Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	B39DA – Applied Machine Learning
Type of assessment:	**Individual**
Coursework Title:	Coursework 2: Final Project Report
Student Name:	Tracey Hughes
Student ID Number:	H00338688

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature: *Tracey Hughes*

Date: *01/07/2022*

Copy this page and insert it into your coursework file in front of your title page.
For group assessment each group member must sign a separate form and all forms must be included with the group submission.

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

NORTH LANARKSHIRE COUNCIL

B39DA – Applied Machine Learning

TRACEY HUGHES (H00338688)

COURSEWORK - MACHINE LEARNING TECHNIQUES

Table of Contents

Introduction

North Lanarkshire (NL) is the fourth largest council by population in Scotland, covering around 181 square miles. Following a local government restructure, the authority was established in 1996, hosting a population of around 340,000 people and offering employment to 16,000 staff across its four main services: Chief Executive, Enterprise and Communities, Education and Family, Adult Health and Social Care. Although, the percentage of the staff recognised as 'ICT users' is approximately 30% of that overall staff, the consumption of digital services is increasing across the council due to transformation programmes and increased digital adoptions, including HR and field services solutions. Support for ICT users is managed through an ITIL service delivery function, acting as a single-entry point for all the councils ICT incidents and request, utilizing the IT Service Management (ITSM) system: ServiceNow. This report will assess the service desk incident data collected over the past 24 months.

Impact

The recent insourcing of the service delivery function at North Lanarkshire Council, supports The Plan for North Lanarkshire (North Lanarkshire Council, 2019) and the subsequent vision for ICT service delivery. The service has been outsourced for over a decade and as such, this has caused some disruption and uncertainty. As seen in Figure 1.1, the open incidents volumes have shown a significant increase over the past two years. Although staff have been insourced, many of the processes, standards and Service Level Agreements need to be reviewed. As such, this report will review the helpdesk statistics for a two-year period to analyse the correlation of incidents parameters and frequency to support the rebuilding of an internal service delivery function.

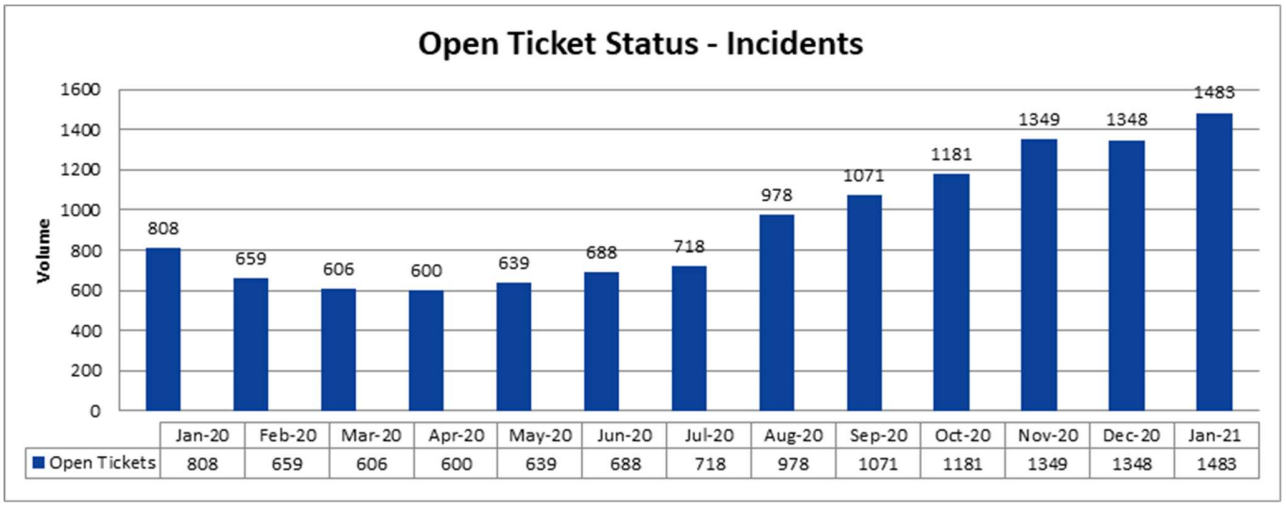

Figure 1.1 *NLC increased volumes of open Incident 20-21*

The Machine Learning Problem

There are several definitions on machine learning, from describing it as a 'Field of study that gives computers the ability to learn without being explicitly programmed' (Samuel, 1959) to the more detailed description of 'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance a task in T, as measured by P, improves with experience E.' (Jordon & Mictchell, 2015). Essentially machine learning performs pattern matching on data, producing algorithms based on the patterns to generalise and predict information about the data.

Focusing on the analysis of incident data, the aim is to predict an incident's P1 status based on priority, severity and urgency. Subsequently the correlation to meeting SLA for the same incident parameters will also be explored. This will be based on a dataset provided by North Lanarkshire Councils Service Desk Solution: ServiceNow.

Before progressing, the chosen problem statement should be assessed against any relevant criteria to ensure Machine Learning is the right approach to address the problem. The flow chart in Figure 1.2, highlights some of the common challenges in machine learning.

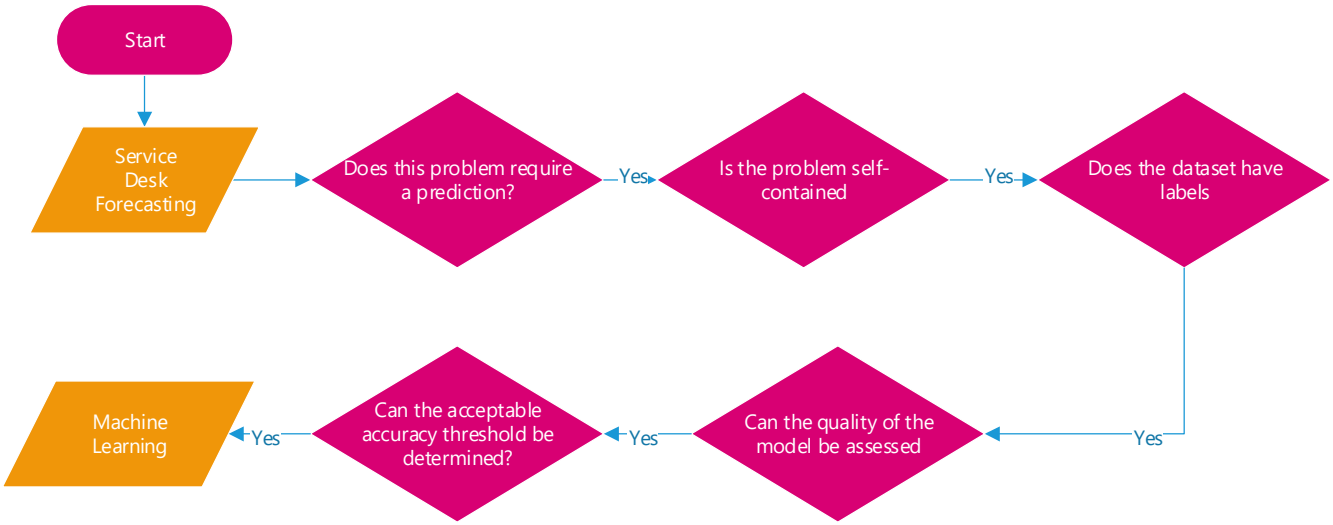

***Figure 1.2** defining alignment to machine learning problem* (LinkedIn, 2019).

- In this case, the aim is to analyse data to better understand trends allowing the prediction of P1 calls based on severity, urgency and priority, thus aligning to the first step of a machine learning problem.
- Further, the problem can assess as self-contained, with the required data to address the problem accessible within to the dataset.
- The data is structured, with identifiable input and output labels, which allows direct feedback to be achieved.
- Due to the size of the dataset, it's reasonable to suggest that the quality of the model can assessed, through application of an evaluation metric which is suitable to the model, for example support vector machines (SVM). Validation of the model can be achieved by splitting the dataset for training/test/validation, by 60% of data to training, 20% of data to test and 20% of data to validation.
- Understanding the business impact of this model, in terms of the benefits to rebuild the service desk function more accurately, can determine the success criteria of the model.
- Based on these factors, the project appears aligned to a machine learning approach.

Given that service desk dataset (or feature set) is pre-categorised with values which can be transformed from categorical to numeric values more suitable for machine learning, the labels are clearly defined in the form of input data and desired outputs. Where supervised learning is applied, the machine learning problem is best aligned to when the dependant (or target) variable is well-defined, meaning the expected prediction is clear (Vieira, et al., 2020). In this case the input data will be 'Severity', 'Priority' and 'Urgency' with the output data being 'P1 Incident' as seen in Figure 1.3.

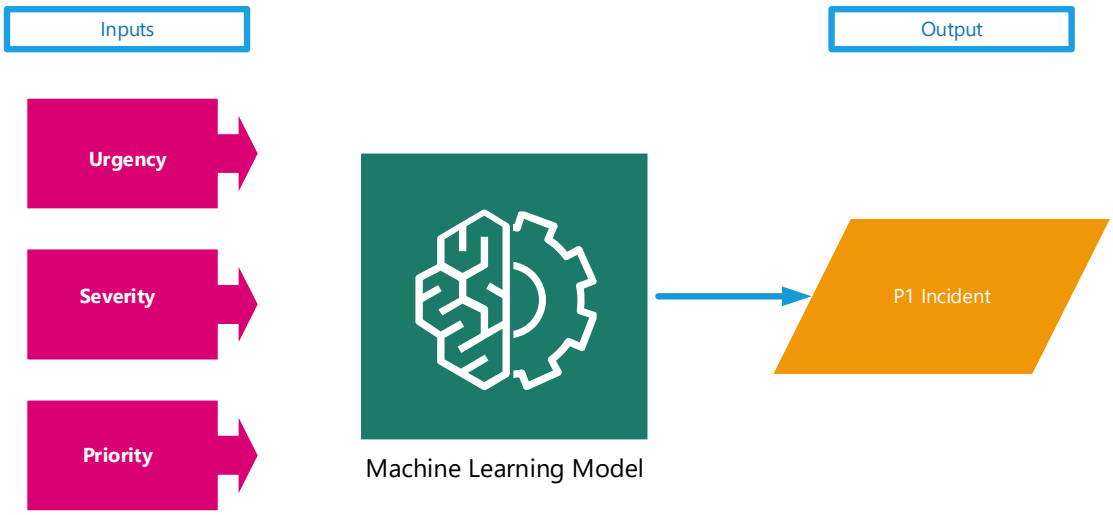

Figure 1.3 *Defined input and output data (labels)*

It is, therefore, proposed to implement a supervised learning model based on regression, starting with a simple linear regression model. Regression analysis is recognised as the most widely used method for the analysis of dependences. Linear regression can be singular (simpler) or based on multiple linear regression. Regression analyses the relationship between independent variables (x, features) and a single dependant variable (Y, labels) (Rong & Bao-Wen, 2018).

The aim of linear regression is to fit a line to the relationship between features and labels, a simple formula to express linear regression model is $y = mx + b$

Where y is the single value of the dependant variable, m is the slope, x is the value of the independent variable.

In the following equation, $Y = \beta_0 + \beta_1 X + \epsilon$, Y represents all observed values for the dependant variable, β_0 is the y-intercept or bias, β_1 is the slope or co-efficient, X is all observed values of independent variable and ϵ is the error.

Methodology

Data Extraction & Exploration

The approach to collecting the dataset(s) will be based on the steps outlined in Figure 1.4.

1. Preparing the Machine Learning Environment
2. Data acquisition: gather the right data to develop a machine learning model,
3. Data Cleaning: ensuring quality data to produce a quality model. Anonymising the data by ensuring the removal of personal data, process outliers, autonomizing sensitive data, addressing missing values.
4. Data Exploration: Discover and assess the data, provide an understanding of the type of data, data correlations, features and duplicate/missing values.

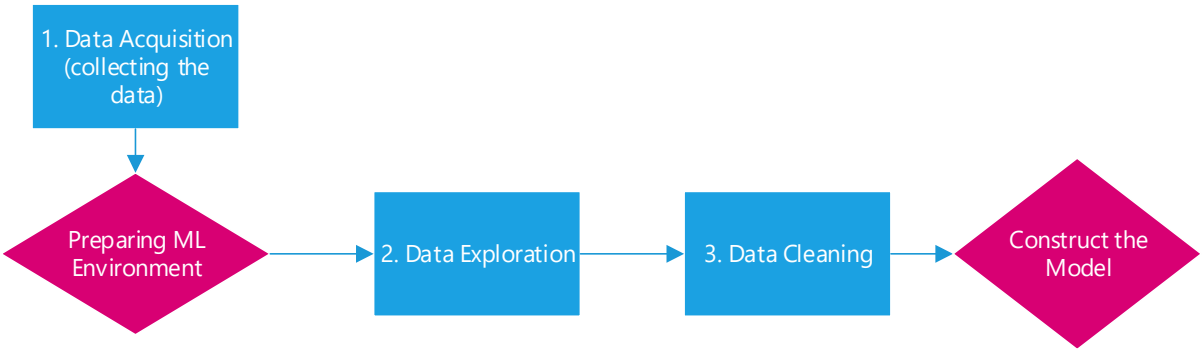

Figure 1.4 *Data acquisition and pre-processing flow*

The Machine Learning Environment

This project will be based on Python Jupyter Notebook (version 3.3.3), with files saved and shared from the Github repository - https://github.com/tah8/AML-Project-Service-Desk.git

To prepare for the data pre-processing steps, the required libraries must be imported for use. When attempting to import the libraries, a syntax error was received indicating libraries weren't installed, therefore the pip install feature was used to install all required libraries.

In [1]:
```
pip install numpy
```

```
Collecting numpy
  Downloading numpy-1.21.6-cp37-cp37m-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (15.7 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 15.7/15.7 MB 37.2 MB/s eta 0:00:0000:0100:01
Installing collected packages: numpy
Successfully installed numpy-1.21.6
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
pip install pandas
```

```
Collecting pandas
  Downloading pandas-1.3.5-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (11.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.3/11.3 MB 35.3 MB/s eta 0:00:0000:0100:01
Requirement already satisfied: python-dateutil>=2.7.3 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from pandas) (2.8.2)
Requirement already satisfied: numpy>=1.17.3 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from pandas) (1.21.6)
Requirement already satisfied: pytz>=2017.3 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from pandas) (2022.1)
Requirement already satisfied: six>=1.5 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from python-dateutil>=2.7.3->pandas) (1.16.0)
Installing collected packages: pandas
Successfully installed pandas-1.3.5
Note: you may need to restart the kernel to use updated packages.
```

In [18]:
```
pip install matplotlib
```

```
Requirement already satisfied: matplotlib in /srv/conda/envs/notebook/lib/python3.7/site-packages (3.5.2)
Requirement already satisfied: fonttools>=4.22.0 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (4.33.3)
Requirement already satisfied: cycler>=0.10 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: python-dateutil>=2.7 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: packaging>=20.0 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (21.3)
Requirement already satisfied: numpy>=1.17 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (1.21.6)
Requirement already satisfied: kiwisolver>=1.0.1 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (1.4.3)
Requirement already satisfied: pillow>=6.2.0 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (9.1.1)
Requirement already satisfied: pyparsing>=2.2.1 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from matplotlib) (3.0.7)
Requirement already satisfied: typing-extensions in /srv/conda/envs/notebook/lib/python3.7/site-packages (from kiwisolver>=1.0.1->matplotlib) (4.1.1)
Requirement already satisfied: six>=1.5 in /srv/conda/envs/notebook/lib/python3.7/site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [4]:
```
pip install sklearn
```

```
Successfully installed joblib-1.1.0 scikit-learn-1.0.2 scipy-1.7.3 sklearn-0.0 threadpoolctl-3.1.0
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
#These libraries are reqyured to manage the servicedesk dataset
# Numpy: used for large, multi-dimensional arrays and matrices, and for high-level mathematical functions
# Pandas: used for data manipulation and analysis
# matplotlib: used for visualisation and plotting graph/image/etc
# sklearn for splitting training data
# sklearn for linear regression model
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import sklearn.linear_model as skl_lm
from sklearn.model_selection import train_test_split
```

Data Acquisition

Data acquisition has been progressed through a collection of structured data from the service desk system (Service Now), derived from database tables and the last two years' service reports. This report will focus on a dataset with a single file, in csv format.

The file (Dataset 2) provides the detail of all incidents logged for North Lanarkshire Council, consisting of 14 fields with the dataset containing over 4500 records. A summary of each field is provided in Figure 1.5.

Field	Description	Data Type	Comments
Incident Number	Unique Identifier	Text	Drop – not required
Company	Due to managed service, company identifier	Text	Drop – not required
Summary	Description of issue	Text	Drop – not required
State	Current state of the incident (Active, Closed, Resolved)	Categorical Data	Required
Priority	Priority assigned to incident (1 – Critical, 2-High, 3 – Moderate, 4 – Low)	Categorical Data	Required
Assigned Group	Resolver group assigned to progress the incident	Categorical Data	Possible of interest
Severity	Severity of impact of the incident	Categorical Data	Required
P1 Incident	If the incident is categorised as a Priority 1 Incident with enhanced SLA	Boolean, Categorical Data	Required
Problem No	The problem number associated if the incident is part of a problem record	Text	Drop – not required
Vender Ticket Ref	The ticket reference associated if a call has been raised with 3rd party support	Text	Drop – not required
Urgency	The urgency of the incident (1 – Critical, 2-High, 3 – Medium, 4 – low)	Categorical Data	Required
Made SLA	Did the call meet the associated SLA	Boolean, Categorical Data	Required
Business Service	The impacted business service	Text, possible Categorical Data	Drop – not required
Reassignment Count	The number of times the call was reassigned to a different resolved group	Numeric	Possible of interest

***Figure 1.5** – Dataset 2– Detailed list of Incidents*

As seen above, the dataset contains a variety of different data types (text, numerical, Boolean, categorical), an extract of the raw dataset can be found in Figure 1.6.

Number	Company	Summary	State	Priority	Assigned Group	Severity	P1 Incident	Problem	Vendor Ticket Ref	Urgency	Made SLA	Business Service	Reassignment count
INC0028452	North Lanarkshire Council	NLC - No Network Connectivity	Closed	4 - Low	Schools Telecoms	3 - Low	FALSE			4 - Low	TRUE	Network Service	1
INC0148122	North Lanarkshire Council	NLC SCH - Telephone not operating Cathedral Primary	Closed	3 - Moderate	Schools Telecoms	3 - Low	FALSE			3 - Medium	TRUE	Telephony Service	2
INC0259518	North Lanarkshire Council	NLC - Urgent - Website access denial	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	1
INC0324680	North Lanarkshire Council	NLC - WinScp	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	4
INC0309371	North Lanarkshire Council	NLC - pages not loading	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	1
INC0259593	North Lanarkshire Council	NLC - Certificate has expired	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	1
INC0057391	North Lanarkshire Council	NLC SCH - STQ access forbidden	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Desktop Service	2
INC0093258	North Lanarkshire Council	NLC - 403 Error	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Desktop Service	1
INC0136834	North Lanarkshire Council	NLC - no outbound calls Calderhead	Closed	3 - Moderate	Telecoms	3 - Low	FALSE			3 - Medium	TRUE	Telephony Service	2
INC0237041	North Lanarkshire Council	NLC - Wifi Issue - Fleming House	Closed	3 - Moderate	Telecoms	3 - Low	FALSE			3 - Medium	TRUE	Network Service	1
INC0176540	North Lanarkshire Council	NLC - INC5065973 - Caird Data Centre - Firewall	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	2
INC0268778	North Lanarkshire Council	NLC - Windows Edge error message for external sites.	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	2
INC0343207	North Lanarkshire Council	NLC SCH - Seemis not available	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	8
INC0315624	North Lanarkshire Council	NLC SCH - Internet access	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	1
INC0156433	North Lanarkshire Council	NLC - Symology Error	Closed	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Desktop Service	1
INC0383800	North Lanarkshire Council	NLC - Unable to access Unity2016	Resolved	3 - Moderate	Secure Access	3 - Low	FALSE			3 - Medium	TRUE	Network Service	1
INC0136952	North Lanarkshire Council	NLc - Cisco Fast Dial	Closed	3 - Moderate	Telecoms	3 - Low	FALSE			3 - Medium	TRUE	Telephony Service	1
INC0018463	North Lanarkshire	NLC SCH - Intermittent	Closed	4 - Low	Telecoms	3 - Low	FALSE			4 - Low	TRUE	Network Service	3

***Figure 1.6** – Dataset 2 extract– Detailed list of Incidents*

Data Encoding

To improve the dataset, those highlighted as categorical data, have been manipulated to drop the text, leaving the numeric representation (Priority, Severity and Urgency). In Boolean fields (Priority 1 Incident and Made SLA), 'False' was changed to 0, and 1 represents 'True'. Further, the Assigned group has been converted to numbers (1 - Schools Telecoms, 2 - Secure Access, 3 – Telecoms). The procedure of changing the raw data enhance its suitability to machine learning can also be referred to as 'feature engineering'. This results in a transformed dataset or 'feature set' which will be used to train the model (Vieira, et al., 2020).

The dataset was saved as dataset3, an extract of the resulting dataset is shown in Figure 1.7.

State	Priority	Assigned Group	Severity	P1 Incident	Urgency	Made SLA	Reassignment count
Closed	4	1	3	0	4	1	1
Closed	3	1	3	0	3	1	2
Closed	3	2	3	0	3	1	1
Closed	3	2	3	0	3	1	4
Closed	3	2	3	0	3	1	1
Closed	3	2	3	0	3	1	1
Closed	3	2	3	0	3	1	2
Closed	3	2	3	0	3	1	1
Closed	3	3	3	0	3	1	2
Closed	3	3	3	0	3	1	1
Closed	3	2	3	0	3	1	2
Closed	3	2	3	0	3	1	2
Closed	3	2	3	0	3	1	8
Closed	3	2	3	0	3	1	1
Closed	3	2	3	0	3	1	1
Resolved	3	2	3	0	3	1	1

***Figure 1.7** – Dataset 3 extract –Following data pre-processing*

Data Exploration

Exploratory data analysis is used to better understand raw data, build knowledge of which features in the dataset will be useful and support the data cleaning in the steps that follow. The first thing to point out is that Priority, Assigned Group, Severity, Urgency and Made SLA are all categorical features, with the only continuous feature being Reassignment count and the target variable being P1 Incident.

The first step can be performed by ascertaining counts or distribution of the variables, supporting the understanding of the shape of the data for input features and target variables. Subsequently the data types of each feature would be assessed, identify missing and duplicate data as well as correlations. Involving the reformatting of data, creating/combining datasets, removing outliers and finding missing values, the goal of this data preparation is to increase the data quality and eliminate bias.

The dataset must first be uploaded and read into the python Notebook.

```
In [3]: # Import the service desk dataset
        # head visualise the first five rows of the dataset
        servicedesk = pd.read_csv('Dataset3.csv')
        servicedesk.head()
```

Out[3]:

	State	Priority	Assigned Group	Severity	P1 Incident	Urgency	Made SLA	Reassignment count
0	Closed	4	1	3	0	4	1	1
1	Closed	3	1	3	0	3	1	2
2	Closed	3	2	3	0	3	1	1
3	Closed	3	2	3	0	3	1	4
4	Closed	3	2	3	0	3	1	1

To understand the shape of the data, continuous features were explored

```
In [4]: #explore continious feature
        servicedesk.describe()
```

Out[4]:

	Priority	Assigned Group	Severity	P1 Incident	Urgency	Made SLA	Reassignment count
count	4518.000000	4518.000000	4518.0	4518.000000	4518.000000	4518.000000	4518.000000
mean	3.027003	2.066180	3.0	0.014830	3.026782	0.948207	2.180832
std	0.362238	0.663614	0.0	0.120884	0.362560	0.221633	1.779602
min	1.000000	1.000000	3.0	0.000000	1.000000	0.000000	0.000000
25%	3.000000	2.000000	3.0	0.000000	3.000000	1.000000	1.000000
50%	3.000000	2.000000	3.0	0.000000	3.000000	1.000000	2.000000
75%	3.000000	3.000000	3.0	0.000000	3.000000	1.000000	3.000000
max	4.000000	3.000000	3.0	1.000000	4.000000	1.000000	17.000000

Figure 1.8 *– Explore continuous features*

In Figure 1.8, the count is equal across all fields, which validates there are no missing values in the dataset. The target variables of 'Made SLA' and 'P1 Incident' are binary (either 0 or 1) which means we can use this what percentage of calls achieved there SLA (94%) and which were assigned P1 (0.14%).

Grouping by 'Made SLA' and 'P1 Incident' in Figure 1.9, shows that 'Severity' doesn't change value, and is always '3'. This feature can be dropped as there is no correlation to explore.

```
In [8]: servicedesk.groupby('Made SLA').mean()
```

Out[8]:

Made SLA	Priority	Assigned Group	Severity	P1 Incident	Urgency	Reassignment count
0	2.948718	2.064103	3.0	0.047009	2.948718	2.132479
1	3.031279	2.066293	3.0	0.013072	3.031046	2.183473

```
In [9]: servicedesk.groupby('P1 Incident').mean()
```

Out[9]:

P1 Incident	Priority	Assigned Group	Severity	Urgency	Made SLA	Reassignment count
0	3.052348	2.066277	3.0	3.052123	0.949899	2.178836
1	1.343284	2.059701	3.0	1.343284	0.835821	2.313433

```
In [10]: servicedesk.groupby('Severity').mean()
```

Out[10]:

Severity	Priority	Assigned Group	P1 Incident	Urgency	Made SLA	Reassignment count
3	3.027003	2.06618	0.01483	3.026782	0.948207	2.180832

```
In [11]: servicedesk.groupby('Urgency').mean()
```

Out[11]:

Urgency	Priority	Assigned Group	Severity	P1 Incident	Made SLA	Reassignment count
1	1.000000	2.000000	3.0	1.000000	0.826923	2.211538
2	2.012195	2.036585	3.0	0.085366	0.817073	2.463415
3	3.000000	2.092225	3.0	0.001962	0.953642	2.063527
4	4.000000	1.739414	3.0	0.000000	0.931596	3.657980

```
In [12]: servicedesk.groupby('Priority').mean()
```

Out[12]:

Priority	Assigned Group	Severity	P1 Incident	Urgency	Made SLA	Reassignment count
1	2.000000	3.0	1.000000	1.000000	0.826923	2.211538
2	2.037037	3.0	0.086420	2.000000	0.814815	2.481481
3	2.092202	3.0	0.001962	2.999755	0.953654	2.063266
4	1.739414	3.0	0.000000	4.000000	0.931596	3.657980

Figure 1.9 *– Explore data using group by*

To plot the data, continuous features can be dropped as displayed in Figure 2.0, leaving only numerical values for all features.

```
In [7]:
#drop all excess and continious features
cont_feat=['State', 'Reassignment count']
servicedesk.drop(cont_feat, axis=1, inplace=True)
servicedesk.head()
```

Out[7]:

	Priority	Assigned Group	Severity	P1 Incident	Urgency	Made SLA
0	4	1	3	0	4	1
1	3	1	3	0	3	1
2	3	2	3	0	3	1
3	3	2	3	0	3	1
4	3	2	3	0	3	1

```
In [8]:
servicedesk.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4518 entries, 0 to 4517
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Priority        4518 non-null   int64
 1   Assigned Group  4518 non-null   int64
 2   Severity        4518 non-null   int64
 3   P1 Incident     4518 non-null   int64
 4   Urgency         4518 non-null   int64
 5   Made SLA        4518 non-null   int64
dtypes: int64(6)
memory usage: 211.9 KB
```

Figure 2.0 *– Drop features and display data type*

This would allow scatter plots to be produced for more complex datasets where outliers could be more difficult to ascertain. This is where normalisation could also be applied to a more complex dataset if required through data scaling using min/max normalisation or standardisation (z-score). Linear regression models can be sensitive to feature scaling.

Data Cleaning

Through data collected, the aim is to develop a dataset containing details on incidents and the association of P1 Incident.

Data cleaning is imperative to ensure the machine learning model is successful, bias and incomplete data will impact the predictions. Data cleaning should also remove irrelevant data or if necessary, fill in missing data. Columns which do not contribute to the desired forecasted outcomes and outliers can be removed. This will be achieved by analysing the extracted columns and rows of data to ascertain any data pertinent to the forecast model. A variety of techniques will be applied, including manual assessment and review and assessing the shape/correlation of the data using python features (describe group by).

Based on the Figure 1.4, the highlighted fields were dropped as they did not present relevance to the desired outputs. Further, the remaining data was checked for missing values and found to have no missing values (other than Problem No and Vendor Ticket Ref).

Although this dataset has no missing values, (as validated in Figure 2.0), the impact of any missing values would be assessed in terms of the volume, category/pattern and reason to determine the right approach. Missing or incomplete data could be addressed through various data removal techniques or retained through a 'technique denoted as imputation' (Emmanuel, et al., 2021). A 'missingness mechanism' was introduced by Rubin in 1976, is widely used today (Rubin, 1976), with some of the 'common mechanisms including Missing Not at Random (MNAR), Missing at Random (MAR, and Missing Completely at Random (MCAR)' (Mostafa, 2021). These techniques replace missing values with statistical estimates, time series data or categorical data techniques.

Model Validation

To ensure the model is validated with new, unseen data, a separate dataset must be used to test the model's accuracy. It is common to split the dataset into three separate sections; Training dataset (to train the model), Validation dataset (data used to select the best model) and Testing dataset (evaluate the model).

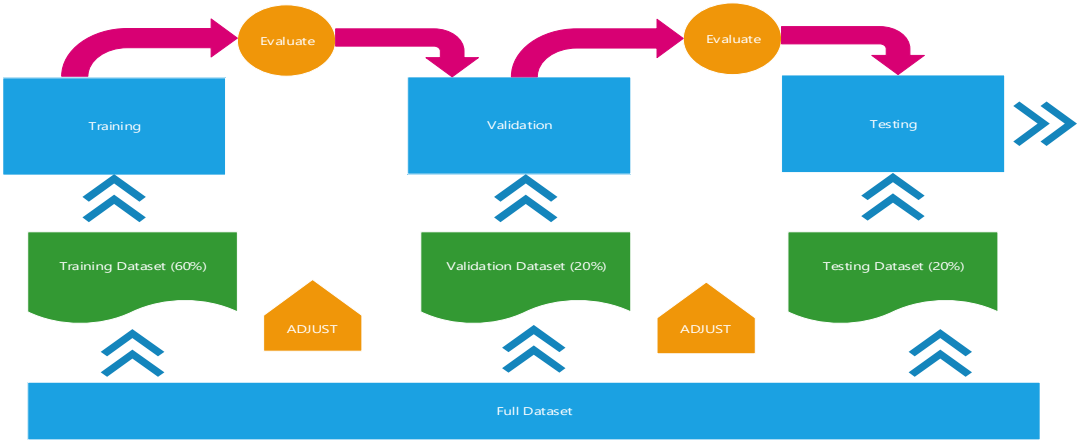

Figure 2.1 *Splitting data*

```
In [14]: features = servicedesk.drop('P1 Incident', axis=1)
         labels = servicedesk['P1 Incident']

         #Stage one split 60% to training set and 40% to training set
         x_train, x_test, y_train, y_test = train_test_split(features, labels, test_size=0.4, random_state=42)
         # Stage 2 split training set into 20% validation and 20% Test
         x_test, x_val, y_test, y_val = train_test_split(x_test, y_test, test_size=0.5, random_state=42)

In [16]: print(len(labels), len(y_train), len(y_val), len(y_test))

         4518 2710 904 904
```

***Figure 2.2** – Two stage data separation*

In Figure 2.2, the data is split in two stages due to the limitation in the algorithm used. Stage one, splits the data between training data and test data in a 60/40 ratio. The training dataset is then split again with half for testing and half for validation. This produces an overall 60/20/20 split of the dataset, validated through the 'print length' feature, total length is 4518, training dataset 2710, test and validation set are both 904.

The selection of the model complexity can lead to inadequate generalisation, which in the case of lack of complexity can lead to underfitting, generating high number of errors or over-fitting where the model is too complex, where low error volumes occur but the model does not generalise.

Results

Linear regression is based on the input variable(feature) and output variable (label) having a 'one-to-one relationship', which is often use in analysis of data but can lead to oversimplifications in practical use as it is unable to deal with complexity. (Ray, 2019).

The service desk dataset has multiple features, these could be analysed independently to the label, or for greater accuracy these would be applied using Multiple linear regression, the fore increasing the complexity of the model.

Due to several issues running code in the notebook, I was unable to run the code required visualising the linear regression model. Due to the simplicity of the linear regression model, size of the dataset and research completed, it is likely that the model may present some

underfitting impacting on the overall results. This can not be substantiated as no other models were applied to the data and no cross validation was completed.

Discussion

It's unclear if the size of the data set impacted the challenges in running the dataset through the preparation and visualisations attempted, or if this was due to corporate restrictions on the device used or simply lack of expertise in python.

There are several other areas within the service desk data where machine learning could prove beneficial and could be later explored using other models. These would include analysis of patterns/relationships of the number of open calls in each of the service desk queues in correlation to the number of staff or the disciplines involved in the queue, perhaps more aligned to the classification method of supervised learning.

Further, this could be extended to analyse the incident description to align it to the correct queue, although this would be more complex, based on unsupervised learning based, perhaps on clustering.

Conclusion

Although the exercise to understand the service desk dataset did not lead to a reliable prediction through the machine learning model, a better understanding of the data was achieved. The lack of correlation between severity and any of the other features was evidenced. Despite the limitations of the study, the benefits of using machine learning on this type of data is evident.

References

Emmanuel, T. et al., 2021. A survey on missing data in machine learning.. *Journal of Big Data,* 8(1), pp. 1-37.

Jordon, M. I. & Mictchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects.. *Science,* 349(6245), pp. 225-260.

LinkedIn, 2019. *Applied Machine Learning: Foundations.* [Online]
Available at: https://www.linkedin.com/learning-login/share?account=2374954&forceAccount=false&redirect=https%3A%2F%2Fwww.linkedin.com%2Flearning%2Fapplied-machine-learning-foundations%3Ftrk%3Dshare_ent_url%26shareId%3DMPRr%252FIvySa2gIvxbujANbw%253D%253D
[Accessed 05 2022].

Mostafa, S. M., 2021. Towards improving machine learning algorithms accuracy by benefiting from similarities between cases. *Journal of Intelligent & Fuzzy Systems,* Volume 20, pp. 947-972.

North Lanarkshire Council, 2019. *The Plan for North Lanarkshire.* [Online]
Available at: https://www.northlanarkshire.gov.uk/your-council/council-strategies-and-plans/council-strategies/plan-north-lanarkshire
[Accessed 01 June 2022].

Ray, S., 2019. A quick review of machine learning algorithms. *International conference on machine learning, big data, cloud and parallel computing (COMITCon),* pp. 35-39.

Rong, S. & Bao-Wen, Z., 2018. The research of regression model in machine learning field. *MATEC Web of Conferences,* Volume 176, p. 01033.

Rubin, D. B., 1976. Inference and missing data. *Biometrika,* 63(3), pp. 581-592.

Samuel, A., 1959. *Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed..* s.l.:s.n.

Vieira, S., Pinaya, W. H. & Mechelli, A., 2020. Main concepts in machine learning. . In: *In Machine learning.* s.l.:Academic Press, pp. 21-44.