**SUICIDE ANALYSIS AND PREDICTION**

**1.0 Introduction**

This iteration is a deep dive into the suicide dataset to learn much more about reasons for the thousands of suicides that occur each year around the world. Even though various studies on suicide have already been done previously, such as John et al. (2018), this study aimed to produce new insights that can help government bodies better grasp the problems that lie beneath them. This research could also benefit them in developing new strategies to minimize mortality rates over time. This research will look at a variety of suicide attributes and predict how many more fatalities will occur in various countries in the next years.

The purpose of this study is to understand the reasons behind suicides in each country. As per Wikipedia (2012), 800,000 people die by suicide every year. For example, suicide is becoming an increasing and significant issue in India, according to the World Health Organization (WHO). 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 of the suicide dataset and also there will be options for page admins to report new suicides to the dataset. I believe this dashboard can make a significant impact on society for government to identify and help who is in need which would help them mitigate the suicide figures each year in every country. By taking appropriate measures from the outcome of this study, the government is going to not only save people's lives but also make this world a better and safer place for people to live.

**2.0 Background Problem**

Different social, economic, and cultural contexts exist in different countries. Russia and Ukraine, for example, are two of the most commonly mentioned countries recently. The world is aware that the two countries are involved in a major dispute. When you see that kind of observation and data insights in Explanatory data analysis (EDA), it's always suspicious (there is a presumption that there is a relationship between the conflict and suicides in two countries); the two countries, among others, have high suicide rates.

In 2011, 554 people in Ireland committed suicide, according to the CSO statistical release (2011). In terms of the country's population, this is a large figure. Each suicide will have its own set of motives. Have you given any thought to the various reasons of these figures? You won't know the answers to these questions unless you start studying and researching suicides, like Zetzsche et al's (2007) research when they sought to figure out why people commit suicide in Western and Central Europe and came up with a few extremely interesting findings. moreover, all of the suicides might be due to a number of factors that we are not aware of. To put it differently, collecting all of that information is challenging, but there are some elements that push individuals to commit suicide in every country. Some of these common factors are included in the suicide dataset as features, which can be used to dig deeper into and analyze data from multiple countries.

**2.1.0 Research questions**

1. Examine the link between GDP per capita and the suicide rate.

2. Which country has the greatest rate of suicides in relation to its population?

3. Which age groups are at a higher risk of suicide?

4. Estimate the number of suicides that will occur in each continent during the following five years.

5. Predict the top five nations with the fewest suicides in the next five years.

6. What age group of persons is most prone to commit suicide?

**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 of 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 ,trthat .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:*** *Framework to evaluate the forecasting model from Kumar and Susan* (2020)

In Qingdao, China, another study was conducted on forecasting cancer-related deaths in 2021. They've also employed the ARIMA Model to forecast fatalities. The ARIMA method integrates the autoregressive and moving average models into one. Airiti and her team conducted another study on exchange rate forecasting. This research makes use of ARIMA and Artificial Neural Networks.

The study of 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

SVM has been applied in Time series analysis, such as when Huang et al (2017) looked at classification issues in the Breast cancer dataset. Different kernel functions employed in the SVM Classifier were also investigated in this work. RBF kernel-based SVM ensembles based on boosting outperform other classifiers on large size datasets, according to their findings. Cortes and Vapnik (1995) were the first to develop SVM, which proved to be more effective for two-group classification issues.

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 the VAR model 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**

The visualizations in this project made use of all of the data. For the time series modeling and forecasting phase, I only used data from the "Russian Federation." My dissertation includes a lot of work with time series forecasting. My dissertation has numerous objectives, including dashboard visualization, forecast modeling, database administration, and so on. I've been trying to figure out how to forecast the number of suicides in the following years. My interest in time series and machine learning led me to go into advanced time series models such as SARIMA and VAR in order to create models using suicide data and anticipate future suicides in various nations. The ARIMA model is a hybrid of many models, including the Autoregressive, Moving Average, and Autoregressive Moving Average models.

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 Database 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

Finally, 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 were some concerns about how well we can make modifications to the existing model and maintain them. So, our model must work dynamically 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 as shown in fig 4.1. Also using visualizations and some statistical analysis I have cleaned the dataset.

Graphical user interface

Description automatically generated

**Fig. 4.0:** showing information about the dataset using *info()* function in pandas

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. 4.1:*** *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 visualisation 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. 4.2:*** *Checking for NaN or null values in the dataset*

Figure 4.2 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**

The dashboard about suicide was inspired by Johns Hopkins COVID-19 Dashboard (Coronavirus Resource Center n.d.). This dashboard displays real-time Covid-19 data that is regularly updated, with corresponding graphical representations created automatically. The python dashboard, which visualizes the data and insights for the public, will be the most appealing aspect of this project. Python dash and streamlit have emerged as the most capable frameworks for web-based visualisation projects in recent years. This project provides both static and dynamic visualisations. Before the real web dashboard app, individual static graphs are produced to obtain insight from the data. A final dashboard app with dynamic visualisations will be constructed when the initial static models are completed in Jupyter Notebook.

**Suicide per hundred thousand around the world**

Map

Description automatically generated

**Fig. 4.2:** 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. Colored regions represent the rate of suicides per hundred thousand.

**Suicide per hundred thousand in Male and Female**

Chart, bar chart

Description automatically generated

***Fig. 4.3:*** *Suicide per hundred k among males and females in different age groups in different countries*

As per fig 4.3, 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.

**Suicide per hundred thousand Vs GDP Per capita**

A screenshot of a computer

Description automatically generated

***Fig. 4.4:*** *Suicide per hundred thousand around the world in different countries*

As seen in the fig 4.4, nations such as Russia, Ukraine, and Hungary have some form of relationship in terms of the number of suicides per 100,000 people. According to a BBC News article, the causes of suicide in Ukraine are the consequences of Russia-Ukraine hostilities and the ongoing war. Ukraine is undoubtedly one of the most afflicted countries in terms of suicide, according to our statistics.

**Top ten countries with the highest suicide averages**

Chart

Description automatically generated

***Fig. 4.5:*** *Top ten countries with the highest suicide rates*

The extent of suicides in various nations is seen in Figure 4.5. It is apparent that the Russian Federation has the highest suicide rate of all the countries studied. According to earlier research by Bellman and Namdev (2022), Russia has a considerable problem with suicidal behavior among men, and their drinking habits have a substantial impact on their decision to commit suicide when compared to other nations.

**Population, Suicide, Suicide in Hundred Thousand Vs Gender and Age**

Graphical user interface, chart, histogram

Description automatically generated

***Fig. 4.6:*** *Suicide, Suicide per hundred thousand, Population in different genders and age in all countries*

As seen in fig. 4.6, Females outnumber men in terms of population. In relation to total suicides among men and women, the majority occurred between the ages of 25 and 34. It is undeniable that the male population has a higher suicide rate per 100,000. Furthermore, the majority of those who died were above the age of 75.

**Population, Suicide, Suicide in Hundred Thousand Vs Gender and Age**

Chart, line chart

Description automatically generated

***Fig. 4.7:*** *Total suicides in the world in each year distribution*

Fig 4.7 shows that there is a quick rise in the suicide rate from the year 1990. After that, it continued increasing until the next ten years. From 2000 it started to decline.

**Fig. 4.8:** Distribution of suicide per hundred thousand and GDP

Chart, line chart

Description automatically generated

According to the fig 4.8 above, the number of suicides per 100,000 climbed from 1985 to 1995, then rapidly fell. At the same time, global GDP per capita climbed gradually from 1985 to 1995, remained stable until 2003, then increased abruptly until 2014, before falling precipitously in 2015. (World Economic Situation UN). Life expectancy and unemployment have been influenced by missing data and imputation, making the lines appear irrelevant.

Chart

Description automatically generated

**Fig. 4.10:** correlation matrix of suicide dataset

Normally, checking for features in the correlation matrix is necessary to see whether there are any features with strong correlation. We usually eliminate such variables from the dataset since they might cause the model to overfit. As seen in fig. 4.10, there is a significant correlation between suicide and suicide per hundred thousand. As a result, before modeling, the suicides feature was deleted from the dataframe.

**4.3 Modeling and Forecasting**

This project has various objectives, including creating a dashboard for visualizations, forecasting using Machine Learning Models, and creating an Admin Panel Portal. Working with a time series model and projecting future values would be the most intriguing and challenging aspect of this endeavor. A decision tree classifier was employed to categorize the risk and non-risk groups while dealing with numerical and categorical information. Brunello et al . seem to have effectively utilized them in their research, which prompted me to extend the concept to suicide analysis. Checking seasonality, trends, and stationarity, as well as assessing prediction tests for identifying the optimal model for prediction, such as AIC and BIC, are all part of this research.

Chart, line chart

Description automatically generated

**Fig. 4.11:** ARIMA Forecast and observed value graph on Russian data

The output of the ARIMA forecasting model employing the one step forward strategy is shown in Figure 4.11. The blue line depicts the actual result from the dataset, which is the number of suicides per 100,000 in Russia. The red line represents the expected values based on historical values, whereas the blue line represents the actual values. The conf\_int() method from forecast is used to determine the confidence interval, which returns a pandas object. Lower and upper uncertainty limits for a Dataframe. The prediction's uncertainty limit is indicated by the pink shaded region.

AIC and BIC graphs were made for checking the order of the ARIMA model. In my research I am trying to work on predictions so, I will be looking at the AIC. A lower AIC score means a better predicting model. If the order is set too high, it could result in a high AIC value, this stops us from overfitting the training data. BIC is similar to AIC, lower BIC indicates a better model. BIC likes to choose a simple model with the lower order.AIC is better at predictive models but BIC is choosing a good explanatory model.

Text

Description automatically generated

**Fig. 4.12:** AIC and BIC Score in ascending order

Here Fig 4.12, I am looking at a better predicting model so, I will be choosing AIC with the least score. This is an ARMA(2,0) Model

Graphical user interface, chart

Description automatically generated

**Fig. 4.13:** ACF and PACF to choose the model in ARIMA

From fig 4.13, In the above ACF and PACF, we can see ACF tails off and PACF cuts off since we have a MA(q) model. So this is an AR(1) Model.

A screenshot of a computer

Description automatically generated with medium confidence

**Fig. 4.13:** Sarimax Model Results with order (1,0,0)

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

We can automate the process of training and evaluating the ARIMA Models on different combinations of model hyperparameters. In machine learning, this is called grid search.

Text

Description automatically generated

**Fig. 4.14:** grid search result from ARIMA Model

In fig 4.14, we can see I have implemented the grid search and evaluated the ARIMA Model. Also, I have evaluated a set of different parameters.

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

Another Model used for the time series data is the 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

**Fig. 4.15:** predicting future values in VAR Model

You can see in fig 4.15, that we have predicted the number of suicides for the year 2016 using VAR Model on the time series sequential data.

Text

Description automatically generated

**Fig. 4.16:** Regression result summary from VAR Models

As I mentioned earlier, I was able to make predictions on multi-variate time series data using VAR Models. Fig 4.17 shows the actual value of and predicted values distribution in VAR Model.

Chart, line chart

Description automatically generated

**Fig. 4.17** Plot of Forecast vs Actuals from VAR Models

From fig 4.17 you can see how the model forecasted different variables in the Russian suicide dataset concerning actual values.

**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

**Fig. 4.18** Plot of Forecast vs Actuals from VAR Models

As you can see in the diagram above, the suicides per hundred thousand are distributed throughout the year in ‘Russian Federation’ is shown.

Chart, line chart

Description automatically generated

**Fig. 4.19** Plot of Forecast vs Actuals from AR Models

In this fig 4.18, the blue line is the test data and the red line is the predicted values. 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

**Fig. 4.20** Plot of Forecast vs Actuals from AR Models

Fig 4.20 shows the result from the Decision Tree classification algorithm. Lee and Oh (1996) have done studies on Neural networks using a Decision Tree classifier to distinguish between complex features. Here my goals are to separate the risk group and non-risk groups based on the feature called risk. I got 99.62 training accuracy and a hundred per cent testing accuracy.

**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. Despite the fact that the model below was constructed using a Russian dataset. The live dashboard and modeling for the final product will be based on the whole country's master dataset.

Chart, line chart

Description automatically generated

**Fig. 4.21** ARIMA Diagnostic plots

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)

Graphical user interface, text

Description automatically generated

**Fig. 4.22** Plot of Forecast vs Actuals from VAR Models

4.22

**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. To begin, consider some recent unethical events. When a researcher's action has a negative impact on participants or society, it is considered harm. There are a number of reasons why researchers' actions cause so much harm for society or people. Likewise, one of the most obvious examples of such accidents was Human Radiation Experiments (2017) was one of the biggest examples of such incidents. 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 of Benefits of Electric Cars (2016) 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.

**5.0.1 SWOT Analysis**

The suicide dataset, as previously indicated, did not contain any personal information. I strongly believe that in future research, we should include more humans in the trial in order to get data from individuals in real time. The most critical step is to obtain written consent from everyone who wishes to participate in the study. Consider what would happen if these human individuals who are participating in the study/experiment had not given their consent. They may eventually take us to court (Facebook Data Breach BBC) and file a complaint accusing us of misusing personal information. In addition, we must explicitly disclose what activities or hazards are involved in the research. So that people 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 hard to describe how much we can learn from suicide data analysis. The government is attempting to determine the causes of suicides as well as ways to lessen the number of suicides each year. Through the study and interpretation of current data, we can design innovative techniques to reduce the impact of suicides. Machine Learning models may be used to create smart applications that assist mobile users based on their activity data. Suicide analysis ushers in a new era of AI in which we can monitor who is at risk of dying.

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 MACHINE LEARNING 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.

**6.0 Conclusion**

In my study I have used for models including ARIMA, VAR and AR to analyse and predict the effect of suicide in different countries. As an initial step, all the models were created on the ‘Russian Federation’ suicide dataset. Out of the four models, ARIMA model performed well compared to other algorithms. I would also like to include the SVM model also in the coming days to further improve or identify more efficient models if there are any. ARIMA and Decision tree classifiers gave me better accuracy compared to other models. From this stage, I will be working on the main dashboard which works online to make live predictions based on the input data.

**7.0** **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.

*COVID-19 Map - Johns Hopkins Coronavirus Resource Center*. Available from: https://coronavirus.jhu.edu/map.html [accessed 11 June 2022].

*Facebook Files: 5 things leaked documents reveal - BBC News*. Available from: https://www.bbc.com/news/technology-58678332 [accessed 12 June 2022].

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., Glendenning, A.C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., Lloyd, K. and Hawton, K. (2018). Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review. *J Med Internet Res 2018;20(4):e129 https://www.jmir.org/2018/4/e129* [online], 20(4), p.e9044. Available from: https://www.jmir.org/2018/4/e129 [accessed 9 June 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.

Lee, K.C. and Oh, S.B. (1996). An intelligent approach to time series identification by a neural network-driven decision tree classifier. *Decision Support Systems* [online], 17(3), pp.183–197.

*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.

Värnik, P. (2012). Suicide in the World. *International Journal of Environmental Research and Public Health* [online], 9(3), pp.760–771.

*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).

*World Economic Situation and Prospects 2015 | Department of Economic and Social Affairs*. Available from: https://www.un.org/development/desa/dpad/publication/world-economic-situation-and-prospects-2015/ [accessed 12 June 2022].

Zetzsche, T., Bobes, J., de La Fuente, J.M., Pogarell, O., Norra, C., Schmidtke, A., Wasserman, D., Löhr, C. and Rihmer, Z. (2007). Changing suicide rates in western and central Europe. *European Psychiatry* [online], 22(S1), pp.S35–S35. Available from: https://www.cambridge.org/core/journals/european-psychiatry/article/changing-suicide-rates-in-western-and-central-europe/1B2C943626D6D31150E00FCD3CECFDA9 [accessed 11 June 2022].