycle though the deployment stage will not be included. The data acquisition and data model execution are iterative and from time to time, overlapping. The key stakeholders to establish this project are project supervisor, a project manager, a developer, and a scrum master. In data visualization, the progress and the primary result will be presented and consulted to the project supervisor to re-examine the outcome.

The primary goal of this documentation is to demonstrate our project plan and explain the method of developing a crime predictive model. We aim to build an Artificial Intelligence (AI) model that will predict the location and time of a possible criminal offense, which is first-ever in New Zealand.

2 Data science life cycle

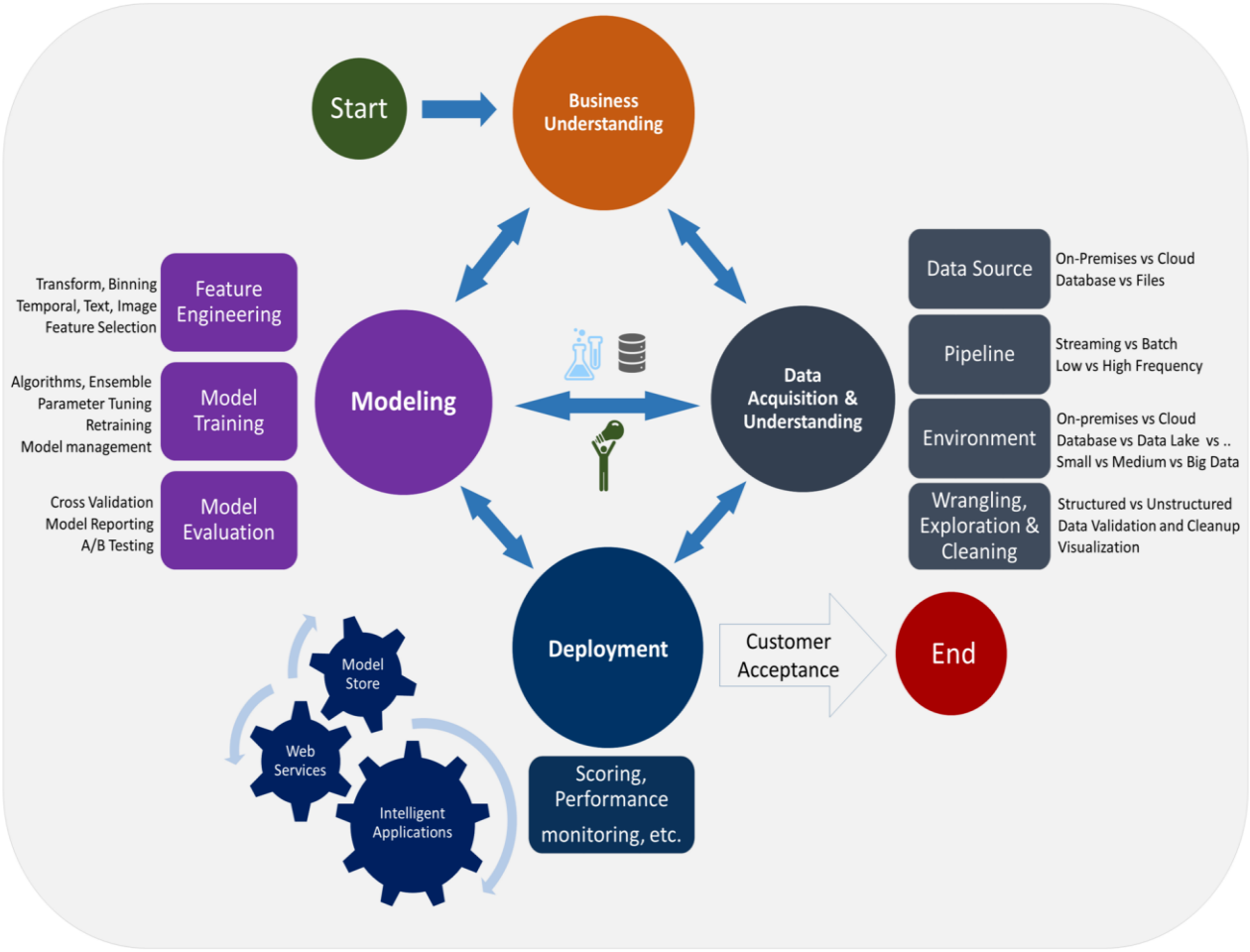

Figure 1. Data science life cycle Image Source: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Data science life cycle phases are an iterative process, where we can evaluate how the success of our model related to the original goal of the project.

The data science life cycle has the following different phase:

1. Business understanding
2. Data acquisition and understanding
3. Modelling
4. Deployment

2.1 Business understanding

This stage mainly focuses on understanding the goal and objective of the project:

- The main goal of this project is to develop a Machine Learning (ML) model that can predict the location and time when a crime may occur in New Zealand.
- To acquire relevant data from different sources and to engage in the process of creating our expected data product.
- To identify variables that could be used in the process of analytics and modeling.
- To discover and recommend different predictive modelling technique

2.1.1 Business Requirements

A. Dataset

- Acquire dataset from different resources.
- Appropriate datasets for the analysis should be learned and verified.

B. Model

- A model to predict the crime location and particular time with reasonable accuracy.

C. Business Report/ Dashboard

- There will be a dashboard for visualising the predicted crime.
- There will be a simple presentation of the progress and updates

2.1.2 Scope of the project

This project will concentrate on developing a crime predictive model specific for New Zealand's location and will focus on the South Auckland region. The crime type will be based on the ANZOC Group and will not present all the types of crime. The prediction of the model will be based on historical data and does not cover any unforeseen events. This project will include the data analysis, building and improving the model, model evaluation, and data visualization while the deployment in API will be part of the future work.

2.1.3 SWOT Analysis

The SWOT Analysis is used to determine the strengths, weaknesses, opportunities, and threats of the project strategy.

Table 1. SWOT analysis for Predictive Policing tool development.

	Positive	Negative
Internal Environment	**Strength** • There is a tool for visualization of the total number of victimization and the frequency of the crime for an hour and day of the week (*NZ Herald Crime Map*[1]) • Existing tools like Auror can prevent crime for retails stores (*Auror*[2]) • Have visibility in Victimisation Time and Place (*Policedata.nz*[3])	**Weakness** • There is no implemented tool that predicts crime in New Zealand • The existing tool like Auror is not designed for crime prediction • Restricted to a few crime types like theft or robbery only.
External Environment	**Opportunities** • Introduce to machine learning and artificial intelligence that is potential to predict crime in New Zealand region • Opportunity to effectively allocate resources of police officers	**Threats** • An existing predictive policing application like *PredPol*[4] • Unpredictable crime (ex. Mass shooting) • Sustaining the model

2.1.4 Tools for project execution

This project will be managed by using Jira to identify the stories, subtask, and tracking issues in each sprint. For coding and developing the predictive model, we will use Spyder, Pycharm, and Jupyterlab. To easily monitor the progress of the project, we have set up the WBS Gantt chart for Jira. It integrates the activities in Jira and allows all team members to view the entire project timeline. And for storing codes and version control, we will use Bitbucket because it is integrated into Jira. Having most of the tools in one application helps us to easily navigate, track, and update the activities without leaving the current tool.

[1] https://insights.nzherald.co.nz/apps/crime_maps/crime/index.html
[2] https://www.auror.co/
[3] https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place?nondesktop
[4] https://www.predpol.com/

2.1.5 Project Workflow

This activity diagram shows the development life cycle of the data science project with high-level processes.

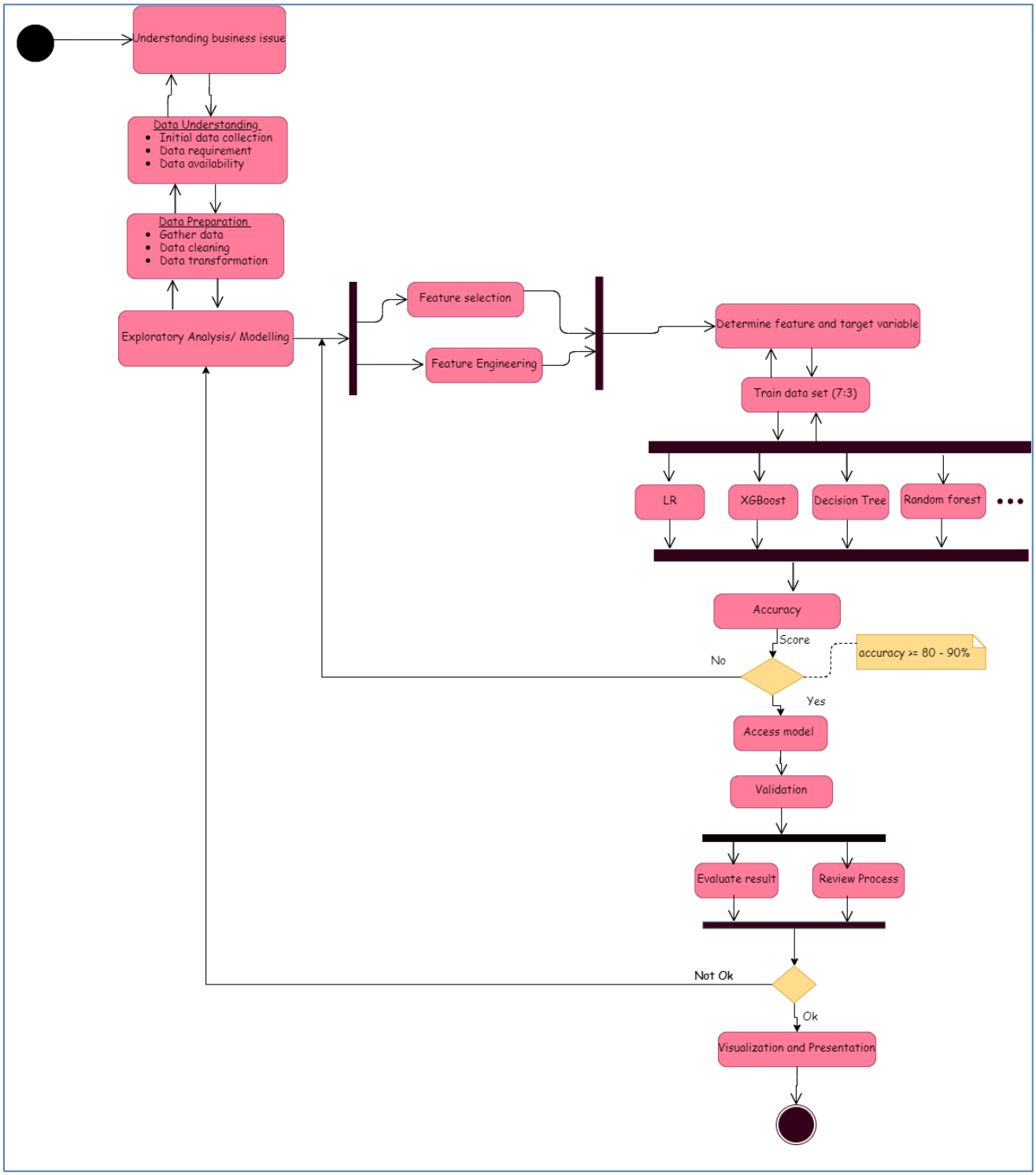

Figure 2. Activity diagram of the predictive modelling in a data science project.

The following individuals comprise the data analysis team. They are responsible for the analysis and development of the project.

Table 2. Team and responsibilities. RACI is an acronym derived from the four key responsibilities most typically used: Responsible, Accountable, Consulted, and Informed.

Team	Description	RACI
Project Supervisor	Defines the purpose and objective of the project Review and provides feedback to the team	C, I
Project Manager	Accountable in fulfilling the actual need for the project Organize the structure, tasks, and timeline of the project	A, C, I
Developer and Tester	Responsible for developing and evaluating the prototype Research and review factors the will affect the result	R, C, I
Scrum Master	Facilitate collaboration with the stakeholders Maintains the project tools	R, C, I

2.2 Data Acquisition and understanding

- In this stage, we have defined the objectives of the project, for that we used specific tools to understand data.
- In this stage, we performed different steps, such as data mining, where we collect different datasets from multiple resources.
- Data cleaning, where we identify
- Perform the data exploration stage is like the brainstorming of data analysis. This is where we can understand the patterns and bias in the dataset. Exploratory data analysis to understand the correlation between each variable, clean the dataset handle null values, and duplicate data.

2.2.1 Dataset Analysis

- In the data analysis stage, we collect data from different resources such as policedata.nz and the government websites.
- In this stage we perform data cleaning where we remove the missing values and delete duplicate data from datasets.
- Based on different crime datasets, we visualized and compare different graphs based on time variables (Month, year, and day) and location with respect to higher crimes.

2.2.2 Process Data

- After performing different cleaning operations and make the dataset meaning full the next stage is to transform data. In this stage we check each column variables are highly correlated or not by applying several feature selection techniques such as, SelectKBest, Extra Tree Classifier, Chi-square, and Mutual Information.
- We compared the score of these techniques and decided on the target and feature variables.

- We have applied the 7:3 ratio of data set based on that feature, and we have applied various machine learning algorithms on a train and test data sets.
- To improve the model accuracy, we performed feature engineering. This step helps to extract more information from existing data. Feature creation is part of feature engineering. We extract from date-time variable extract the hour of the day, day, month, week, and year which are created as a new feature. These new features may have a higher ability to explain the variance in training data which helps to improve the model accuracy.

2.3 Predictive Modelling

In this phase, we perform different machine learning algorithms aiming to collect the results for comparison. The model validation technique will be used to assess the accuracy of the models. We also use a variety of metrics to measure classification performance. After that, we will iterate the whole Machine Learning development process until the models reach acceptable accuracy.

2.3.1 Algorithm identification

In order to identify the algorithms for the classification problems, we reviewed research papers related to crime prediction to find the recommended algorithms. The accuracy, processing time, and mechanism of algorithms are the chosen criteria. Jain (2017) and Martin-Short (2019) have discussed the pros and cons of the algorithms, which shows in the following table.

Table 3. The algorithms for predictive modelling.

Algorithms	Types	Pros	Cons
Logistic Regression	Supervised	Overfitting can be addressed through regularization.	Can only learn linear hypothesis functions so are less suitable to complex relationships between features and target.
Naïve Bayes	Supervised	Fast and simple which is suitable for the baseline prediction.	Assumes that the features are independent, which is rarely true.
K-Nearest Neighbors (k-NN)	Supervised	Naturally handles multiclass classification and can learn complex decision boundaries.	Expensive and slow to predict new instances because the distance to all neighbors must be recalculated.
Support Vector Classifier (SVC)	Supervised	Fairly robust against overfitting, especially in higher dimensional space.	It doesn't perform well when we have a large data set because the required training time is higher.
Decision Tree Classifier	Supervised, Tree-based	Requires little data preparation - data typically does not need to be scaled.	It can be very non-robust, meaning that small changes in the training dataset can lead to quite

			major differences in the hypothesis function that gets learned.
XGBoost	Supervised, Tree-based, Ensemble	Robust to missing data, highly correlated features, and irrelevant features in much the same way as random forest.	May be more prone to overfitting than random forest - the main purpose of the boosting approach is to reduce bias, not variance
Random Forrest	Supervised, Tree-based, Ensemble	Highly flexible and generally very accurate.	Overfitting in case of noisy data.
Apriori	Unsupervised, Association analysis	It does not require labelled data as it is fully unsupervised.	Calculating support is also expensive because it has to go through the entire database.

2.3.2 Building the predictive model

In this stage we will perform the selected machine learning algorithms using the given datasets. There will be data sets for testing, training, and validation. The crime data will be divided into 70%-30% between the train and test parts. We fit the different models with the same datasets, so the results of the prediction can be compared. We will have a baseline model and compare the performance from the other model before selecting the best one.

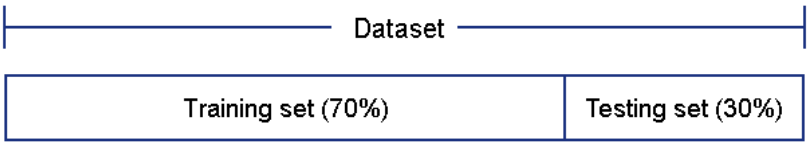

Figure 3. Divide the training and testing sets.

2.3.3 Model Validation

The basic problem of the machine learning model is the underfitting and overfitting, which can happen when we train the models with all records from the dataset.

- Underfitting: the model fails to capture the underlying trend of the data.
- Overfitting: the model is too sensitive and captures random patterns.

We can avoid this problem by using the cross-validation technique such as k-fold to reserve the data before performing the train/test split method. The reserved data will be a sample of unseen data for the models. This approach can simulate the situation when the models have to predict the result by using real-world inputs.

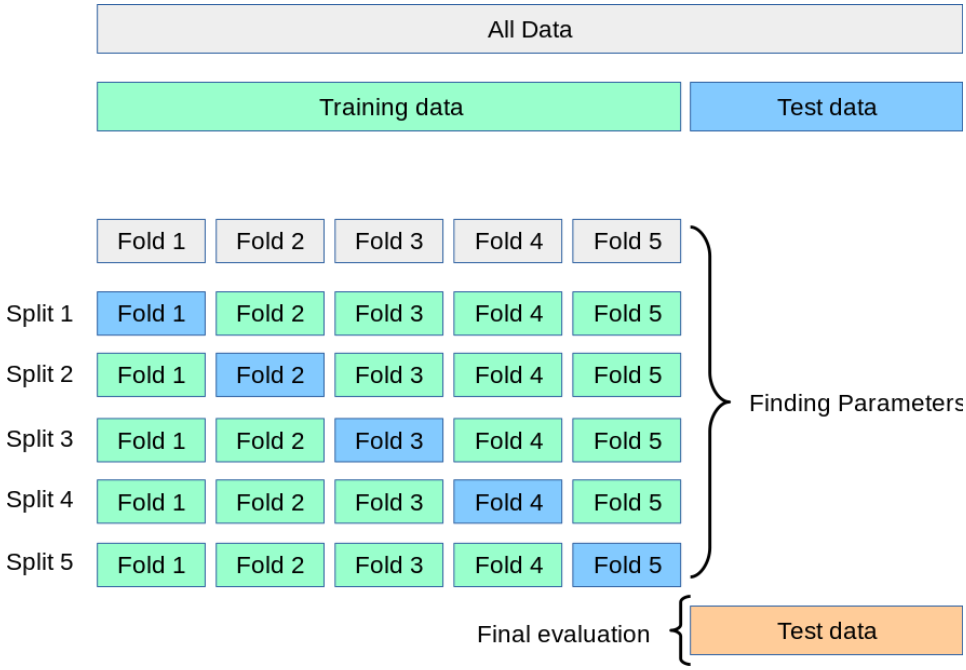

Figure 4. K-fold cross-validation. From Cross-validation: evaluating estimator performance by scikit-learn, 2019, https://scikit-learn.org/stable/modules/cross_validation.html.

2.3.4 Model Evaluation

The evaluation metrics for classification problems in case of multilabel tasks are suggested by scikit-learn. Each metric has different advantages that can verify the model's performance from different aspects.

Table 4. The evaluation metrics and their advantages.

Metrics	Advantages
Classification Report	• Show the precision, recall, f1-score of each class. And summarise the accuracy score.
ROC AUC Curves	• Summarize the trade-off between the true-positive rate and false-positive rate for a predictive model using different probability thresholds (Brownlee, 2018). • Be appropriate when the observations are balanced between each class (Brownlee, 2018).
Precision-Recall Curves	• Summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds (Brownlee, 2018). • Be appropriate for imbalanced datasets (Brownlee, 2018).
Confusion Metrix	• The number of correct and incorrect predictions are summarized with count values and broken down by each class (Brownlee, 2020).
Log Loss	• Heavily penalises classifiers that are confident about an incorrect classification (Collier, 2015).

2.3.5 Model iteration and improvement

In this stage, the development team will review the datasets, source code, algorithms, and results of crime prediction. We aim to collect feedback and look for the opportunity to improve the accuracy score by iterating the whole process starting from business understanding, data acquisition and understanding, modelling, and result visualisation.

2.3.6 Functional Requirements

A. General Requirements

- The tool will be built with Machine Learning or Deep Learning technologies.
- The tool will read the datasets from the CSV file.
- The tool will utilize historical crime datasets from various regions.

B. Data Mining and Analytics Requirements

- The tool will perform descriptive analysis.
- The tool will investigate the data and drop the outlier.
- The tool will perform the feature selection methods to reduce the inputs of the predictive models.
- The tool will transform and scale data to improve the performance of model fitting.
- The tool will divide data into train, validation, and test sets to apply to the predictive models.
- The tool will measure the accuracy of each model with various methods.
- The tool will have acceptable accuracy at around 80 – 90 %

2.3.7 Non-Functional Requirements

A. Operation

- The tool will have a special focus on the New Zealand crime dataset.
- The tool will compare the accuracy of the models' algorithm.
- The tool will incorporate additional datasets into the analysis to improve the accuracy of the prediction.
- The tool will visualise the result of crime prediction with geographic locations.

B. Usability

- The tool should run on both Windows and Linux platforms.
- The tool should provide an Application Programming Interface (API).
- The tool should provide a web or mobile application

C. Performance

- The tool should save and load the trained models.
- The tool should handle large datasets.
- The tool should handle imbalanced datasets.

2.4 Results Visualization

2.4.1 Communicate results

In every iteration all the results will be communicated and consulted to the project supervisor. This is to validate how the model performs and meet the expected outcome.

By the end of this project we will present a narrative summary of all the findings and outcomes in building the predictive model. We will illustrate the challenges, realization, and how we were able to attain the desired accuracy result.

3 Reference List

Brownlee, J. (2018, August 31). How to Use ROC Curves and Precision-Recall Curves for Classification in Python. Retrieved from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Brownlee, J. (2020, January 12). What is a Confusion Matrix in Machine Learning. Retrieved from https://machinelearningmastery.com/confusion-matrix-machine-learning/

Collier, A. (2015, December 14). Making Sense of Logarithmic Loss. Retrieved from https://datawookie.netlify.app/blog/2015/12/making-sense-of-logarithmic-loss/

Jain, R. (2017, March 24). A beginner's tutorial on the apriori algorithm in data mining with R implementation. Retrieved from https://www.hackerearth.com/blog/developers/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/

Martin-Short, R. (2019, February 24). Pros and cons of classical supervised ML algorithms. Retrieved from https://rmartinshort.jimdofree.com/2019/02/24/pros-and-cons-of-classical-supervised-ml-algorithms

Staff GP. Understanding the Lifecycle of a Data Analysis Project [Internet]. Northeastern University Graduate Programs. 2019 [cited 2020May23]. Available from: https://www.northeastern.edu/graduate/blog/data-analysis-project-lifecycle/

4 Appendices

4.1 Appendix 1

- **NZ Herald Crime Map**

https://insights.nzherald.co.nz/apps/crime_maps/crime/index.html

- **Auror**

https://www.auror.co/

- **Policedata.nz**

https://www.police.govt.nz/about-us/publications-statistics/data-and statistics/policedatanz/victimisation-time-and-place?nondesktop

- **Predpol**

https://www.predpol.com/