**SUICIDE ANALYSIS AND PREDICTION**

**1.0 Introduction**

Thisiteration is a deep dive into the suicide dataset for knowing the reasons behind suicides each year happening around the world, though several studies like John et al (2007) had been done in the past, this study is going to make new findings that are useful for government entities to better understand the problems that are hidden under. This analysis is also going to help them for making new policies that could reduce the mortality rate in the future. I will be looking at different aspects of suicides and predicting how much more suicides are going to happen in the coming years in different countries.

The purpose of this study is to understand the reasons behind suicide. The suicide rate in many countries is higher than the total mortality rate. To make a change to these problems, we need to study different patterns and clusters in the data and understand what factors are triggering the tendency for someone to think of committing suicide. Furthermore, a web-based system will be created that can dynamically produce insightful visualizations on a python dashboard. I believe this dashboard can make a significant impact on society and government vice-versa to help each other to mitigate the suicide figures each year in different countries.

**2.0 Background Problem**

Different countries have different social, economical, and cultural backgrounds. For example, I can tell you the names of the two ‘recent most heard names’ Russia and Ukraine. We know there is a huge conflict going on between the two countries. It is always suspectable seeing that kind of observation and insights from data. As per the Explanatory data analysis (EDA), I could see those two countries have high suicide rates among many other countries.

CSO statistical release committedseveral circumstancestohundredsall of thes which we don’t know yet. I mean it’s very hard to collect all those information together, Some of such common factors are available in the form of features in the suicide dataset to look deep into it and analyse to get more insights from the data of different countries. I think there is a need for government entities to step in to identify those threats to human life and try to predict future figures and mitigate these deaths by taking measures in advance. My research is going to look at different countries' data separately to precisely understand the most affected features in the dataset.

**3.0**

are normally, the study of Kumar and Susan (2020) is a great example of the effective use of these approachess The main goal of this study was to identify the future infected cases and virus spread rate for the preparation in the healthcare services to avoid deaths. In this study, they have used day level information on covid-19 spread for cumulative cases from the whole world. The top ten most affected counties were the US, Spain, Italy, France, Germany, Russia, Iran, United Kingdom, Turkey, and India. They have used temporal data of coronavirus spread from January 22, 2020 to May 20, 2020. They have used ARIMA and Prophet models are effectively used for forecasting future infected cases and evaluation of the model is done using mean absolute error, root mean square error, root relative squared error, and mean absolute percentage error. This study has proved that ARIMA Model was more effective in forecasting covid-19 prevalence. been ,rthat .D The trend analysis showed a rapid increase in the affected cases and the prediction study showed a great increase in the expected active, recovered, and death cases worldwide. However, as per their research, containment policies and lockdowns might affect the prediction results.



**Fig. 1.0.0:** Framework to evaluate the forecasting models.(Kumar and Susan 2020b)

Another study I can point out was done on predicting mortality of predicting attributable to cancer in 2021 in Qingdao, China. They have also used ARIMA Model for the prediction of deaths. ARIMA model is a combination of the autoregressive model and moving average model. Another study was done by Airiti and her team on the prediction of the exchange rate. Artificial Neural Network and ARIMA are used for this study.

The study of Singh et al. Singh et al. (2020) for predicting the covid-19 pandemic on time series data is also a similar example. But they have used the Support Vector Machine (SVM) model for the prediction. The objective of the research was to produce a real-time forecast using SVM Model. The purpose of this study was to investigate the coronavirus disease in the year 2019 prediction of confirmed, diseased and recovered cases. This prediction is used to plan resources, determine government policy, provide survivors with immunity passports and use the same plasma for care. The findings from the author indicate that Covid-19’s daily mortality rate is positively correlated with several confirmed cases. The study also showed that it is dependent on the dietary routine and immune system. As per the author, an emergency can be awakened before the proper vaccine is invented. from Włodarczyk (2021) has used the same SVM modelthis has alsothe along with SVM

in Huang et al(2017)sin thes Cortes and Vapnik(1995)showed

e(Vector Autoregressive Models for Multivariate Time Series 2006)VAR Model is mainly used when we have to deal with Multivariate time series data. The suicide dataset contains more than twenty variables, which needs such a complex model like VAR. VAR is a systematic but flexible approach for dealing with complex real-world behaviour. VAR is also popular amongst data scientists because of its high forecasting performance. Most importantly VAR can capture the intertwined dynamics of time series data. We have to understand the lag using calculating AIC and BIC. AIC is considered in our case and looking at different lags from one to nine, we need to look where the AIC value is dropping quickly, we will be using that particular lag for the model fitting. Filho and Valk (2020) have implemented VAR model-based control charts for batch process monitoring In the field of Statistical Process Control (SPC). There were many approaches to deal with monitoring the batch process. A three-way data structure (batches x variables x time-instants). For each batch, there are multivariate time series data available. In traditional approaches, they do not take the nature of the time series data into account. They used multivariate techniques on the reduced two-way data. Recent developments in SPC have proposed to use VAR concerning the original three-way structure. However, they are restricted control approaches focused on VAR Models. This study has suggested a new method to deal with the batch process focusing on VAR.

**4.0 Technologies and methods**

Secondly, working with time series forecasting is an important part of my dissertation. I have several different targets in my dissertation including dashboard visualisation, forecast modelling, database management etc. I have been looking for ways to predict the number of suicides in upcoming years. My interest in time series and ML made me dive deep into sophisticated time series models like SARIMA and VAR to make models on the suicide data and forecast future suicides in different countries. The ARIMA model is a combination of multiple models, including the Autoregressive model, the Moving average model, and the Autoregressive Moving Average model. The form of the ARIMA model is represented by ARIMA (p, d, q), where p is the autoregressive order, d is the number of differences, and q is the moving average order. Vector Auto Regressive Model is mostly used in finance and econometrics because they offer a framework for achieving important modelling goals, including data description, Forecasting, Structural Inference, and Policy Analysis. VAR Model is a workhouse time series multivariate model that relates current observations of a variable with past observations of itself and past observations of other variables in the system.

Thirdly, we need a database server for data to be stored on the server. I will be using PSQL or MySQL servers for data storage and management. I want the data in my DB to be updated from time to time and my model has to be updated based on the new data injected in each time. The reason

for choosing these DB’s is the flexibility of usage and its syntax matching with Structured Query Language (SQL) minute differences. the fascinatingcanmanycana to the sStreamlitsean

Firstly, I want to talk about the python dashboard. It’s always wonderful to see how we can make models and interpret them. But it is also important to note, recently there are some concerns about how well we can make modifications to the existing model and maintain them. So, our model has to work dynamic and make a prediction based on the available data. In recent years programmers used to use VueJS or web-based languages for making dashboards, we now have the most advanced packages like Streamlit has made these processes easier and more efficient. I am going to use some of the python packages like plotly to make an interactive dashboard and make models that can make great predictions.

**4.1 Data Preparation**

In every data analysis, about seventy percentage the total time is spent on preparing the data and making it ready for doing analysis. Initially, I had to explore the dataset using describe () method. Also using visualizations and some statistical analysis I have cleaned the dataset. Imputation was carefully done based on the time, context, and importance of the variable. I have chosen a dataset which was simple and aggregated. But, later on, thinking about the complexity and wide range of the reasons behind committing suicide I did a thorough research about how much additional information I can incorporate into the existing dataset. There have been several variables like continent missing in the dataset. So, I have added additional columns for continent names. Also, I have received another dataset which is similar to the suicide master sheet I have previously received and contained much more information. The main reason behind taking this dataset into account is that those variables were very meaningful concerning the context I am working with, for example, I assume there could be some relation between suicide rates and unemployment or the number of internet users and suicides in any country.

A screen shot of a computer

Description automatically generated with low confidence

**Fig. 4.0:** Describing the dataset using *describe()* function in pandas

Many such variables were making my research firm on the ground in terms of working with useful and meaningful information for machine learning modelling. The second main reason is that Data Visualization is a major part of my final project. If I had a greater number of variables in the dataset, I would get more opportunities of making more visualizations. Outliers in the data are one main thing we need to carefully do. Replacing the outliers without thinking about why they occur is a dangerous practice. Chart

Description automatically generated

**Fig. 4.1:** Checking for NaN or null values in the dataset

Figure 4.1 gave me a clear idea of how much data is missing in each column of the dataset. Also used visualizations like boxplot and histogram, I have explored the dataset for data preparation.

**4.2 Data Visualization**

Firstly, I want to talk about the python dashboard. It’s always wonderful to see how we can make models and interpret them. But it is also important to note, recently there are several concerns about how well we can make modifications to the existing model and maintain them. So, our model has to work dynamic and make a prediction based on the available data. In recent years programmers used to use VueJS or web-based languages for making dashboards, we now have the most advanced packages like Streamlit has made these processes easier and more efficient. I am going to use some of the python packages like plotly to make an interactive dashboard and make models that can make great predictions.

Map

Description automatically generated

**Fig. 5:** Suicide per hundred thousand around the world - Timeline in plotly

The global suicide rate is visualized using plotly. It is an animation frame page visitors can see the information based on the year. Coloured regions represent the rate of suicides per hundred thousand.

Chart, bar chart

Description automatically generated

**Fig. 5:** Suicide per hundred thousand among males and females in different age groups

As per the fig, Most suicides are happening between the age of 35 and 54. And out of the majority are Males. In all the age groups females are less affected groups. Also, we can see from the age of five to fourteen children are less likely to commit suicide.

A screenshot of a computer

Description automatically generated

**Fig. 5:** Suicide per hundred thousand around the world in different countries

As per the fig, we can see there are countries like Russia, Ukraine and Hungary that have some sort of relation in terms of the number of suicides per hundred k. As per the report from BBC News, the reasons for suicides in Ukraine are the aftermath of the conflicts between Russia and Ukraine and continuous war. Our data clearly shows that one of the most affected countries in terms of suicide is Ukraine.

Chart

Description automatically generated

**Fig. 5:** Top ten countries with the highest suicide rates

The above figure shows the scale of suicides in different countries. It is very clear to see that the Russian Federation is showing the most suicide rate among all the other countries. As per the previous studies by Bellman and Namdev (2022) it’s very clear that Russia is facing issues with suicidal behaviour from the male population and also their drinking habits have a significant effect on leading them to commit suicide.

Graphical user interface, chart, histogram

Description automatically generated

**Fig. 5:** Suicide per hundred thousand in different genders and age

Chart, line chart

Description automatically generated

**Fig. 5:** Total suicides in the world in each year distribution

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

**4.3 Modeling and Forecasting**

Secondly, working with time series forecasting is a crucial part of my dissertation. I have several different targets in my dissertation. I have been looking for ways to predict the number of suicides in upcoming years. My interest in time series and ML made me dive deep into sophisticated time series models like SARIMA and VAR to make models on the suicide data and forecast future suicides in different countries. Dealing with numerical and categorical variables I have used a decision tree classifier capable of classifying purposes. I could see Brunello et al have successfully implemented them in their research which inspired me to adapt the idea to suicide analysis. My test includes checking seasonality, tends, and stationarity and testing statistical models for finding the best model for prediction for example AIC and BIC.

Chart, line chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

AIC and BIC graphs made for checking the order of ARIMA Model. In my research I am trying to work on predictions so, I will be looking at the AIC. From the graph we can see we need to make an AR1 Model.

A screenshot of a computer

Description automatically generated with medium confidence

**4.4.1 grid search ARIMA parameters for time series**

Text

Description automatically generated

**4.4.2 Prediction using Vector Auto Regression Models (VAR Model)**

Another Model used for the time series data is VAR model. Vector Auto Regression. The reason behind using this model is that it helps in forecasting models based on multiple variables in time series. Usually, we use single variable and sequential time for time series analysis. But here I was able to include multiple variables in the model as you can see in the figure. Vector Autoregressive Models are one of the best models we could use to choose for time series.

A picture containing graphical user interface

Description automatically generated

You can see in the above example we have predicted the number of suicides for the year 2016 using VAR Model on the time series sequential data.

Chart, line chart

Description automatically generated

Graphical user interface, text

Description automatically generated

As you can see in the fig, we have calculated the accuracy of each variable. This helps us understand how well our model is performing.

**4.4.3 Prediction using Auto Regression Models (AR Model)**

The next model I have created is the Auto regression model, Train and Test were split into seventy and 30 per cent. Seventy per cent of the data was used for training the model and the rest thirty per cent was used for testing. I have got an 11.792 Root mean squared error. Also, I could save different models to the local and I was able to load the models later and update them accordingly.

Chart, line chart

Description automatically generated

As you can see in the diagram above, the suicides per hundred thousand is distributed throughout the year is shown.

Chart, line chart

Description automatically generated

In this fig, the blue line is the test data and red line is the predicted data. I have AR 1 Model with one window.

**4.4.4 Decision Tree Classifier**

I have added a new column called risk where I split the data into two classes, class 1 stands for high risk and class 0 for low risk. Using the decision tree model, I have made a classification.

Text

Description automatically generated with low confidence

**4.4 Evaluation of models**

Evaluation of the model is as important as making the model. I have created 3 models in ARMA, Auto regression and Vector Auto regression. Using mean squared error and R-squared error I took the error rates of different models. Accuracy is also calculated to understand how efficient and precise my model is. Initial modelling was done on the time series of “Russian Federation”. Final modelling will be done on a live website for all countries using python Dash or Streamlit.

Chart, line chart

Description automatically generated

After finalizing my models, I would be able to add more evaluation techniques. Above are the four diagnostic plots I have created after running ARIMA Model. Looking at the plots I can say there is no pattern in the standardized model. Looking at the Histogram there is no Gaussian distribution (green and red lines should be almost the same for gaussian). The QQ Plot seems to be not normally distributed, if it is normally distributed all the blue dots will be aligned over the line (except for some values in either end)

**4.5 Data Storage**

Thirdly, we need a database server for data to be stored on the server. I will be using PSQL or MySQL servers for data storage and management. I want the data in my DB to be updated from time to time and my model has to be updated based on the new data injected in each time. The reason for choosing these DB’s is the flexibility of usage and its syntax matching with Structured Query Language (SQL) minute differences. Initially, data was stored in the CSV format in different files. Later I uploaded them into mochahost Psql server.

**4.6 Web Server and Hosting**

We know there are thousands of hosting companies providing hosting services, I have chosen Mochahost as one of the best service providers for small business websites. My goal is to make a highly dynamic web application on the server. I have purchased a VPS service which allows running PIP Packages on their server making the server IDE more suitable for the Dash App. Mochahost cPanel will be connected to the Github repository where my application is updated from time to time. Using git technology for the hosting makes the process more sophisticated and professional in terms of version control.

**4.7 Data Security**

Data Security has become an important concern in this era. Even though the suicide dataset is publicly available on the internet, I have followed the best practices in data security to ensure there is no data leakage. I have used encrypted windows drive to store the data. Whole project codes are updated from time to time to GitHub private repository. Any information related to this study has been considered for data security and ethical practices before actually using them. No personal information is used in this study. For web applications, files are kept in a private repository and used that repo to pull changes to the live mocha host server.

**4.8 Applications and software**

Microsoft Visual Studio is the main IDE used for coding. For version control, I have used Github. I also have used other tools like Jupyter Notebook, Spyder, and Atom for coding purposes. All the testing is done in the local Anaconda environment. Python version 3.8.8 is used for the whole research. PIP package is used for configuring the IDE. Microsoft Excel is also used for minor CSV file inspections. Visual Studio’s inbuilt git version control features are used from time to time for managing the branches in the repository.

For reports and notes, MS Word and notepad are used. PowerPoint is used to create slides for presentations. Adobe Acrobat DC is used for managing PDF Files. Visual Studio in-built terminal, Anaconda Prompt and windows terminal are used for running PIP and git commands. Windows Operating System is used for the whole work. Google Chrome and Mozilla Firefox are the two browsers used in this project.

**5.0 Ethical Considerations**

* + 1. Harm is inflicted when an action of a researcher affects the participants or society. There are several reasons behind the actions of researchers that create huge trouble for society or individuals Furthermore (Human Radiation Experiments | Atomic Heritage Foundation n.d.) was one of the biggest examples of this. In 1994 US President Clinton created an advisory team to research human radiation that has been conducted over the years. In this study, doctors injected Plutonium into the body of many patients and many of them did not consent to be part of this study. Also, there was a company called Quaker Oats (2020) which is also part of this study included radioactive components in oatmeal and were unknowingly fed to the children.
    2. In my study, no such experiment is done on humans in the process of data collection or analysis. An aggregated suicide dataset only provides information about the country's general population and related detail as features is used throughout the research. No prior experiment is conducted to gather data for this research. No harm is made to any subject in this regard. There are several benefits related to the data. Data provides an overview of how many suicides are happening from time to time. Talking about the societal impact of this research is enormous. For example, (Study: Benefits of Electric Cars Add Up—in the Billions! | NRDC n.d.) has created a significant impact on how this research has benefited society to help understand the carbon footprint reduction and cost-saving. In suicide analysis, I am trying to make use of data to leverage suicide attempts by helping the govt to take measures or policies from the outcome of my study it’s going to help create plans to tackle such acts in coming years.
  1. My research about ‘suicide analyses was based on the dataset which is open source in Kaggle which anyone can download. Thinking about the data storage and security I would not say it’s a very much sensitive dataset because first of all this dataset is not private on the internet, so the author has kept the access to the public. Secondly, this suicide dataset is not using any confidential information about any individual rather it’s more of a summary dataset providing general information about the country's death rates. Also, information like the age group, internet users, human development index etc who is more likely to commit suicide.

**SWOT Analysis:**

As I previously stated, the suicide dataset did not contain any personal information. I would highly recommend for future studies we need to incorporate more humans in the experimentation to collect data from people in real-time. The most important thing we need to follow taking consent from each person who is willing to participate in the study. There are some major concerns in this regard, let’s imagine if we have not taken any consent from these human beings who are participating in the study/experiment. They might later go to court and file a complaint against us for doing illegal use of personal information. Also, we need to clearly state what are the acts or dangers involved in the experiments. So that they are aware before they participate in these activities. (Bogod 2004) is one of the real-life examples where in 1942 prisoners were asked to undergo dangerous experiments to understand the survival chance of soldiers sometimes even leading to deaths. Understanding personal, social, and business impacts of data practice.

In addition, even sharing information of individual sharing with any other colleagues or third party would be through proper procedure and getting signs on consent forms.

**Strength: -** In my study, I am trying to see suicide rates in different countries from time to time. My research strength is its dynamic nature. Similar weather forecast of google or Microsoft, my model will be run from time to time based on the latest data. This research aims at tackling suicide tendencies in every country’s population. My research is going to predict how many people are going to commit suicide in the next 5 years in different countries or continents. When working with a socially responsible research project, it is going to stand out in the world of the internet. Similar to the websites showcasing covid trends live, my website is also going to show the same impact of suicide numbers and create respective visualizations for any general audience to easily understand what the trend in data would be.

**Weakness: -** The data is aggregated and no specific information about individuals is available for forecasts. So, I think the data must be having more specific features which could make accurate predictions about the suicides. But the model would have been more accurate if more specific features could have been added to the dataset. Things like diagnostic information of subjects, population happiness index, education index, and happiness index each country. So, my prediction would be more of a general understanding of the trends in data.

**Opportunity: -** It’s unexplainable how much we can make use of the suicide data analysis. Govt. is trying to find out the reasons behind suicides or how to reduce the number of suicides every year. We can develop new strategies that can mitigate the effect of suicides through analysis and understanding of existing data. We can make use of ML models to act as smart applications which can guide mobile users based on user activity data. Suicide analysis creates a new era of AI where we can keep an eye on who is more vulnerable to death.

Now let’s look at my data and its opportunities. Have you ever thought of having a suicide prediction model for each country? The wide range of opportunities using AI and the Time Series model on big data is possible using current technologies. Internet of things, cloud computing and ML are the best examples of state-of-the-art technologies. The suicide prediction model and live dashboard visualization is a great analysis model which any growing business can take inspiration from. Just imagine a burger selling vendor creating a live predicting model of a specific kind of burger that is sold at a particular season of a year? or maybe checking bestselling milkshakes each month? Wouldn’t these analyses make them grow? or even predict how many products are going to be sold in the coming months so they can prepare their store for the coming period to avoid lack of materials. Thus, this model is ultimately showing what kind of predictions or analysis our business and health industry need today to go smarter and do smarter businesses.

**Threat**: - Data can be used in many different ways. Some people have used it for good reasons others did differently as well. The suicide dataset could be misused in some way. But in my point of view as long as we are not providing specific information about individuals, they are less likely to occur. In my analysis, what I would say is that in case more features are added to the model in the future, I will have to alter the model statically and make them dynamic using cron jobs. Also, when it comes to storing individual information in in near future, more storage space might be required as well as my model could perform poorly because of the server requirements. Even though we have other options to buy cloud storage space, it will still be costing more money on the other hand I will have to figure out ways to improve the requirements.

Talking about the analysis of suicides in previous years, there could be a political impact because of the difference in counts during different political administration periods.

**4.0 Conclusion**

**References**

*(3) 1/4: What is Streamlit - YouTube*. Available from: https://www.youtube.com/watch?v=R2nr1uZ8ffc [accessed 4 June 2022].

*Airiti Library\_Comparative+Study+of+Artificial+Neural+Network+and+ARIMA+Models+in+Predicting+Exchange+Rate*. Available from: https://www.airitilibrary.com/Publication/alDetailedMesh?docid=20407467-201211-201512080011-201512080011-4397-4403 [accessed 8 April 2022].

Bellman, V. and Namdev, V. (2022). Suicidality Among Men in Russia: A Review of Recent Epidemiological Data. *Cureus* [online], 14(3). Available from: https://www.cureus.com/articles/88128-suicidality-among-men-in-russia-a-review-of-recent-epidemiological-data [accessed 8 June 2022].

Bogod, D. (2004). The Nazi Hypothermia Experiments: Forbidden Data?. *Anaesthesia* [online], 59(12), pp.1155–1156.

Brunello, A., Marzano, E., Montanari, A. and Sciavicco, G. (2019). J48SS: A Novel Decision Tree Approach for the Handling of Sequential and Time Series Data. *Computers 2019, Vol. 8, Page 21* [online], 8(1), p.21. Available from: https://www.mdpi.com/2073-431X/8/1/21/htm [accessed 7 June 2022].

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* [online], 20(3), pp.273–297.

Filho, D.M. and Valk, M. (2020). Dynamic VAR model-based control charts for batch process monitoring. *European Journal of Operational Research* [online], 285(1), pp.296–305.

Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C. and Ferreira, C.E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology* [online], 1.

*How suicide became the hidden toll of the war in Ukraine - BBC News*. Available from: https://www.bbc.com/news/world-europe-60318298 [accessed 8 June 2022].

Huang, M.W., Chen, C.W., Lin, W.C., Ke, S.W. and Tsai, C.F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS ONE*, 12(1).

*Human Radiation Experiments | Atomic Heritage Foundation*. Available from: https://www.atomicheritage.org/history/human-radiation-experiments [accessed 15 April 2022].

John, A., Okolie, C., Eyles, E., Webb, R.T., Schmidt, L., McGuiness, L.A., Olorisade, B.K., Arensman, E., Hawton, K., Kapur, N., Moran, P., O’Connor, R.C., O’Neill, S., Higgins, J.P.T. and Gunnell, D. (2020). The impact of the COVID-19 pandemic on self-harm and suicidal behaviour: a living systematic review. *F1000Research 2020 9:1097* [online], 9, p.1097. Available from: https://f1000research.com/articles/9-1097 [accessed 7 June 2022].

Kumar, N. and Susan, S. (2020a). COVID-19 Pandemic Prediction using Time Series Forecasting Models. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020* [online], 1 July 2020.

Kumar, N. and Susan, S. (2020b). COVID-19 Pandemic Prediction using Time Series Forecasting Models. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020* [online], 1 July 2020.

*Mochahost Review 2022: Mocha Host Details, Pricing & Features | Sitechecker*. Available from: https://sitechecker.pro/web-hosting/mochahost.com/ [accessed 7 June 2022].

Qi, F., Xu, Z., Zhang, H., Wang, R., Wang, Y., Jia, X., Lin, P., Geng, M., Huang, Y., Li, S. and Yang, J. (2021). Predicting the mortality of smoking attributable to cancer in Qingdao, China: A time-series analysis. *PLOS ONE* [online], 16(1), p.e0245769. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245769 [accessed 8 April 2022].

Singh, V., Poonia, R.C., Kumar, S., Dass, P., Agarwal, P., Bhatnagar, V. and Raja, L. (2020). Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. *https://doi.org/10.1080/09720529.2020.1784535* [online], 23(8), pp.1583–1597. Available from: https://www.tandfonline.com/doi/abs/10.1080/09720529.2020.1784535 [accessed 8 June 2022].

*Study: Benefits of Electric Cars Add Up—in the Billions! | NRDC*. Available from: https://www.nrdc.org/experts/luke-tonachel/study-benefits-electric-cars-add-billions [accessed 15 April 2022].

*Suicide Statistics 2011 - CSO - Central Statistics Office*. Available from: https://www.cso.ie/en/releasesandpublications/er/ss/suicidestatistics2011/ [accessed 7 June 2022].

Tang, L., Pan, H. and Yao, Y. (2018). K-nearest neighbor regression with principal component analysis for financial time series prediction. *ACM International Conference Proceeding Series* [online], 12 March 2018, pp.127–131.

*Vector Autoregressive Models for Multivariate Time Series*. (2006). *Modeling Financial Time Series with S-PLUS®* [online], 9 October 2006, pp.385–429. Available from: https://link.springer.com/chapter/10.1007/978-0-387-32348-0\_11 [accessed 4 June 2022].

*When Quaker Oats Fed Children Radioactive Oatmeal | by Calin Aneculaesei | History of Yesterday*. Available from: https://historyofyesterday.com/when-quaker-oats-fed-children-radioactive-oatmeal-5e06faf3ce4d [accessed 9 June 2022].

Włodarczyk, T., Płotka, S., Szczepański, T., Rokita, P., Sochacki-Wójcicka, N., Wójcicki, J., Lipa, M. and Trzciński, T. (2021). Machine learning methods for preterm birth prediction: A review. *Electronics (Switzerland)*, 10(5).