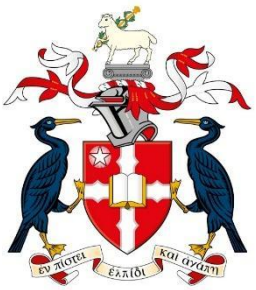
Pandemic Predictive Analytics using Machine Learning



LIVERPOOL HOPE UNIVERSITY

BSc Computer Science

2020

17005609, Ciarán Adams

Abstract

The aim of this study is to determine if machine learning can be effective at predicting pandemics, in this instance the prediction will be the deaths, cases and recoveries of the current Covid-19 Pandemic. This will be created by a single developer and no funding with a limited dataset that does not include factors which may contribute to the real world outcome, as this contributes to the idea that if the system is in any form accurate that systems that are funded and worked by large teams of developers will have a much more accurate system that can help predict Pandemics in the future.

Furthermore, the uses and effects of Machine Learning will be discussed from economics, Warfare, Medicine and even Governments to grasp a better understanding of how powerful machine learning. Two Models will be used to compete against one another helping grasp which Model is more effective The Models are the support vector machine (SVM) and Linear Regression Model (LRM). These models where not chosen as they were the best fit for this role, SVM was chosen as it fits quite well with a prediction such as this however Linear Regression was mainly chosen to compare the differences between these models. Grouping studies together of a similar nature signifies the relevance of the studies and how close they can be compared for a more reliable outcome.

Table of Contents

1. Introduction

For this dissertation the main focus of prediction will be to analyse the current covid-19 outbreak across various regions. This will be visualised using charts and graphs the goal will be to predict the number of upcoming cases for the 10 days to come. The models will be created in python Using the linear regression and SVM models.

1.1 About the Author

Undergraduate Ciarán Adams is a University student currently studying Computer Science (BSc) at Liverpool Hope University. His Degree includes writing reports, articles, journals all on major technologies and theories of technology in today's world such as Artificial intelligence. He also creates programmes using multiple languages C++, C#, python, java and much more

1.2 Covid-19

The Current Covid-19 Pandemic has halted the world, ruined lives, economies, families and so much more. The devastating impacts of this pandemic will affect an entire generation and for this reason Predicting how they play out is essential for society in future. A pandemic is defined as "an epidemic occurring worldwide, or over a very wide area, crossing international boundaries and usually affecting a large number of people" (JM, 2001). Covid-19 first started in 2019 and is believed to have originated in Wuhan, China. 'In Wuhan, researchers will take a closer look at the Huanan meat and animal market, which many of the earliest people diagnosed with COVID-19 had visited.' (Mallapaty, 2020). The high population of china helped the virus spread at an alarming rate and due to how well the world is connected the virus went global within months. The virus itself is an infectious disease with the majority of those who contract the virus having mild symptoms and may not know they have the virus. This is what makes it deadly as most may be capable of surviving due to individuals unknowing, they have the virus could unknowingly spread the virus to those who may not such as those with asthma, a weakened immune system or even the elderly. The virus spreads similarly to that of the common cold/flu through saliva and droplets from sneezing and coughing. To prevent this pandemic from getting any worse governments have banned travel, asked people to wash hands, wear face masks and even introduce lockdowns. As the effects are now understood on how deadly mishandling such a catastrophic event is known, perhaps a new method to stop pandemics could be introduced and that's where this project comes in. Using Machine language to predict pandemic cases, recoveries and deaths could help understand how deadly a disease could be before a pandemic even begins.

1.3 What is Machine learning?

'Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data.' (Mitchel, 1997) The term machine learning was defined by Arthur Samuel in 1959. In today's world data is everywhere from our phones, to websites, watches from work to our homes and this comes with a problem of too much data to the point where old systems where humans made sense of this data are just no longer viable. This is where machine learning comes in. Machine learning has the ability to not only make this process automated but to keep up or even be ahead of the volume of data, this is the true value of Machine Learning unlike a traditional programme machine learning learns from the data and adapts.

1.4 The Methods of Machine Learning

1.4a Un-supervised Learning

This Method is perfect for a project where a user isn't sure what type of pattern they are looking for as with Un-supervised learning the data is not labelled therefore cannot have any feedback.

However, the ability to find a pattern in unlabelled data, a hidden structure or even what connects together is what it is capable of.

1.4b Reinforcement

The reinforcement method is in its infancy with extreme promise, it has the ability to behave in an environment that will perform an action(s) and will behave depending on the result. This is a way that humans learn if/when this method becomes more capable and used with Artificial Intelligence it will create a huge leap in Machine Learning.

1.4c Supervised Learning

Supervised learning is a method that is used to classify and predict an outcome based on the data that is labelled to the machine. This is the type of machine Learning this project will be primarily based around due to the labelled data and direct feedback being a perfect fit for the dataset that will be used.

1.5 Aims and Objectives

The aim of this project will be to predict future deaths, cases and recoveries of the Covid-19 Pandemic using machine learning. The models that will be used are the SVM Model and the Linear Regression Model. The idea is that using datasets from Kaggle the data will be normalised into three sections as tables and graphs Deaths, Recoveries and Cases. Once the data is normalised the next step will be to create the Models to predict the future data. In this case those models will be SVM and Linear Regression. Upon creation the models will be run using the normalised data that will then deliver the outcome for the predictions which can then be normalised itself into tables and graphs to give a better understanding of them. All of the predicted data will be compared using graphs and then all of the data will be cross examined to determine how effective/accurate the predictions were.

1. Obtain Datasets

2. Normalise Data

3. Create Machine learning Models

4. Run Models

5. Examine the data

2. Literature review

Machine learning is used in a number of aspects on a daily basis in today's world and for this literature review the areas of focus upon which machine learning is used in will be economics, warfare, medicine, and Government. These areas of focus have been chosen as they have always been the backbone of humanity in key periods of history, this can be seen in positive and negative ways from wars, plagues to the greatest accomplishments of our species. These 4 areas are almost always in play with one another.

2.1 Economics

'Economics is concerned with helping individuals and society decide on the optimal allocation of our limited resources.' (Pettinger, 2017) Machine learning has shaped the economy over the last decade due to the rapid advances in Artificial intelligence. This can be seen in an abundance of ways According to an article by (McCarthy, 2019) 'Banks are using it to detect fraud and predict changes in the stock markets, Insurance companies are employing AI to help them produce policy quotes and assess claims, It's helping police forces to identify suspects from grainy CCTV images'. These advancements have a direct impact on production and services. As part of a chain reaction this leads to even more implications such as employment and competition.

The Impact of Machine Learning on Economics by (Athey, 2019) states that 'Despite its importance, there is little economics research on the topic… Without a better understanding of how AI might impact the economy, we cannot design policy to prepare for these changes.'. This statement shows a clear picture that a better understanding of Machine learning must be grasped, this is essential as it has already begun to affect the economy. It's no doubt going to continue to grow as time goes by.

Automation is the key when it comes to the economy, 'Alarmists argument… AI and robotics will spell the end of work by humans, while many economists on the other side claim… eventually increased the demand for labour and wages.' (Athey, 2019). On the one side of this argument it can be seen that automation does have the ability to let employers/companies to cut of employees due to an automation that costs less maintenance than the employees salary 'The organization cites that automation will supplant about 85 million jobs by 2025' (Kelly, 2020) This can be seen to be approaching in a number of posts a great example that can be seen in that tesla creating automated cars this over time could replace taxi work being completely run with no employees in addition to deliveries with trucks as tesla are currently testing automated cargo trucks The electric truck, which is due to be unveiled in September by Elon Musk's electric vehicle company, is close to prototype on-road testing ' (Gibbs, 2017)'. This would have a serious impact on the global economy as a whole as 'The transport industry directly employs around 10 million people and accounts for about 5% of gross domestic product' (Hub, 2019).

However as stated by economists this could lead to a demand for more labour. As there are less drivers and at lower costs people and goods will be able to travel at much lower costs. Leaving more money for consumers to spend which in turn will also create more jobs whether it be consumers buying more goods, services or even travel it seems this could be a positive for the economy in the long run. There are many benefits of machine learning to the economy. 'According to a forecast from the International Data Corporation, $12 billion was spent on AI and machine learning development in 2017, and this will reach $57 billion by 2021.' (Fischer, n.d.) This has created an abundance of jobs in the technology sector, helping employ more computer scientists to create and research new models and methods that can improve upon machine learning more specifically Artificial intelligence. As the needs of machine learning become more and more (which they are) this will transfer into more funding from both the public and private sector. In turn creating even more jobs increases the

economy further as time continues. Machine learning is abundantly used in the gaming industry for Artificial intelligence in both games and the hardware that's used to run them. Nvidia makes use of AI within their GPU's, a company that tailors to the gaming community this is pretty big towards the economy as 'Reports show that the video game industry revenue was $78.61 billion in 2017' (Dobrilova, 2021).

Machine learning has shown to have a major impact on global economics, a trend that will continue for a long time. With automation on the rise the very society upon which we have become so use to may rapidly change. As the growth continues no doubt the research itself will too, overtime new ways to regulate, repair and adapt to the effects will be discovered. The short-term screams of negative and positive impacts within the current economic state of the world however over time.

2.2 Warfare

The implications of Machine learning in warfare are seen throughout the current modern armies of the world. Warfare is changing, tactics and armies alike with the new technologies of machine learning cyberwarfare is taking its place. As artificial intelligence is used more commonly and as the technology becomes cheaper the defensive and offensive capabilities could have serious implications. Machine Learning and Artificial intelligence can be applied in training programmes that can help soldiers become trained to a higher standard/have more experience. This aspect is sure to grow as VR is introduced to train troops even better. 'Microsoft will deliver more than 120,000 devices based on its HoloLens AR headset to the US defence for over ten years ' (Beatrice, 2021) in addition to this it has been used in wargames for decades to help the military detect battle/war outcomes with other nations in a multitude of scenarios from ground invasion, sea invasion and even nuclear war the most famous instance of this is Able Archer 83 'The war game, which was staged over two weeks in November of 1983' (Gallagher, 2020). It became famous for almost starting a third world war due to it being mistaken for a real-life invasion as opposed to a simulation. This is a prime example of how dangerous Machine learning can be used in warfare.

Drones whether its unmanned, manned weaponless or fully loaded have made use of machine learning in many ways the main one being aerial detection for enemy aircraft the advantages of an aircraft having this as opposed to one without is a clear advantage not just for the aircraft but for the entire army/air force as this can reduce the loss of friendly aircraft not only helping win engagement with the enemy but also in reducing the loss of funds as aircraft today can cost as much as '$131million' (Mayfield, 2019) the advantages of this alone cannot be underestimated. Machine learning has been used in warfare for decades, the most known use of the technologies would be the Target recognition used for missiles. The aim of the target recognition is to detect enemy vehicles with the use of laser radar sensors however this is an area that needs improvement for some time with the use of machine learning. This is due to some of the issues with radar as the target's appearance 'changes' depending on the time, weather, and orientation of the sensor. Machine learning aims to improve by using synthetic aperture radar images that allows for the ability to autonomously locate targets and to predict enemy behaviour, environment changes and much more.

Applying machine learning into warfare is extremely dangerous as 'Vulnerabilities exist in machine learning models implemented to generate information from data. An important source of vulnerability lies in the faulty assumptions made while designing and training the ML model.' (Duddu, 2018) This can be extremely dangerous in the area of warfare creating false pretences which can lead to unnecessary catastrophize.

2.3 Medicine

Machine learning always operates better the more data it has and the amount of medical data that's stored is in abundance from patient files, research to case studies and much more. it's use is seen more and more throughout modern medicine in almost every way. Machine learning could potentially begin a new era for modern medicine with the ability to optimise research data, predict illnesses and not only this but help with recoveries, organising patient data more effectively, robotic limbs and much more.

Similar to the topic of this paper machine language being used in prediction and as the predictive methods of machine learning become more accurate the demand for such a technology is extremely desirable for medical diagnostics and researchers. Of course, before this can be widely used to its full potential the models must be tested and reliable. If this can be achieved the benefits will be priceless. One such study was done 'Machine learning in medicine: a practical introduction' (Sidey-Gibbons & Sidey-Gibbons, 2019) the study was based on the accuracy of three machine learning models' ability to predict cancer diagnosis. The algorithms were first trained on data from evaluations before they were used. The results were quite promising as all of the models achieved an accuracy of over 80% with SVM in the lead. 'Maximum accuracy (.96) and area under the curve (.97) was achieved using the SVM algorithm.' That said, according to (Chen & Asch, 2018) 'Machine-learning methods are particularly suited to predictions based on existing data, but precise predictions about the distant future are often fundamentally impossible.' Essentially making it impossible to completely remove the need for a Dr's diagnosis. However, despite this, that doesn't remove machine learning from the equation as it can help spot things a Dr may not, help confirm a diagnosis and reduce times waiting for results. 'Although predictive algorithms cannot eliminate medical uncertainty, they already improve allocation of scarce health care resources' (Chen & Asch, 2018)

The benefits amongst research are not to be overlooked as the use of Machine learning in Project Hanover developed by Microsoft is using ML-based technologies for multiple initiatives including developing AI-based technology for cancer treatment and personalising drug combination for Acute Myeloid Leukaemia Bringing biomedicine and computer science together. These initiatives cannot be underestimated. Cancer is one of the leading deaths in the US according to the CDC 'Cancer: 599,601' (CDC, n.d.) machine learning could catapult cancer research forward exponentially. Should this be a successful treatment, pharmaceutical drugs will boom. Finding new ways to cure a multitude of diseases helping people all over the globe. Radiotherapy is one of the top healthcare practises to make use of machine learning. This is due to the medical image analysis using the technology that helps detect cancers and lesions which could be near impossible with the human eye. Due to machine learning's ability to learn more and more over time this increases its accuracy as it makes use of the data of previous detections becoming more efficient at diagnosing and detecting results Oxford's P1vital® Predicting Response to Depression Treatment (PReDicT) project is using predictive analytics to help diagnose and provide treatment '' (Faggella, 2020).

It's quite clear that the benefits of using machine learning in medicine today are in abundance and in the not so distant future this will only improve. From helping diagnosticians having more efficient methods to discover diseases and viruses to helping researchers solve new anomalies and treatments to most importantly reducing the number of patients deaths.

2.4 Governments

Machine learning is used continuously in the daily life of most people whether they realise it or not from home (Siri, google nexus, Alexus etc) to work (payroll, satnav, weather etc). Since machine learning has first started to be used it has birthed a multitude of technology whether directly or even inspiration. However, machine learning in politics isn't something many people think of and they should. Politics is the backbone of the world shaping nations almost daily whether it be communism, capitalism, fascism, a democracy, or a dictatorship . As technology has progressed the old ways of a politician meeting individuals in person by door to door could soon be coming to an end. As politicians are becoming more reliant on new technology a new approach is needed.

Using machine language voters could become much more engaged and informed of elections, political issues, and their representatives. Machine learning algorithms have the ability to recognise patterns in data and by analysing data this happens every single second particularly used by social media **'**Social media monitoring tools that make use of machine learning are great for analysis' (Perakakis, 2018). If this were used correctly it could re-establish a trust to voters with the information they consume as currently the trust in media is at an all-time low ' politicians are trusted less than any other profession' (Gelblum, 2019). Essentially by analysing the data, machine learning could use targeted advertising for policy's, laws, parties, and elections to be sent to each voter individually.

There are plenty of advantages of using this technology in the public sector which is backed by the government also makes use of this such as the postal service makes use of machine learning to automatically sort letters by reading the handwriting on envelopes helping reduce the costs as there are less staff needed to process the letter along with the benefit of faster processing teams lead to faster delivery of post at a reduced cost. Implementing document processing automation solutions within government agencies allows these entities to better serve their constituents.' (Labs, 2020) .In addition to this the national weather services make use in order to predict the weather, a service that assists agriculture and how citizens will plan their week whether it be to go out and eat or stay at home . Furthermore machine learning is used to help with city planning in some aspects for example analysing traffic flow this can be used to help identify more effective timing being used on traffic lights or to identify heavy road usage to assess the condition of roads that may need maintenance. These services are essential as the more efficient they are the population will be more content with the government that's improving them which will not only help improve the approval rating of said government but also assist in keeping them in power.

The biggest fear among voters particularly in the west is surveillance combined with machine learning and that fear only grows. 'Studies have also demonstrated how mass surveillance crushes free expression and free thought' (Greenwald, 2016). By using video analysis software learns to identify, extract, classify and index objects in video to enable the operator to have the ability to make video data searchable, actionable, and quantifiable. As governments make use of this technology the fear that citizens will have is that they are being stripped of freedom and privacy which could have devastating effects leading a police state or even in the event of a successful invasion as the infrastructure would be in place the invader would have practically full control over the region. These fears are not only a possibility but also a reality, this is only thanks to the whistle-blower Edward Snowden who leaked the program known as PRISM which lead to him having to flee to Russia. 'The unconstitutional surveillance program at issue is called PRISM, under which the NSA, FBI, and CIA gather and search through Americans' international emails, internet calls, and chats without obtaining a warrant' (Toomey, 2018). However, there are plenty of benefits particularly in crime

prevention and counter terrorism spotting culprits who are on the run before they have a chance to strike again. This could prevent disasters that create fear or even take innocent lives.

The rise of deep fakes which using machine learning has the ability to create fake videos with the faces of individuals therefore allowing the ability to ‘fake crimes’ putting innocent citizens in danger in addition to public figures with the possibility of fake videos demonising them. This technology has the ability to destroy lives

These topics all have pros and cons its quite clear that the benefits of machine learning will be pivotal in the coming years for humanity. Laws will be necessary for it to be truly great and to ensure that the fear from citizens won’t be an issue.

3. Methodology and Research Design

The purpose of this section is to introduce the research methodology explaining how machine learning will be used to get the desired outcome, the models we have used and why along with explanation of sections of code.

3.1 Hypothesis

The Hypothesis for this project will be that the linear Regression model will give a bland prediction in a linear fashion without considering any chance of a dip in cases, deaths, or recoveries. All of these outputs will be a slow increase. However, the SVM model. When using these data sets, there are a variety of criteria that can be essential for predicting Covid-19 cases, deaths, and recoveries. Some of these criteria aren't part of the chosen datasets therefore reducing the credibility of said results. However to eliminate bias to allow for a more accurate result the data set has records from countries all across the globe, this helps show the results on a more global scale be predicted much more accurately as opposed to using a data set from a single country.

These models were not chosen as they were the best fit for this role, SVM was chosen as it fits quite well with a prediction such as this however Linear Regression was mainly chosen to compare the differences between these models. Grouping studies together of a similar nature signifies the relevance of the studies and how close they can be compared for a more reliable outcome. When two or more studies that set out to achieve the same hypothesis have a positive correlation in the data it gives confidence in the results. Concurrently, it also allows for flaws to be identified where one study may excel over another whether it be in sample size, method of research or data collection. This was done as the end result Hypothesis is anticipated that the SVM Model will be close to the anticipated figures however not 100% accurate whereas the linear Regression Model will be less accurate than the SVM while being on the correct track.

3.2 The Language Used

Using Anaconda Navigator, the code will be written in python on JupyterLab. Python was chosen as it offers a wide range of tools to simplify the process of creating the Machine learning necessary in this project. An AI based project requires an abundance of research, python can aid in this as with a large community providing plenty of libraries and tools this only makes the research that much simpler. Python is so popular it's been declared the 2nd most popular programming language of all time. (Stephens, 2020)The flexibility and stability of using python can also not be underestimated. Python offers the ability for code that is quite easily readable to a programmer whether it be new or experienced. A code that can be easily read makes it more likely that the developers will create a reliable system. For example, a team that's been creating a code with experienced developers for a very long time to then be dropped onto a team of new inexperienced programmers will be much more reliable to continue the system as the code is easily readable as opposed to one that's not. Also, Python is supported on all major platforms Windows, MacOS and even Linux. Python has always appealed to new and upcoming developers not only as it's easy to read but as it is also easy to learn. This is due to a number of reasons it's a very intuitive language with an abundance of libraries, frameworks and so on. The online repositories for python are in abundance well over 100,000 software packages that are custom created. However, we have chosen python for this project in particular due to certain packages, more precisely Scientific packages. Numpy and Matplotlib will be essential for the machine learning models used in this project. In addition to this python also offers platform independence, this essentially means developers can implement changes to a code of a program on one machine and use it without making any or very little changes on a completely

different machine. This makes training ML models much cheaper and easier as a data scientist can use their own machine (if powerful enough) as opposed to relying on a company/research owned machine.

3.3 The Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import random
import math
import time
from sklearn.model_selection import RandomizedSearchCV, train_test_split
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error, mean_absolute_error

import datetime

import operator

plt.style.use('seaborn') for data visualisation

%matplotlib inline
```

Figure 1, Representation of the code used to implement the libraries

The libraries that are used can be seen above in Figure 1. Numpy will be an essential library to import for this project as this project will heavily rely
on arrays. This library has functions for the ability to work in the domain of linear algebra, Fourier transform and matrices. Pandas is also a no-brainer with the open source data analysis and manipulation tool, it's fast, flexible, and easy to use. matplotlib.pyplot and matplotlib.colors are a collection of functions that make matplotlib work like MATLab. And of course, the core libraries which are used in almost every programming project random, maths, time

3.4 Data Normalisation

The Datasets obtained from Kaggle are in abundance of data however as there is so much it must be normalised. Data normalisation in short is simplifying data, breaking it down into a structure and organising it. ‘Data normalisation gets rid of a number of anomalies that can make analysis of the data more complicated. ’ (import.io, 2019) Normalisation is the process of reorganising data in a database so that it meets two basic requirements:

1. There is no redundancy of data, all data is stored in only one place.
2. Data dependencies are logical, all related data items are stored together. (techopedia, 2021)

Normalisation is quite necessary for Machine Learning as it can reduce disk space and processing times, although not entirely essential in the case of this project it does offer one necessity which will be made use of and that’s the organisation of the data. With the data being much more structured it’ll make the code much easier to write and read for users.

In order to do this for this project the first step will be to load the datasets this will be done by using this line of code:

confirmed_cases = pd.read_csv(‘FILE LOCATION’)

the names of each dataset will be different (deaths reported, cases_revovered). Once this is completed the .keys function will be used to extract all of the columns. keys() function returns the 'info axis' for the pandas object, in this instance as the pandas object is a data frame the output will return as columns

Cols = confirmed_cases.keys()

Next, using the .loc function from Pandas provides a unique method to retrieve rows from a Data frame. The date columns that have information of the confirmed cases, deaths and recoveries will be extracted.

Confirmed = deaths_cases.loc[:, cols[4]:cols[-1]]

The first parameter that's colon means beneath all the rows while the 2nd parameter means that the data is needed from the 4th column to the last column which are the date values. This process will be repeated for deaths and recovered cases. Next the data will be organised into 4 lists this will be done by obtaining the total confirmed deaths and recovered cases then appending them into 4 empty lists in addition to this calculating the mortality rate which is death_sum/confirmed cases will also be achieved. The 4 lists:

world_cases = []
total_deaths = []
mortality_rate = []
total_recovered = []

To obtain the data for the empty lists, a for loop has been created which can be seen in figure 2 below. It takes the date variables into consideration and for each value summing the total confirmed, deaths and recovered cases then appends the values to the empty lists that have been created.

```
for i in dates:
    confirmed_sum = confirmed[i].sum()
    death_sum = deaths[i].sum()
    recovered_sum = recoveries[i].sum()
    world_cases.append(confirmed_sum)
    total_deaths.append(death_sum)
    mortality_rate.append(death_sum/confirmed_sum)
    total_recovered.append(recovered_sum)
```

With the core of the data normalised the data is now prepared for the prediction with the machine learning models.

3.5 Independent and Dependent Variables

When making the predictions the independent and dependent variables must be set and just as importantly understood. An independent variable is a variable with a value that won't change by the effect of other variables. However, it is used to manipulate the dependent variable. In the case of this project X will be the days since 22/01/2020 (independent). A dependent Variable value changes

when an independent value manipulates it. Y will be the Number of Cases or number of deaths or number of recoveries (dependent). This will depend on the prediction that is being attempted to simplify the explanation the table 1 will be used to demonstrate:

Prediction	Independent (X)	Dependent(Y)
Deaths	days since 22/01/2020	Total_deaths
Cases		Total_Cases
Recoveries		Total_Recoveries

Table 1, clarification of independent variable

Essentially X won't change however Y will change depending on X

3.6 The Machine Learning Models

Two different models will be used for this dissertation: the SVM Model and Linear Regression Model.

Both SVM and Linear Regression Models are Machine learning Algorithms based on supervised learning. Both of these models operate in a different manner and in this instance will be used to achieve the same goal, the predictions of the desired outcomes. All of the predicted data outcomes will be compared over one another to discover which of the Models are more suited for this type of prediction.

3.6a SVM Model

An SVM model otherwise known as A support vector machine is a supervised learning model. It uses classification algorithms 'Support vector machine is highly preferred by many as it produces significant accuracy with less computation power.' (Gandhi, 2018). They're able to categorise new text once the model has obtained 2 sets of labelled training data. SVM is a fast-dependable classification algorithm that performs well with limited amounts of data. This is another core reason upon which the data had to be normalised at the very beginning of this project. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes if possible. The algorithm will look for the ideal lean between the data/classes. These points are named the support vectors (thus the name). The distance of the line and support vectors will then be computed, the aim is to maximise margin. The hyperplane for which the margin is maximum is the optimal hyperplane.

This is the model which has been created for this project in particular as it will be predicting the deaths, cases, and recoveries each time the SVM will be re-run with the necessary changes for the particular value which can be seen in Figure 3.

```
kernel = ['poly', 'sigmoid', 'rbf']
c = [0.01, 0.1, 1, 10]
gamma = [0.01, 0.1, 1]
epsilon = [0.01, 0.1, 1]
shrinking = [True, False]
svm_grid = {'kernel': kernel, 'C': c, 'gamma': gamma, 'epsilon': epsilon, 'shrinking' : shrinking}


svm = SVR()
```

```
svm_search = RandomizedSearchCV(svm, svm_grid, scoring='neg_mean_squared_error', cv=3, return_train_score=True,
n_jobs=-1, n_iter=40, verbose=1)
svm_search.fit(X_train_confirmed, y_train_confirmed)
```

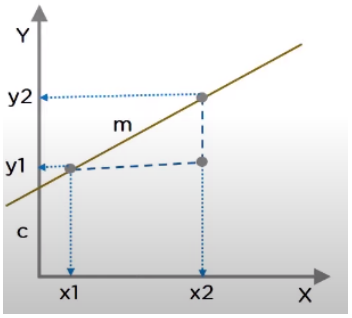
These parameters are kernel, c, gamma, epsilon shrinking and svm_grid, Kernel will transform the input data to the necessary form, it specifies the kernel type to be used in the algorithm its must be one of linear, poly rbf they create the separation line in the higher dimension. If nothing is given rbf will be used. c is a declaration parameter, gamma, and epsilon. The SVM module is the imported and create support vector classifier object by passing argument kernel as the linear kernel in SVC() function, the number of jobs to use for the computation. This will only provide speedup for n_targets

> 1 and sufficient large problems. SVM_search will searches for the maximum marginal hyperplane then the model is fit on train set using svm_search.fit().

3.6b Linear Regression Model

'Linear regression is perhaps one of the most well-known and well understood algorithms in statistics and machine learning' (Brownlee, 2020). Linear regression is used to predict the relationship between the dependent and independent variables, a statistical model. This is achieved by examining 2 factors, the first important factor is to determine the important variables for predicting the outcome variable. The second is to make sure the regression line is as reliable as possible in order to make the predictions as accurate as possible. The simplest form of linear regression equation with one independent and dependent variable is as shown in Figure 4 which is then explained using table 2.

m	Slope of the line
x	Independent Variable
c	Coefficient of the line
y	Dependant Variable



Y=m*x+c $m = \frac{y2-y1}{x2-x1}$

Table 2, explaining Figure 4

Figure 2.6b2

```
from sklearn.linear_model import LinearRegression
linear_model = LinearRegression(normalize=True, fit_intercept=True)
linear_model.fit(X_train_confirmed, y_train_confirmed)
test_linear_pred = linear_model.predict(X_test_confirmed)
linear_pred = linear_model.predict(prediction_forecast)
```

Figure 5 is the code that was used for the linear model of this project. It's a simple form of code and similar to the SVM Model certain variables will change depending on the variables necessary.

3.7 Data Output

Once the data has been output, the data will all be put into graphs that show the difference between the original confirmed data and the new predicted data. This will show basic trends and accuracy to the human eye in addition to this with the required data now obtained the ability to cross examine, work out the accuracy and to compare with the actual results of the next ten days will be able to begin

4. Results and Discussion

4.1 Normalised Data

Upon normalisation of the data, the data has been simplified even more to give a better understanding of all of the information around the current pandemic into pie charts, bar graphs and linear graphs. As all of the datasets have been normalised into the variables, breaking them down even more helps to show a better picture to the user.

As mentioned in section 3.4 Data Normalisation the datasets had to be loaded so that the system could obtain the data that was necessary. To clarify this step was completed the head of the data set was displayed this can be seen in Figure 6. This step was of crucial importance not only for the normalisation of data but for the core data to be read by the system to allow for the predictions to be backed by a large amount of accurate data.

To further normalise this data and give a better understanding of how the current pandemic is unfolding a bar graph showing the number of cases in china compared to the countries outside of china in Figure 7. This was due to China being ground zero for this outbreak to help give a better understanding of how hard hit a country that is ground zero for a pandemic may be. This can also be seen in figure 9 helping visualisation of what countries other than China have been hit the hardest as of early 2020. Figure 8 helps see exactly all of the nations hit the hardest globally in a simple yet narrative bar graph. The drastic amounts of cases that can be seen in China shows how devastating being ground zero for the pandemic has hit the nation, not to mention one of the highest populations in the world. The Pie Chart in Figure 10 offers a similar visualisation giving a better understanding that China has been hit hard with a simple and clear visual showing the nation having almost half of the cases worldwide.

```
latest_confirmed
```

```
0      114
1      839
2      226
3        1
4      428
      ...
445      1
446      1
447      1
448      1
449      1
Name: 3/15/20, Length: 450, dtype: int64
```

```
latest_deaths
```

```
0       1
1      22
2       0
3       0
4       0
       ..
445     0
446     0
447     0
448     0
449     0
Name: 3/15/20, Length: 450, dtype: int64
```

```
latest_recoveries
```

```
0       35
1      118
2      105
3        1
4       42
      ...
445      0
446      0
447      0
448      0
449      0
Name: 3/15/20, Length: 450, dtype: int64
```

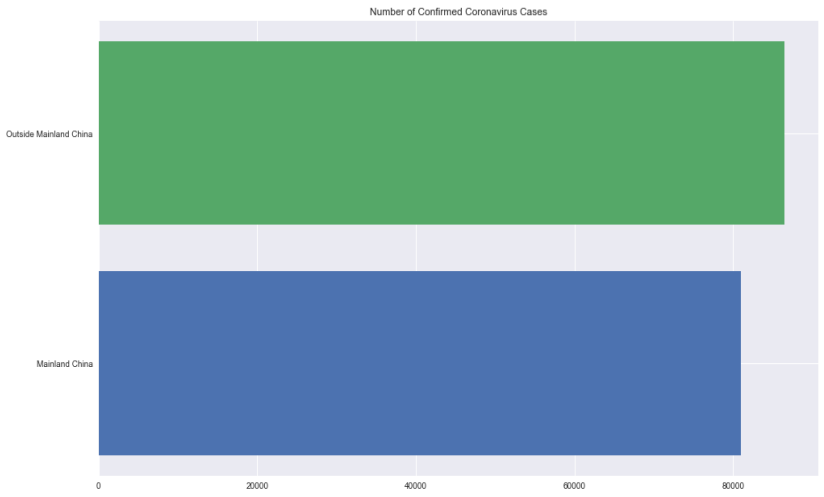


*Figure SEQ Figure * ARABIC 7, Barr graph depicting case numbers globally against cases in china*

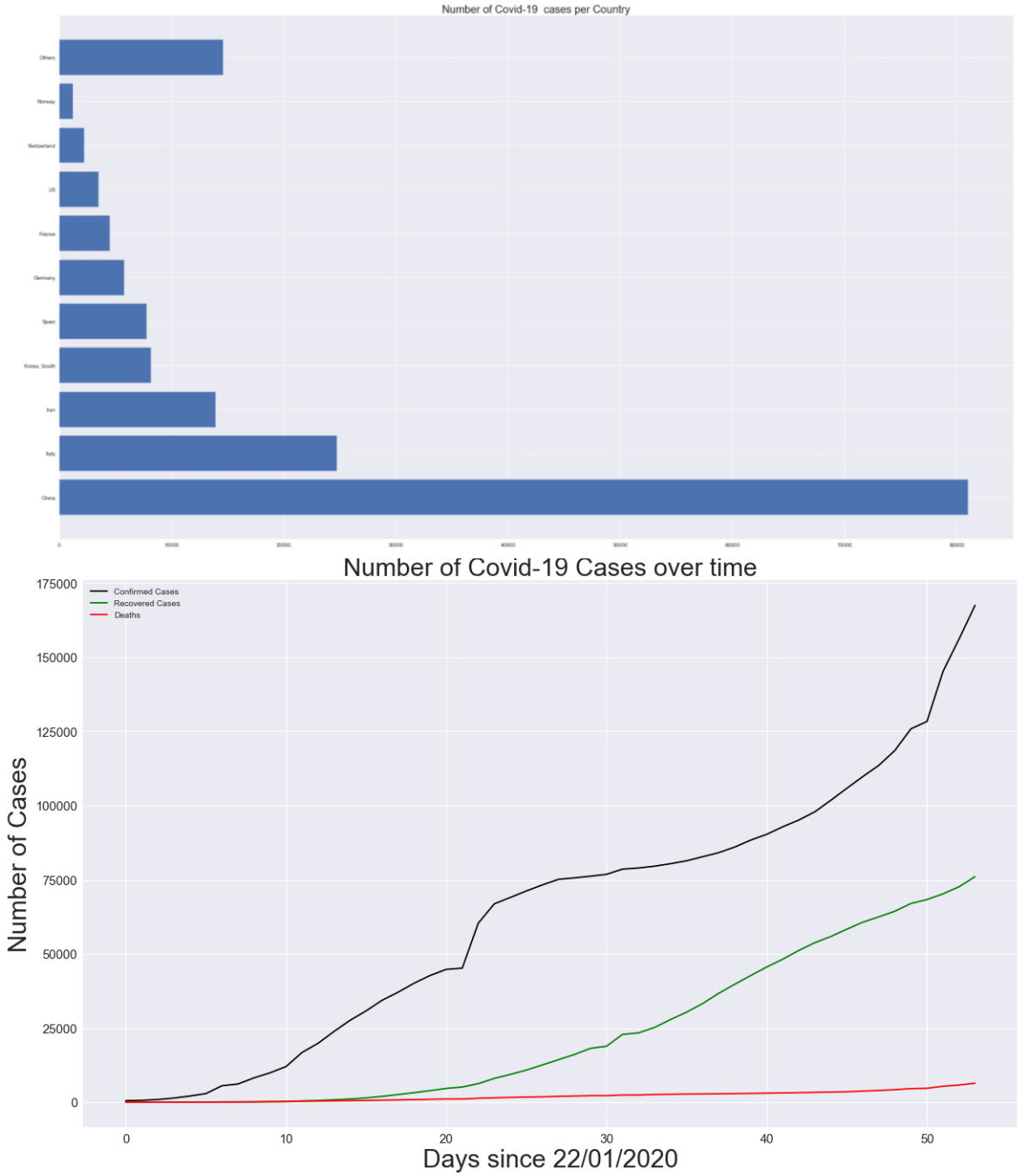


Figure 11, Line graph that depicts cases, deaths, and recoveries of Covid-19

This line graph in Figure 11 shows us the rate at which all of the variables are increasing over the course of 50 days. Each colour co-ordinated to simplify it further creating a better way of informing individuals on how the data can be seen with a clear idea of how the cases may be rising at a high rate the recoveries are too with the deaths increasing slowly. This chart shows that there is a clear upwards linear trend that all of the factors are rising with deaths at a slow and minor upwards pace. The confirmed cases show that the cases have started out slow and have begun to rapidly increase more and more, around the 20th day there is a spike which boosts this trend further increasing the cases. As for case recoveries the regression starts off fairly slow just as the deaths are, at first worrying results however a small rise over deaths can be seen around day 15. From this point the

cases recovered have begun to soar all the way to 75,000 completely outpacing the deaths which are still under 10,000.

4.2 Confirmed Cases Prediction Results

Next 10 Day Predictions for Covid-19 Cases using the Linear Regression Model	
Prediction Dates	Prediction
03/16/2020	132336.25252525
03/17/2020	134890.72222222
03/18/2020	137445.19191919
03/19/2020	139999.66161616
03/20/2020	142554.13131313
03/21/2020	145108.6010101
03/22/2020	147663.07070707
03/23/2020	150217.54040404
03/24/2020	152772.01010101
03/25/2020	155326.47979798

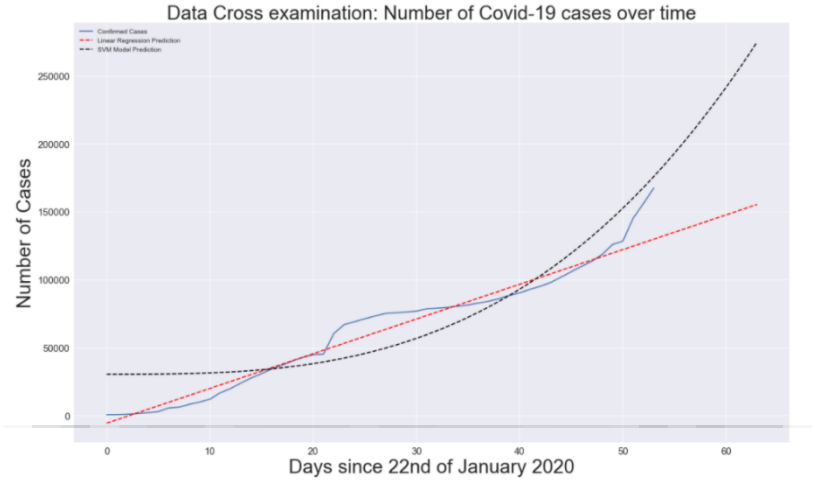
Table 3, SVM Prediction of Covid-19 Cases Results

Table 4, LRM Prediction of Covid-19 Cases Results

The data that was predicted by The Machine learning Models on Cases

Next 10 Day Predictions for Covid-19 Cases using the SVM Model	
Prediction Dates	Prediction
03/16/2020	184302.332923890
03/17/2020	193012.43524916755
03/18/2020	202045.0977548219
03/19/2020	211406.18517140602
03/20/2020	221101.56222947178
03/21/2020	231137.09365957195
03/22/2020	241518.64419225854
03/23/2020	252252.0785580839
03/24/2020	263343.2614876004
03/25/2020	274798.0577113601

As the tables of cases predicted of the next ten days show there has been quite a large difference between the two models that have been used to predict. The SVM model appears to be growing at a much greater rate which can be seen in table 3more rapid rate showing a much more dire and dangerous outlook of the future cases. The linear regression model, while it also shows a growth in the number of cases, appears to be much lower than the SVM model, growing at a slower pace and almost half the numbers. For comparison the 1st day of prediction the SVM model has predicted 184,302 cases whereas the Linear regression model has predicted 132,336 cases. This is a difference of 51,966 cases on the first day of prediction. This similarity can be seen throughout the table with the average increase in the linear regression model as 2,500 new cases per day as opposed to the average increase of the SVM model which is averaging at 10,000 new cases per day. This shows a difference of 400% between both models on daily increase.



As shown in the figure 12 we can see that the modes have all overlapped quite well at some point during the prediction. The SVM model appears to have lined up perfectly at the end whereas the linear regression model appears to have slowed down substantially. This is due to a much more linear approach as we can see the model was quite accurate until around the 10,000 cases mark here is where its downfall lies as the model didn't detect a spike in cases around this time bringing down its overall performance to predict the next 10 days. However

the SVM model appears to have been relatively accurate throughout and this can be seen quite well as of approximately 16,000 cases where the prediction appears to line up with the confirmed cases in a linear fashion almost as if it took the spike in new cases into consideration.

4.3 Confirmed Deaths Prediction Results

The data that was predicted by The Machine learning Models on Deaths

Table 5, SVM Prediction of Covid-19 Deaths Results

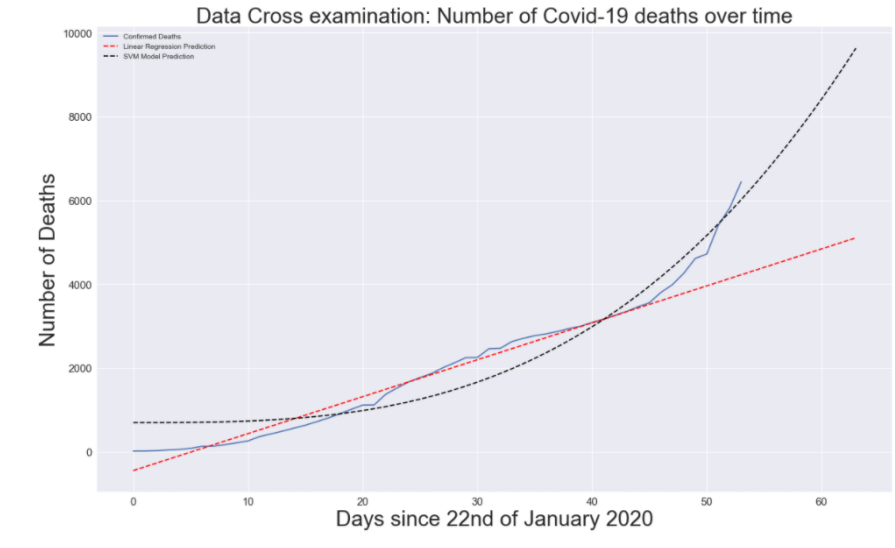
Table 6, LRM Prediction of Covid-19 Deaths Results

Next 10 Day Predictions for Covid-19 Deaths using the SVM Model	
Prediction Dates	Prediction
03/16/2020	6323.5029478476845
03/17/2020	6641.868975937685
03/18/2020	6972.02501516781
03/19/2020	7314.185429376984
03/20/2020	7668.564582404109
03/21/2020	8035.376838088095
03/22/2020	8414.836560267859
03/23/2020	8807.158112782316
03/24/2020	9212.555859470365
03/25/2020	9631.244164170927

Next 10 Day Predictions for Covid-19 Deaths using the Linear Regression Model	
Prediction Dates	Prediction
03/16/2020	4311.92709706
03/17/2020	4400.08454106
03/18/2020	4488.24198507
03/19/2020	4576.39942907
03/20/2020	4664.55687308
03/21/2020	4752.71431708

03/22/2020	4840.87176109
03/23/2020	4929.02920509
03/24/2020	5017.1866491
03/25/2020	5105.3440931

As tables 5 and 6 show there is a drastic change than that of the tables 3 and 4 as the previous figures saw a drastic change of 400% between the end results in this instance the change is different by 4,526 which approximately With the SVM predicting 9831 deaths by march 25th 2020 opposing the 5105 of the linear regression model. That said a similar trend can be seen as of the SVM model growing at a faster pace almost doubling the end result of the linear regression model.



As shown in the figure 13 the SVM Model seems to be lined up quite well with the official stats particularly at the end appearing to show an upwards trend with the official results. The Linear Regression Model shows the most accuracy around the mid-range of the results, almost pinpointing the exact trend of the official results; however the unexpected spike has left its results relatively inaccurate as seen at approximately day 45.

4.4 Confirmed Case Recoveries Prediction Results

The data that was predicted by The Machine learning Models on Recovered Cases

Next 10 Day Predictions for Covid-19 recovered cases using the SVM Model	
Prediction Dates	Prediction
03/16/2020	112389.18425440212
03/17/2020	118766.38545076159
03/18/2020	125379.75277829339
03/19/2020	132233.58016665262
03/20/2020	139332.16154548322
03/21/2020	146679.79084446264
03/22/2020	154280.76199321804
03/23/2020	162139.36892144737
03/24/2020	170259.90555874613
03/25/2020	178646.66583487001

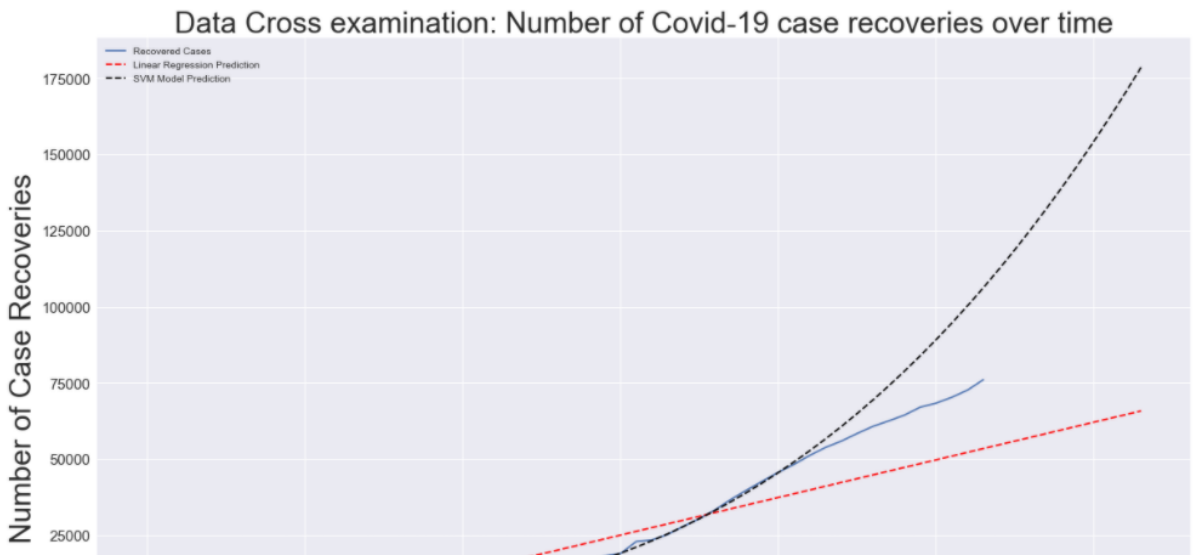
Table 7, SVM Prediction of Covid-19 Recoveries Results

Table 8, LRM Prediction of Covid-19 Recoveries Results

Next 10 Day Predictions for Covid-19 recovered cases using the Linear Regression Model	
Prediction Dates	Prediction
03/16/2020	54648.13236715
03/17/2020	55884.34830918
03/18/2020	57120.56425121
03/19/2020	58356.78019324
03/20/2020	59592.99613527

03/21/2020	60829.21207729
03/22/2020	62065.42801932
03/23/2020	63301.64396135
03/24/2020	64537.85990338
03/25/2020	65774.07584541

Tables 7 and 8 show a drastically different story to one another with a quick glance it's completely understandable one of these models are completely inaccurate. The SVM has an end result of 178,646 recoveries as of March 25th, 2020 (Table 7) as opposed to the Linear Regression Model at a mere 65774 recoveries (Table 8) , a difference of 112,872 recoveries total.



As shown in the figure 14 the SVM Model once again seems to be rather accurate in the form of a pattern of trend going at a particularly high upwards trend. The Linear Regression Model shows tin this instance exactly what was expected: a linear result with a drift between the official results as the model is under achieving in accordance.

4.5 Discussion

Essentially the programme created has the ability to predict the deaths, cases, and recoveries of individuals across the globe for the purpose of predicting these variables in the pandemic of Covid-19. This program is compatible for any pandemic for it to be used in other projects the datasets would have to be compatible or a recreation of the code will be necessary for a drastically different type of dataset. When the programme is run the dataset is then converted into three variables which are deaths, recoveries, and cases. Using this data, the machine learning models that have been created SVM and Linear Regression will use this data to attempt to predict the next fortnight of the variables. This is essential for predicting how devastating pandemics may be which then alerts governments, Dr's etc how drastic of an action they should take. The effects of a pandemic are understood particularly in these times with devastating effects to the economy, healthcare and most importantly lives.

4.6 Comparing the Data

The aim of this section is to compare the models to determine and average accuracy throughout all of the predictions that have been made for this project. Once this has been established the models themselves separately will be compared to determine which model was more accurate throughout this experiment.

4.6a Case Prediction Accuracy

To determine the accuracy of the models we must compare the predicted results with the official results of the date of 25/03/2020 which is '438,746 confirmed cases globally' (Jennings, 2020). This data shows that the SVM model predicted 163,951 cases less than what had actually occurred that day. This leaves the model at an in-accuracy of 37.36% with an accuracy of 62.74%. Furthermore, the Linear Regression Model appears to be even more inaccurate predicting 283,420 less than the official stats having a drastic in-accuracy of 64.59% leaving the model at a measly accuracy of 35.41%.

4.6b Death Prediction Accuracy

With the confirmed deaths as of march 25th 2020 being at 19 675 deaths (Jennings, 2020) both Models appear to be less accurate than that of the case prediction. As the SVM predicted 9831 deaths this leaves it at an accuracy of 49.96%, the Linear Regression Model once again at a lower accuracy of 25.94% with the low number of deaths at 5105. These results are quite low with the SVM being somewhat reliable and the Linear Regression Model continuing to agree with the hypothesis that the accuracy is lower than that of the SVM, that said this result is an inaccuracy of almost 75% rendering it completely inept.

4.6c Recoveries Prediction Accuracy

The prediction for Recoveries had an anomaly when Looking at the SVM end result with a figure of 178,646 recoveries as opposed to previous results that have been inaccurate by underperforming results on this occasion the model was too optimistic. As the official Results came in at 111 895 recoveries (Jennings, 2020), the SVM predicted an additional 66,571 this is an additional 59.49% leaving it at an accuracy of 40.51%. The most surprising outcome however is the Linear Regression Model at a total of 65774 recoveries coming in at an accuracy of 58.78% the most accurate of the model has been thus far and for the first time more accurate than the SVM.

4.6d Overall Accuracy

As seen in table 9 The results of this dissertation were different to what was expected within the hypothesis, the cases were exactly as expected with the SVM model being almost twice as accurate as the linear regression model with an accuracy of 63%. The prediction for deaths was extremely disappointing with both models achieving less than 30% accuracy which damages the results overall by quite a lot. That said, recoveries are where the model has the most shocking results as the linear regression model is over 30% more accurate than the SVM model at a whopping 59% accuracy.

Due to the limiting factors that have been used in this model the results were rather impressive. The SVM had an overall accuracy of 43.06% as opposed to the Linear Regression Model had an overall accuracy of 40.05%. These results were better than expected considering the time frame and single developer. This in a way proves that predictive analytics have a serious ability to function in a way to predict how deadly a pandemic may be. Overall agreeing with the aim of this project in proving that machine learning should be an essential tool utilised within the prediction of pandemics.

Table 9, Here this tables breaks down the accuracy of the models prediction per variable

Predictions	Models Accuracy	
	SVM	Linear Regression
Cases	62.74%.	35.41%
Deaths	25.94%	25.94%
Recoveries	40.51%	58.78%

Despite the accuracy of these models being so low the project's goal has been a success, as the models and system have all been created by a single developer this was to be expected. Considering the accuracy was above 40% for both Models this is somewhat impressive and shares an insight into what a highly funded dull development team could achieve answering the question of this experiment it appears that machine learning does have a strong ability to be used as predictive analytics for pandemics, this research will further be backed as time goes on and the technology advances, more factors are taking into account and so on.

5. Conclusion

5.1 What was Achieved

After a long time and a lot of work this project has finally come to it's end achieving everything it has aimed to do. A Machine Learning program that predicted pandemic cases, deaths and recoveries is what has been created and from this the analysis has been able to be completed. Challenging and time consuming are just some of the words to describe this project however the word best to describe the project is proud.

The programme itself in the end was able to read all of the datasets with ease and return those values into normalised data. This allowed the creation of charts, graphs and tables achieving the goal of bringing an understanding of the pandemic figures globally and regionally. With this data both of the Machine Learning Models were brought to life with an ability to predict the future outcomes of the desired criteria. Although not entirely 100% accurate this was never the goal which was aimed to achieve, the real achievement was to discover just how accurate these models could be, something that could not be achieved without the programme itself and the analysis which followed after.

The results yielded by this project have returned an outcome surpassing the expectations of the hypothesis. With the accuracy itself being relatively low the results are more than high enough to pursue Machine learning prediction analytics for future pandemic predictions. With even more dedicated developers and funds it's almost certain that machine learning will become an essential part of predicting pandemics in the future.

5.2 Future of This Project

This project has a lot of promise that can be grown upon for a multitude of reasons. The next step for this project is to add additional Machine Learning Models to see if they can perform with any more accuracy. Models such as Logistic Regression, Decision Tree, Naive Bayes, kNN and K-Means just to name a few.

Adding More data will of course increase the accuracy of the Models, one such idea could be to add a 'Lockdown' field which shows a decrease in cases and deaths this would show the effectiveness of lockdowns in the case of future Pandemics. There is an abundance of extra data types to add to help predictions for pandemics, economic decrease, additional data for current fields, government approval and so on.

Using this model for future pandemics is also an option, with the appropriate data necessary both for the pandemic in question and too best fit the program which has been created.

All of these additional fields which could be added do have an effect on the accuracy of the models in predicting this can be a case for pretty much anything which is why it's practically impossible for Machine Learning to be 100% accurate in predicting Pandemic Outcomes entirely.

6. Bibliography

1. Athey, S. (2019). The Economics of Artificial Intelligence. *The Impact of Machine Learning on Economics*, 507.

2. Beatrice, A. (2021, April 6). *HOW COUNTRIES USE AR AND VR IN DEFENCE TO ATTAIN WARFARE SUPREMACY?* Retrieved from Analytics Insight: https://www.analyticsinsight.net/how-countries-use-ar-and-vr-in-defense-to-attain-warfare-supremacy/

3. Brownlee, J. (2020, August 15th). *Linear Regression for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/linear-regression-for-machine-learning/

4. CDC. (n.d.). *National Center for Health Statistics*. Retrieved from Center for disease control and prevention: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

5. Chen, J. H., & Asch, S. M. (2018). Machine Learning and Prediction in Medicine. *Beyond the Peak of Inflated Expectations*.

6. Dobrilova, T. (2021, March 19). *How Much Is the Gaming Industry Worth in 2020*. Retrieved from techjury: https://techjury.net/blog/gaming-industry-worth/#gref

7. Duddu, V. (2018). A Survey of Adversarial Machine Learning in Cyber Warfare. *Defence Science Journal*.

8. Faggella, D. (2020, March 4). *7 Applications of Machine Learning in Pharma and Medicine*. Retrieved from Emerj: https://emerj.com/ai-sector-overviews/machine-learning-in-pharma-medicine/

9. Fischer, M. (n.d.). *MACHINE LEARNING WILL DRIVE ECONOMIC GROWTH IN EVERY INDUSTRY*. Retrieved from Guidepoint: https://www.guidepoint.com/machine-learning-will-drive-economic-growth-in-every-industry/

10. Gallagher, S. (2020, 11 29). *WarGames for real: How one 1983 exercise nearly triggered WWIII*. Retrieved from arsTechnica: https://arstechnica.com/information-technology/2020/11/wargames-for-real-how-one-1983-exercise-nearly-triggered-wwiii/

11. Gandhi, R. (2018, June 7). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved from towards data science: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

12. Gelblum, B. (2019, November 27). *Trust in politicians has fallen to an all time low in the UK*. Retrieved from The London Economic: https://www.thelondoneconomic.com/news/trust-in-politicians-has-fallen-to-an-all-time-low-in-the-uk-168748/

13. Gibbs, S. (2017, August 10). *Tesla seeking to test driver-free electric trucks on public roads*. Retrieved from The Guardian: https://www.theguardian.com/technology/2017/aug/10/tesla-test-autonomous-electric-trucks-without-drivers-public-roads-nevada-elon-musk

14. Greenwald, G. (2016, April 28). *New Study Shows Mass Surveillance Breeds Meekness, Fear and Self-Censorship*. Retrieved from The Intercept: https://theintercept.com/2016/04/28/new-study-shows-mass-surveillance-breeds-meekness-fear-and-self-censorship/

15. Hub, E. S. (2019, May 23). *Transport sector economic analysis*. Retrieved from EU Science Hub: https://ec.europa.eu/jrc/en/research-topic/transport-sector-economic-analysis#:~:text=Transport%20plays%20an%20important%20role,gross%20domestic%20product%20(GDP).

16. import.io. (2019, may 7). *What is Data Normalisation and Why Is It Important?* Retrieved from import.io: https://www.import.io/post/what-is-data-normalization-and-why-is-it-important/#:~:text=The%20Importance%20of%20Data%20Normalization&text=Data%20normalization%20gets%20rid%20of,information%2C%20or%20updating%20existing%20information.

17. Jennings, S. (2020, March 25). *COVID-19 Update: Global Confirmed Cases as of March 25, 2020*. Retrieved from PatientCare: https://www.patientcareonline.com/view/covid-19-update-global-confirmed-cases-march-25-2020

18. JM, L. (2001). A dictionary of epidemiology . *A Dictionary Of Epidemiology*.

19. Kelly, J. (2020, October 27). *us-lost-over-60-million-jobs-now-robots-tech-and-artificial-intelligence-will-take-millions-more*. Retrieved from Forbes: https://www.forbes.com/sites/jackkelly/2020/10/27/us-lost-over-60-million-jobs-now-robots-tech-and-artificial-intelligence-will-take-millions-more/?sh=17d3a071a525

20. Labs, H. (2020, January 15). *how-document-processing-can-streamline-government-efficiency*. Retrieved from Hyper Labs: https://hyperscience.com/library/e-books/how-document-processing-can-streamline-government-efficiency-pm/?utm_source=google&utm_medium=cpc&utm_campaign=emea%20-%20ga%20-%20government&utm_adgroup=emea%20-%20ga%20-%20government%20-%20tofu%20-%20ai&utm_term=

21. Mallapaty, S. (2020, November 11). *Where did COVID come from? WHO investigation begins but faces challenges*. Retrieved from nature: https://www.nature.com/articles/d41586-020-03165-9

22. Mayfield, M. (2019, April 26). *Fighting Over Fighter Jets: Pentagon Plan to Buy F-15EX Sparks Controversy*. Retrieved from National Defense: https://www.defenseone.com/ideas/2020/08/nine-reasons-congress-should-nix-air-forces-f-15ex-purchase/167603/

23. McCarthy, A. (2019, August 13th). *Why artificial intelligence is shaping our world*. Retrieved from Machine MInds: https://www.bbc.com/future/article/20190812-how-ai-powered-predictive-text-affects-your-brain

24. Mitchel, T. (1997). *Machine Learning.* New York City: Hill McGraw.

25. Perakakis, M. (2018, june 12). *4 ways machine learning will affect social media monitoring*. Retrieved from Mentionlytics: https://www.mentionlytics.com/blog/4-ways-machine-learning-will-affect-social-media-monitoring/

26. Pettinger, T. (2017, June 27). *The Importance of Economics*. Retrieved from economics help: https://econ.economicshelp.org/2008/06/importance-of-economics.html

27. Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology* .

28. Stephens, R. (2020, July 27th). *RedMonk Top 20 Languages Over Time: June 2020*. Retrieved from RedMonk: https://redmonk.com/rstephens/2020/07/27/redmonk-top-20-languages-over-time-june-2020/

29. techopedia. (2021, 02 10). *normalisation*. Retrieved from techopedia: https://www.techopedia.com/definition/1221/normalization

30. Toomey, P. (2018, August 22). *The NSA Continues to Violate Americans' Internet Privacy Rights*. Retrieved from aclu: https://www.aclu.org/blog/national-security/privacy-and-surveillance/nsa-continues-violate-americans-internet-privacy

7. Figures and Tables

7.1 Figures

7.2 Tables