Analysis of Covid-19 data and predicting future coronavirus cases by using Machine learning algorithms

by

Uldana Zhaksybay

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

Apr 2023

Author . . . .
Department of Computer Science
Apr 27, 2023

Certified by . . . .
Askar Boranbayev
Assistant Professor
Thesis Supervisor

Accepted by . . . .
Vassilios D. Tourassis
Dean, School of Engineering and Digital Sciences

Analysis of Covid-19 data and predicting future coronavirus cases by using Machine learning algorithms

by

Uldana Zhaksybay

Submitted to the Department of Computer Science
on Apr 27, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Background: The Covid-19 pandemic has posed significant challenges to healthcare systems worldwide. Effective strategies to manage the pandemic require accurate and timely forecasting of the spread of the virus. Machine learning (ML) algorithms offer a promising approach for predicting the number of Covid-19 cases.
Objectives: This thesis work aims to analyze the Coronavirus data, and the number of cases and predict the future behavior of Covid-19 in Kazakhstan which helps to make key decisions related to the virus and prevent the country from the global economic crisis.
Methods: The study utilized publicly available data sources to create a comprehensive Covid-19 dataset. The dataset included daily counts of confirmed Covid-19 cases, deaths, recoveries, and tests across multiple countries and regions worldwide. This work used four ML algorithms in our study, including a decision tree, random forest, linear regression (LR), and polynomial regression. Evaluation of the performance of the models based on r2 score, MAE, MSE.
Results: Results showed that all four ML algorithms produced reasonably accurate predictions of Covid-19 cases. The random forest and decision tree algorithms outperformed the other models, with an accuracy rate of over 85% and 90% respectively. The linear and polynomial regression models had accuracy rates of approximately over 75%.
Conclusion: In conclusion, this study demonstrates the potential of ML algorithms for predicting the number of Covid-19 cases. Findings suggest that the random forest algorithm is the most effective in forecasting Covid-19 cases. The results of this study may help inform policymakers and healthcare professionals in developing effective strategies to manage the Covid-19 pandemic.

Thesis Supervisor: Askar Boranbayev
Title: Assistant Professor

Acknowledgments

I would like to show my sincere gratitude to Prof. Askar Boranbayev for supervising my thesis and guiding me during the process of writing thesis work and giving helpful feedback.

Also, I would like to thank my co-supervisor Prof. Siamac Fazli who helped to write my thesis work correctly and provide criteria and tips on writing my thesis work properly.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and context

It was almost 3 years since the outbreak of the virus called Covid-19 which bring to a global pandemic situation. The outbreak of the Covid-19 pandemic has brought unprecedented challenges to public health systems worldwide. The rapid spread of the virus has highlighted the urgent need for accurate forecasting of Covid-19 cases to help governments and health organizations make informed decisions regarding resource allocation, disease control measures, and patient care. The situation becomes better because of studying and taking some actions. However, scientists find out mutations of the coronavirus.

In recent years, machine learning (ML) algorithms have shown remarkable success in predicting the spread of infectious diseases. This thesis work aims to analyze and implement prediction or forecasting of Covid-19 future cases by looking at confirmed, recovered, and death datasets. The novelty of this thesis work will be the geographical factor. That means in the thesis work will be done prediction of Covid-19 cases exactly in Kazakhstan. The main goal of the study is to evaluate the performance of these algorithms and identify the most accurate prediction models for forecasting Covid-19 cases. By doing this, I hope to be able to contribute to the study of the impact of the pandemic and support effective decision-making of public health policy.

1.2 Aim

This thesis work aims to analyze the Coronavirus data and the number of cases and predict the future behavior of Covid-19 in Kazakhstan which helps to make key decisions related to the virus and prevent the country from a global economic crisis.

1.3 Research Questions

This research is about analyzing cases of Covid-19 and making some predictions about the future situation. Here are the following questions should be answered in the thesis work:

- What is the best methodology to use for the prediction of Covid-19 cases?
- What kind of prediction methodologies will be used to make an analysis of coronavirus cases?
- What problems may encounter when predicting the Covid-19 situation?

1.4 Research approach and methodology

The study utilized publicly available data sources, including the World Health Organization (WHO) and John Hopkins University, to create a comprehensive Covid-19 dataset. The dataset included daily counts of confirmed Covid-19 cases, deaths, recoveries, and tests across multiple countries and regions worldwide. For the prediction of these datasets will be used Machine Learning algorithms such as Linear Regression, Polynomial Regression, Random Forest, and Decision Tree. As a result, the accuracy of models will be performed and compared with each other to find the best model.

1.5 Scope and limitations

The main task of this thesis work is to develop a prediction model that will forecast accurately coronavirus cases in a specific country which is Kazakhstan. It is a very important task that can be helpful for the contribution of scientific understanding and transmission dynamics of coronavirus. During the writing of the thesis may be limitations in providing a dataset because of geographical and privacy reasons. The second limitation that might occur is the data quality which is the completeness and accuracy of data that will be used for testing and training the model.

1.6 Outline

The structure of the thesis is divided into some parts, which are as follows:

- Chapter 1: This chapter contains the context and background of what this thesis is about, research questions, motivation, and research approach
- Chapter 2: In this chapter, a summary of all papers related to thesis work is written
- Chapter 3: The third chapter discusses and answers research questions. It includes methods and analysis like data preprocessing, data preparation, applications that are used, and a few summaries about tools.
- Chapter 4: This chapter includes results obtained from this research.
- Chapter 5: The last chapter consists of the conclusion about the thesis work and a discussion of possible future work.

Chapter 2

Related works

Alavikunhu Panthakkan et al. suggested a deep learning model built on the VGG16 architecture and predicted COVID-19-positive cases. The authors used performance metrics such as accuracy, precision, recall, and f1 score. As a result obtained an accuracy of 99.5%, which is significantly better than other methodologies. The experiment was performed with 2000 X-ray specimens. For this research authors collected a large dataset of X-ray images and the system identified them by dividing them into Covid-19 positive case and Normal. [1]

Ersin Elbasi et al. in this research made a survey about the intersection of Machine learning, the Internet of things (IoT), and the Covid-19 situation. They presented a comprehensive survey about the usage of IoT and Machine learning in the context of Covid-19. They have emphasized the promise of these tools in improving COVID-19 diagnosis accuracy and speed, predicting disease progression, and remotely monitoring patients. [2]

Saud Shaikh et al. implemented the two regression models as linear and polynomial and for forecasting future trends used the time series forecasting approach of the tableau. The evaluation of two types of regression models used the R-squared score and error values. The paper contains a lot of analysis and results in the format of tables. As a result, the authors decided that polynomial regression is better than linear regression.[3]

Aradhna Saini et al. identified and predicted the spread of Covid-19 by implementing a hybrid model based on deep learning and machine learning techniques. The research collected recovered and death cases from three countries Brazil, India, and the U.S. However, the paper mostly related to the population of India. The models like Naive Bayes and Linear regression were used. [4]

Anvesh Chitturi et al. identified patterns in data to predict whether a person is infected with coronavirus or not. For the prediction of this case were used ML classifications as Decision Tree, Support Vector Machine, Random Forest, Logistic Regression, and Naive Bayes. The highest performance among these classifiers was the Random Forest with an accuracy of 99.03%. This paper not only predicts whether a person has Covid-19 or not but also suggests Anti-Covid strategies like Social distancing and Mask detection. [5]

Harika Bandarupally et al. forecasted the daily, recovered cases, and deaths caused by viruses using Long short-term memory networks. The authors used the dataset that was obtained from John Hopkins University's publicly available datasets. There were other deep learning models like multi-layer perceptrons (MLP), Convolution Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). However researchers decided to take LSTMN by the criteria: simple, well-understood, robust to noise, approximate non-linear functions, can handle multi-step forecasts and multivariate inputs.[6]

Chapter 3

Methodology

In this thesis work, 2 types of research methodologies were used: Systematic literature review and Experiment. As mentioned in the Introduction part we have some research questions that should be answered. To answer research question 1 was made Systematic literature review through which we find out what is the best Machine learning algorithm or methodology for making predictions of Covid-19 cases. The second part of the thesis work conducted an experiment, that forecasted future cases of Covid-19 and was performed accuracy. This part answered research questions of what kind of ML algorithms used for analysis and prediction and what are the future trends of coronavirus cases.

3.1 Systematic literature review

The SLR conducted several steps, which are:

- **Determining the keywords:** The keywords that are most likely to be identified is Machine Learning algorithms, Covid-19, prediction, comparison, and forecasting.

- **Comprehensive search:** To locate all pertinent studies that address the research topic, a systematic search is carried out. Multiple databases are usually searched, specific keywords are used, and inclusion and exclusion criteria are applied.

- **Relevance:** The studies will be screened for relevance based on their titles, abstracts, and complete texts to see if they satisfy the exclusion and inclusion criteria.

- **Quality assessment:** Evaluation of papers using a predetermined set of factors, such as the study design, sample size, and statistical techniques, the quality of the studies.

- **Synthesizing the outcomes:** Combine the findings of the chosen studies by a variety of techniques, including narrative synthesis and meta-analysis.

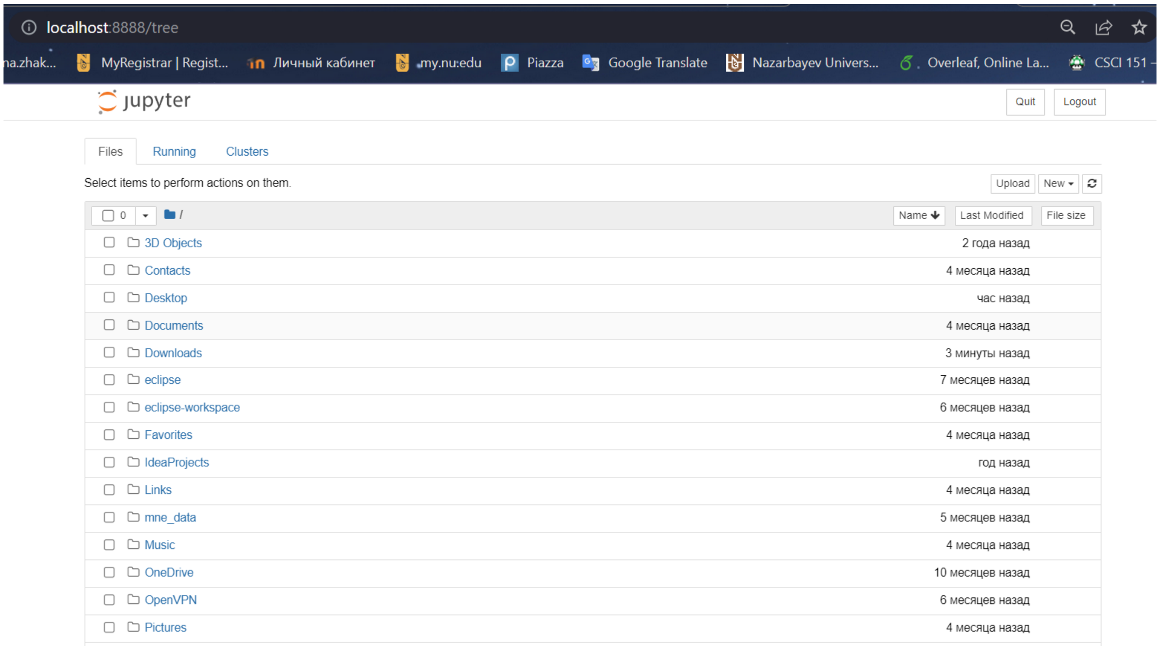
- **Results:** The results of the systematic literature review are presented in a clear, structured way in accordance with accepted reporting standards.

3.2 Experiment

3.2.1 Software Toolset

Jupyter Notebook

Jupyter Notebook is an extremely handy tool for creating beautiful analytical reports because it allows you to store code, images, comments, formulas, and graphs together. Jupyter notepads allow you interactively work with code directly in the browser. A Jupyter notepad consists of independently run code cells that, when run, execute the program code contained in the cell and add the appropriate names of variables, functions, classes, etc. to the namespace. The Jupyter notebook will be opened by writing the command jupyter-notebook in the terminal. However, before that, the package anaconda must be installed. Then, by link, it will go to the web browser. Figure 3.1 below shows the view when you open jupyter notebook:



Select items to perform actions on them.

Name	Last Modified
□ 3D Objects	2 года назад
□ Contacts	4 месяца назад
□ Desktop	час назад
□ Documents	4 месяца назад
□ Downloads	3 минуты назад
□ eclipse	7 месяцев назад
□ eclipse-workspace	6 месяцев назад
□ Favorites	4 месяца назад
□ IdeaProjects	год назад
□ Links	4 месяца назад
□ mne_data	5 месяцев назад
□ Music	4 месяца назад
□ OneDrive	10 месяцев назад
□ OpenVPN	6 месяцев назад
□ Pictures	4 месяца назад

Figure 3-1: Jupyter notebook

Python

In thesis work will be used this kind of python libraries:

- **Pandas** package used for data analysis and cleaning. It is also called a data manipulation package. It has the DataFrame object that lets us work with tabular data which is very similar way what would do in a spreadsheet. Most of the time pandas is helping with data cleansing, merging and connecting databases, and data wrangling. It also includes powerful data visualization and time-series analysis tools.

- **Matplotlib** is a data visualization package that produces 2D plots and graphs. The package contains a lot of plotting tools for creating line graphs, scatter plots, pie plots, bar plots, histograms, and more. Matplotlib has 2 types of tools pyplot and pylab which are used to create charts. Matplotlib is also used in data analysis and scientific computing.

- **NumPy** is a Python library for numerical computing. The library stores and manipulates a large amount of numerical data by offering a multi-dimensional array. NumPy has many mathematical functions such as trigonometric functions, logarithms, and statistics. These functions are used for working with arrays. Like Matplotlib, Numpy is also used in data analysis and scientific computing. However, Numpy is used for numerical calculations.

- **Scikit-learn** is a python package that is used for data analysis and predictive modeling. It includes machine learning algorithms for classification, regression, clustering, and dimensionality reduction. Scikit-learn additionally includes data preparation, model selection, and assessment tools. It is really helpful for data scientists and machine learning engineers who should perform prediction tasks with large datasets.

3.2.2 Dataset

Data collection

The data collection was one of the most important parts of this thesis work. To predict data correctly and give excellent performance metrics we need a really good dataset. There was no access to official databases because they conducted private information about patients. Because of that for this thesis work dataset was collected from open sources. One of the most popular datasets was taken from the website of WHO and from the John Hopkins University site. Also, from this site were taken Covid-19 cases divided by regions of Kazakhstan - https://data.humdata.org/dataset/kazakhstan-coronavirus-covid-19-subnational-cases.

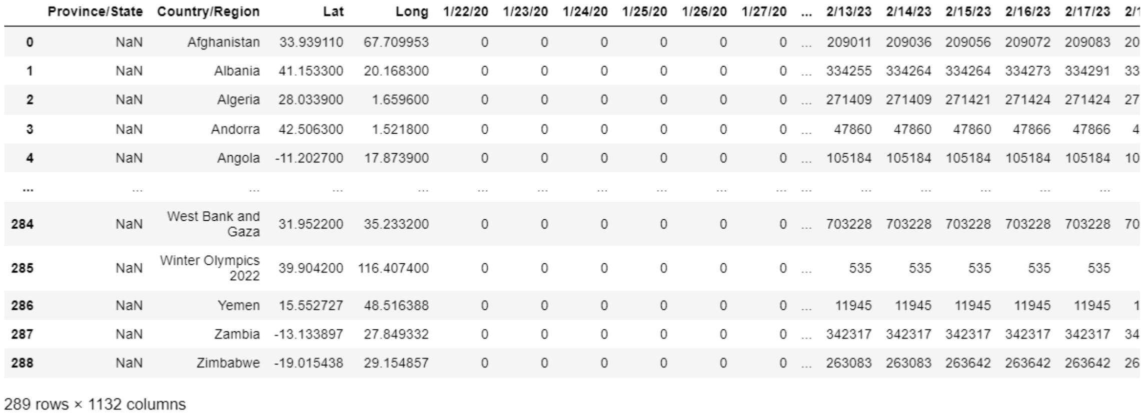
Applied dataset

To predict the future cases of Covd-19 were used 4 types of datasets:

- cases_kazakhstan.csv
- time_series_covid19_confirmed_global.csv
- time_series_covid19_deaths_global.csv
- time_series_covid19_recovered_global.csv

Name	**Type**	**Example**
date	datetime	"2020-03-27", "2022-12-09"
name	string	"Akmola Region", "Jambyl Region"
iso3166-2	char	"KZ-AKM", "KZ-ZHA", "KZ-KUS"
iso3166-1	string	"KZ"
cases	integer	"3", "9", "34", "24", "15"
cumulative_cases	integer	"10", "19", "29070", "29151"

Table 3.1: Features description of dataset 1



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	2/13/23	2/14/23	2/15/23	2/16/23	2/17/23	2/1
0	NaN	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	209011	209036	209056	209072	209083	20
1	NaN	Albania	41.153300	20.168300	0	0	0	0	0	0	...	334255	334264	334264	334273	334291	33
2	NaN	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	271409	271409	271421	271424	271424	27
3	NaN	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	47860	47860	47860	47866	47866	4
4	NaN	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	105184	105184	105184	105184	105184	10
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
284	NaN	West Bank and Gaza	31.952200	35.233200	0	0	0	0	0	0	...	703228	703228	703228	703228	703228	70
285	NaN	Winter Olympics 2022	39.904200	116.407400	0	0	0	0	0	0	...	535	535	535	535	535	
286	NaN	Yemen	15.552727	48.516388	0	0	0	0	0	0	...	11945	11945	11945	11945	11945	1
287	NaN	Zambia	-13.133897	27.849332	0	0	0	0	0	0	...	342317	342317	342317	342317	342317	34
288	NaN	Zimbabwe	-19.015438	29.154857	0	0	0	0	0	0	...	263083	263083	263642	263642	263642	26

289 rows × 1132 columns

Figure 3-2: Time series Covid-19 global dataset

Name	Type	Example
Province/State	string	"Saskatchewan", "NaN", "Zhejiang"
Country/Region	string	"Albania", "Zimbabwe", "Kazakhstan"
Lat	float	"33.939110 ", "42.506300", "-13.133897 "
Long	float	"67.709953", "20.168300", "116.407400"

Table 3.2: Features description of dataset 2,3,4

3.2.3 Data Preprocessing and Analysis

Data cleaning: First of all, the downloaded data was cleaned in all countries except Kazakhstan. As a result, there were 3 datasets which are confirmed, recovered, and death cases with only Kazakhstan data.
Data normalization: The format of datasets was irrelevant to read where date from 2020 to 2023 were columns, not rows. That's why they were normalized to a better format and 3 cases concatenated for better analysis.
Missing values: In the dataset, where conducted data by regions, were replaced NaN values with previous non-NaN values for creating more relevant and understandable analysis.

3.2.4 Model Selection and Implementation

Implementation starts from importing necessary libraries such as pandas, numpy, matplotlib, sklearn, seaborn, and warnings. It also registers matplotlib converter, sets seaborn style, and filters out warnings. Next, it reads data named "combined_Kazakhstan.csv" which is the combination of 3 time series datasets. After we find out daily confirmed, recovered, and death cases by calculating the daily cases from the 'Confirmed', 'Recovered', and 'Death' columns by subtracting the previous day's cases using the shift() method. Then the columns 'Confirmed_pastday', 'Confirmed_2daysago', and 'Confirmed_3daysago' are created to include the previous values of the 'Daily_confirmed' column as features for the prediction of the current 'Daily_confirmed' value. This is done to take into account the trend and pattern of the time series data. By including these lagged values as features, the model can learn from the past patterns in the data and make more accurate predictions. The same operations were performed with recovered and death cases too.

For implementing the prediction model for Covid-19 cases chosen these Machine learning algorithms:

Linear Regression

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

Polynomial Regression

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import train_test_split
```

Random Forest

A random forest regressor model is created with the specified hyperparameters using the DecisionTreeRegressor class from sklearn.

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
```

Decision Tree

A Decision Tree regressor model is also created with the specified hyperparameters using the RandomForestRegressor class from sklearn.

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
```

The features and target variables are defined, and the data is split into training and testing sets using the train_test_split() method. The model is then fitted to the training data using the fit() method, and predictions are made on the test set using the predict() method. A plot is generated to compare the actual and predicted values.

3.2.5 Performance metrics

The proposed work shows the use of Machine learning techniques for prediction purposes. This research tries to predict if there is an increase or decrease in no. of cases. with the help of data analysis, we can clearly see the situation and use them for prediction. This data will help in further decision-making. After comparing the types of data mining techniques I will choose one with better performance. In the paper, the following metrics will be used:

R-squared score ($\mathbf{R}^2$)

R^2, also called as the coefficient of determination, is used for performing the accuracy of regression models. The best possible score is 1.0 which means that your model is perfect. However, it is impossible to take the performance metrics of 1.0 because the prediction could not be 100% correct. Here is the formula of R-squared performance metric:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where, $\mathbf{R}^2$ is the coefficient of determination
RSS is the sum of squares of residuals
TSS is the total sum of squares

Mean Absolute Error (MAE) and Mean Squared Error (MSE) are two commonly used metrics for evaluating the performance of regression models.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is the average absolute difference between the predicted and actual values. It measures the average magnitude of the errors in a set of predictions, without considering their direction. The formula for MAE is:

$$MAE = (\tfrac{1}{n}) * \sum(|y - \hat{y}|)$$

where,

n is the number of samples

y is the actual values

$\hat{y}$ is the predicted values

Mean Squared Error (MSE)

Mean Squared Error (MSE) is the average squared difference between the predicted and actual values. It measures the average of the squares of the errors. The formula for MSE is:

$$MSE = (\frac{1}{n}) * \sum(y - \hat{y})^2$$

where,

n is the number of samples

y is the actual values

$\hat{y}$ is the predicted values

Both MAE and MSE are used to evaluate the performance of regression models, but they differ in their sensitivity to outliers. MSE puts more weight on large errors, while MAE treats all errors equally.

Chapter 4

Results

4.1 Systematic Literature Review Results

The Literature Review was done to answer research question 1 (RQ1). What is the best methodology to use for the prediction of Covid-19 cases? There we should find out what method or model is the best to predict coronavirus cases.

Title	**Findings**
1. Autism Prediction using ML Algorithms [7]	In this research used 3 types of algorithms: The Random Forest, Ada Boost, and Logistic Regression. As a result, the least suitable algorithm with the least accuracy score was the Random Forest. While the Logistic Regression showed the best accuracy result among the three and was decided as the most suitable algorithm
2. An Educational-based Intelligent Student Stress Prediction using ML [8]	This paper implemented the prediction of stress in college and used Naive Bayes and KNN as an ML algorithm. The result showed that Naive Bayes has a better accuracy rate of 94%. The authors also suggested that CNN and Random Forest algorithms also can be used for evaluation
3. Predicting Covid-19 by Referring to Three Supervised ML Algorithms: A Comparative Study using WEKA [9]	This work used three machine learning techniques to predict Covid-19. They are J48 Decision Tree, Random Forest, and Naïve Bayes. As a comparison tool used WEKA and analyzed models in ten-fold cross-validation. The result demonstrates that the Random Forest was the best method with an accuracy of 98.81% and a 0.022 mean absolute error.

4. Machine Learning based Diabetes Prediction using with AWS cloud [10]	The authors of this research used many different types of ML algorithms like artificial neural networks(ANN), XG boosting, Ada boosting, K Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree (DT). In conclusion, the authors did not choose one technique as the best. However, if we take look at the table with a comparison of accuracies, it is clear that Decision Tree has the highest accuracy score among them
5. Performance Evaluation of Supervised ML Algorithms for Elephant Flow Detection in SDN [11]	The purpose of this paper was to predict large-size traffic. For this were utilized algorithms such as Naive Bayes (NB), K-Nearest neighbors (KNN), Logistics regression (RL), Support Vector Machine (SVM), and Decision Tree (DT). The best algorithms with the highest performance were KNN and Decision Tree, while SVM performed the least accuracy result in making a correct prediction
6. AI-Enabled Covid-19 Prediction Methods and AntiCovid Strategies [5]	Aim of this paper was to detect whether a person is infected with Covid-19 or not and performed classification algorithms like Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, and Logistic Regression. Among these ML algorithms, the Random Forest showed better result with 99.31% test accuracy.
7. Analysis And Implementation of a Novel AI-Based Hybrid Model for Detecting, Predicting and Identification Of COVID-19 Spread [12]	In this paper, authors did not compare different Machine learning techniques, they just demonstrate hybrid format of usage. The hybrid format contains the algorithms of Linear Regression (LR) and Naive Bayes (NB).
8. Analysis and Prediction of Covid-19 using Regression Models and Time Series Forecasting [3]	This paper is the most related paper to this thesis work. The authors demonstrate the prediction of Covid-19 cases by using regression models such as Linear and Polynomial regression. The performance metrics like R2 score and MAPE are presented. By looking at the results the authors came to the conclusion that Polynomial regression is better than Linear regression.
9. COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks [6]	In this paper time series forecasting of daily cases, recoveries, and death were performed. The model for forecasting used the Deep learning model long-short-term memory network (LSTM). Besides LSTM authors also analyzed networks such as Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN)
10. Polynomial based linear regression model to predict covid-19 cases [13]	Research to predict the upcoming cases of Covid-19 by current situation using polynomial-based Deep learning model (Linear regression). As a result, the model showed 99.29% accuracy performance which is a really good result.

11. Diabetes Prediction and Classification using Machine Learning Algorithms [14]	Diabetes is one of the most widespread diseases and have no cure from some stage. This paper made a prediction and classification of diabetes using ML algorithms. The authors used 3 different datasets for prediction. As an ML model, they took Logistic Regression, Naïve Bayes, Support Vector Machine and Random Forest. For evaluation metrics were used accuracy, precision, recall, F1-score, and Kappa index. As a result, the Random Forest model showed the highest accuracy with 99%.
12. Prediction of People's Abnormal Behaviors Based on Machine Learning Algorithms [15]	The authors tried to apply Machine learning to Computer vision for detecting improper behaviors such as smoking in a gas stations or talking on the phone while driving. This paper is about using ML algorithms for predicting people's abnormal behaviors. For prediction used these models: Linear Support Vector Machine(LSVM), Decision Tree(DT), Random Forest (RF), Kernel Support Vector Machine(KSVM), and K-means clustering. The result showed that Random Forest had the best accuracy of 82%.
13. Global Prediction of COVID-19 Cases and Deaths using Machine Learning [16]	The paper predicts the number of Covid-19 cases with high accuracy by using SVR and PR models. The authors forecasted the number of recovered cases, deaths, confirmed cases, and daily case count. As a result, the SVR model outperformed other models like linear, polynomial, and logistic regression .
14. Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning [17]	Because of the highest price of newly produced cars, used cars are gaining popularity. In this research were predicted the price of used cars using Artificial Neural Networks And Machine Learning. The models Keras Regressor, Random Forest, Lasso, Ridge, and Linear regressions are built. Among these models, the Random Forest model gave less error with a Mean Absolute Error value of 1.0970472 and an R^2 error value of 0.772584.
15. Heart Disease Prediction Using Supervised Machine Learning Algorithms [18]	Heart disease is one of the most dangerous illnesses for people and it is really important to detect it in the early stages. This paper predicts cardiac illnesses in the human body using KNN, NB, LR, and RF on the basis of medical parameters. Among these ML algorithms, Logistic regression showed the best result of 90.2%.

By conducting SLR and analyzing scientific papers, it became clear which Machine learning algorithms are showing high accuracy and which one of them is not good in prediction. Here are some conclusions from the Systematic literature review. The papers [5], [8], [9], [14], [15], and [17] showed the best accuracy with the Random Forest algorithm which means RF is one of the best models in prediction tasks. The Decision Tree took second place by making also excellent results on evaluation metrics. Also, some papers like [3], [5], [7], [11], [12], [13], [14], [16], and [17] used regression models in their research.

4.2 Results of Experiment

This part represents the results obtained from the experiment. For evaluation of the models used the performance metric that was mentioned in Section 3.2.5. From the literature review part, I find out which ML algorithms were mostly used for prediction tasks. After that, the four Machine learning models chosen for this experiment are:

- Linear Regression
- Polynomial Regression
- Random Forest
- Decision Tree

The above mentioned algorithms were trained and tested with a dataset that was collected and the results were performed. After, for each model performance metrics are applied, and find out the accuracy of algorithms.
Also, the dataset of Covid-19 cases by region performed data visualization for a better understanding of the coronavirus situation in Kazakhstan.

4.2.1 Data Visualization

In Figure 4-1, we can see the chart that illustrates three cases of Covid-19, which are confirmed, recovered, and death, and their increase in numbers from 2020 to 2023. Figure 4-2 shows the number of Covid-19 cases in every region. As you can see, the largest cities Astana and Almaty are the most infested places which are predictable in the number of people living there. The third place by widespread of coronavirus took the Karaganda region which is a little bit close to the largest cities by the number of cases. The number of cases in other regions is similar which each other.

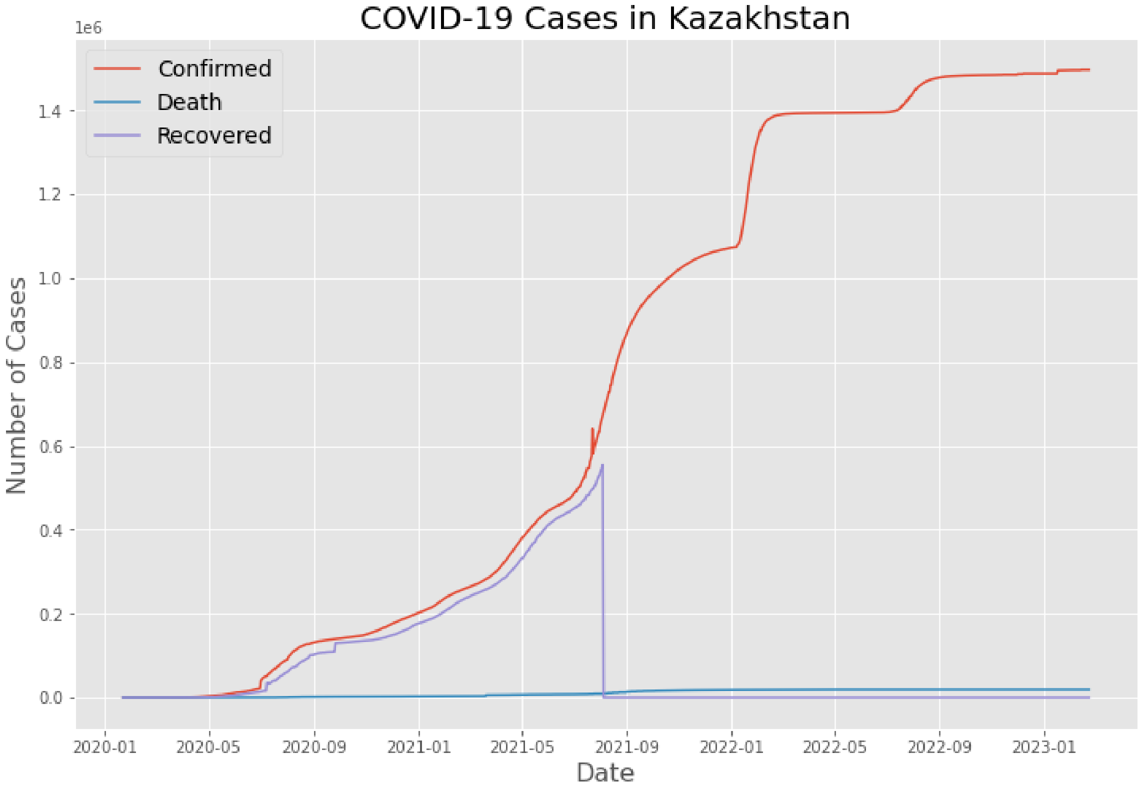


Figure 4-1: Confirmed, recovered, and death cases in Kazakhstan

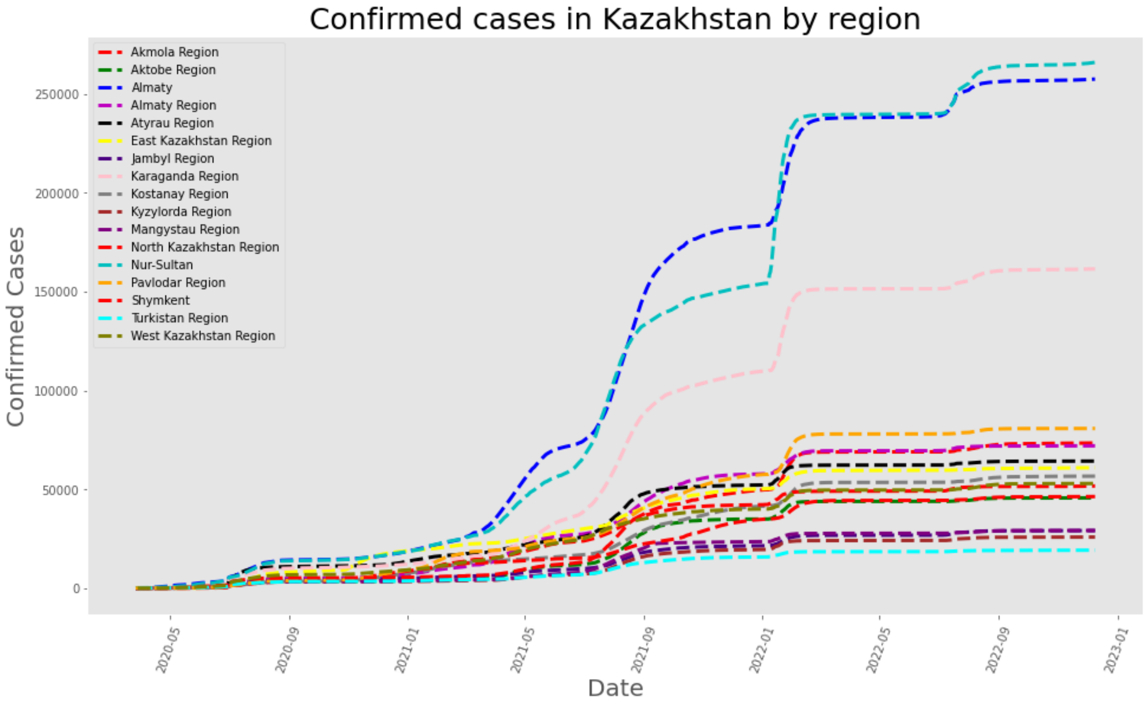


Figure 4-2: Confirmed cases of COVID-19 in Kazakhstan

4.2.2 Machine Learning Algorithms

For predicting the Covid-19 data, firstly, imported libraries for visualization, data manipulation, performance evaluation, and LR, PR, DT, and RF algorithms. After, the dataset was reshaped to a two-dimensional numpy array for model analysis. Then data was split into training and testing datasets with 70% and 30% respectively. The next part was to create the model and fit it to the training data. After the training model is used for predicting the cases for the testing data. The last part was to find out the accuracy of the algorithms by applying performance evaluation. Below you can see the results of these experiments, and the accuracy of confirmed, recovered, and death cases of each ML model. Also, in results represented the prediction of 3 cases in graphical forms.

Linear Regression

The Linear regression model was performed by dividing the dataset into training and testing and forecasting was performed. After, the performance metrics R^2 score, MAE, MSE were used to evaluate the algorithm.

	R2 score	**MAE**	**MSE**
Confirmed	78.23%	617.16864	617.16864
Recovered	44.67%	700.8554	1268209.4
Death	80.40%	4.04464	23.653

Table 4.1: Accuracy of Linear Regression

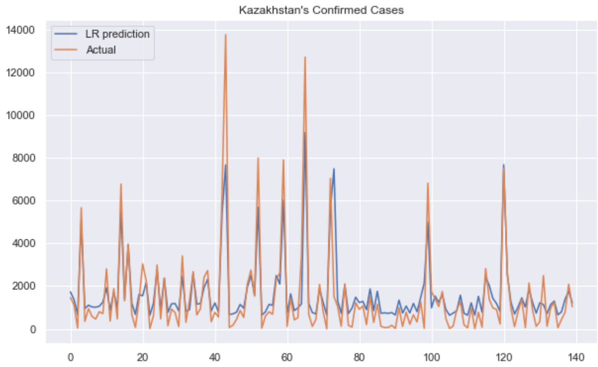


Figure 4-3: LR confirmed cases

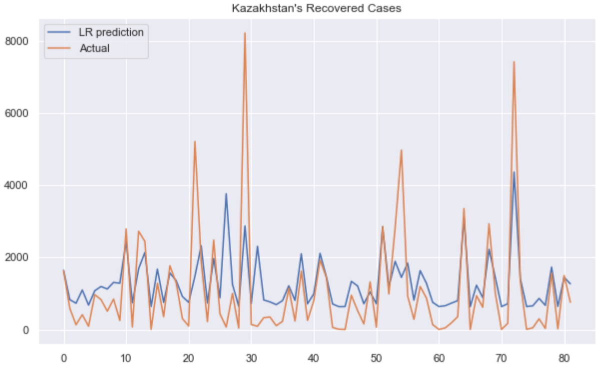


Figure 4-4: LR recovered cases

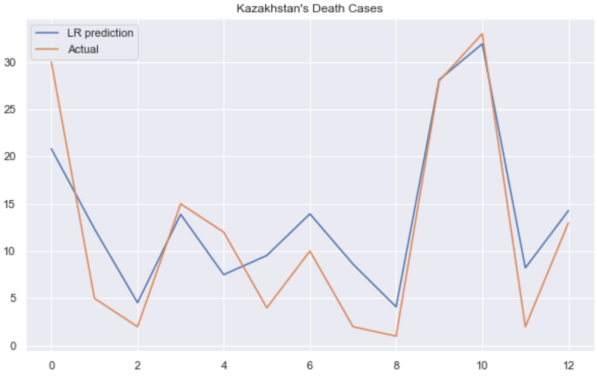


Figure 4-5: LR death cases

Polynomial Regression

The Polynomial Regression model was performed by dividing the dataset into training and testing and forecasting was performed. After, the performance metrics R^2 score, MAE, MSE were used to evaluate the algorithm.

	R2 score	MAE	MSE
Confirmed	75.23%	575.0087	1170918.34
Recovered	28.24%	505.13966	1164299.09
Death	82.51%	4.3394	25.6656

Table 4.2: Accuracy of Polynomial Regression

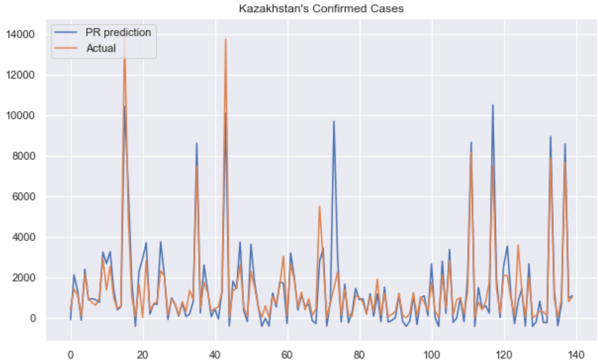


Figure 4-6: PR confirmed cases

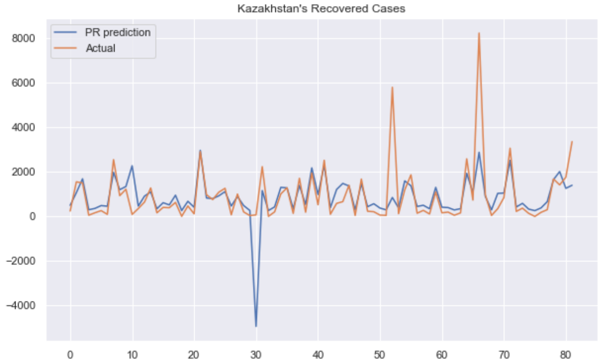


Figure 4-7: PR recovered cases

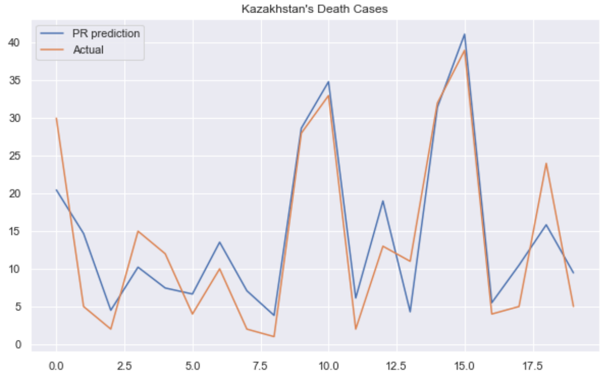


Figure 4-8: PR death cases

Random Forest

The Random Forest model was performed by dividing the dataset into training and testing and forecasting was performed. After, the performance metrics R^2 score, MAE, MSE were used to evaluate the algorithm.

	R2 score	MAE	MSE
Confirmed	88.56%	430.9185	886487.66
Recovered	34.79%	506.2885	958944.84
Death	87.26%	3.5287	18.6492

Table 4.3: Accuracy of Random Forest

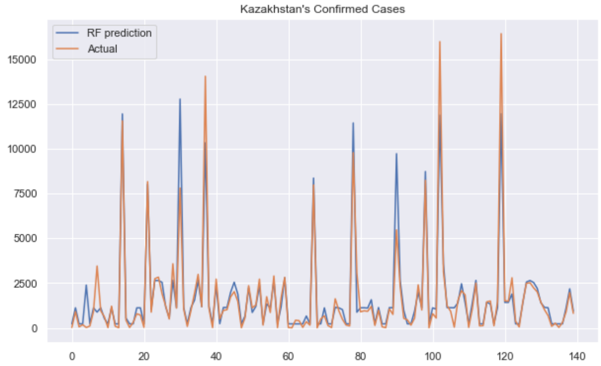


Figure 4-9: RF confirmed cases

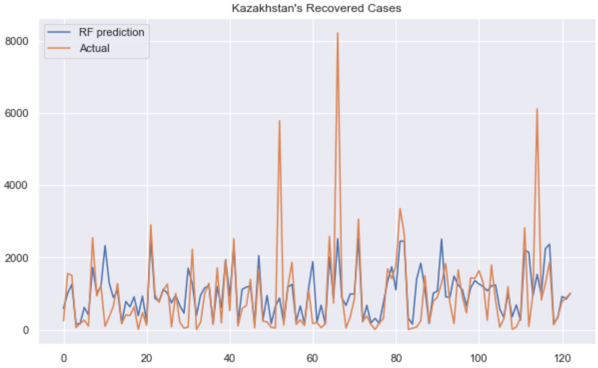


Figure 4-10: RF recovered cases

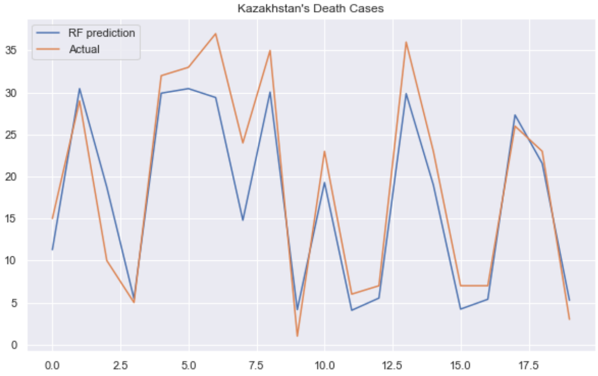


Figure 4-11: RF death cases

Decision Tree

The Decision Tree model was performed by dividing the dataset into training and testing and forecasting was performed. After, the performance metrics R^2 score, MAE, MSE were used to evaluate the algorithm.

	R2 score	**MAE**	**MSE**
Confirmed	90.66%	460.799	723209.36
Recovered	25.76%	566.8943	1091758.69
Death	79.39%	4.5785	31.8276

Table 4.4: Accuracy of Decision Tree

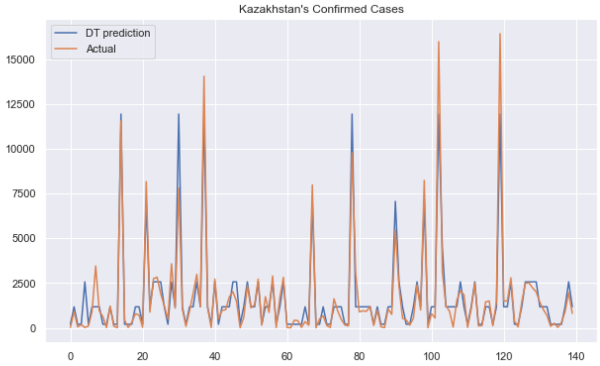


Figure 4-12: DT confirmed cases

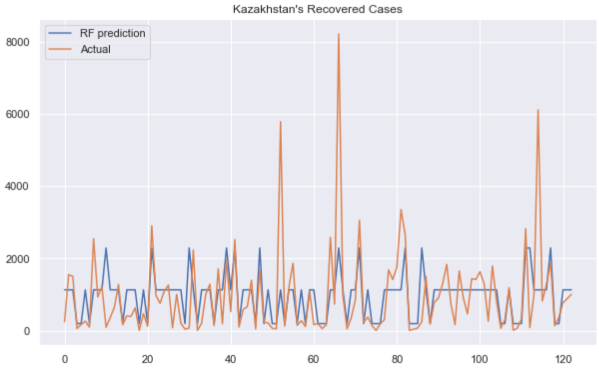


Figure 4-13: DT recovered cases

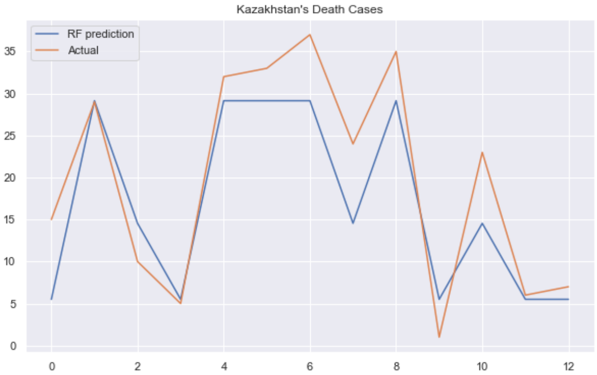


Figure 4-14: DT death cases

4.2.3 Comparison of Results

Based on the experiment, the results of the accuracy are put in Table 4.5 for comparison. In Figure 4-15, you can see the visual representation of the comparison table for a better understanding of which Machine algorithm performed better results. As you can see, because of lacking the dataset related to recovered cases, on recovered cases results showed bad results. However, it was decided to add these results as an example.

	Confirmed	**Recovered**	**Death**
Linear Regression	0.7823	0.4467	0.8040
Polynomial Regression	0.7523	0.2824	0.8251
Random Forest	0.8856	0.3479	0.8726
Decision Tree	0.9066	0.2576	0.7939

Table 4.5: R2 score comparison of ML Algorithms

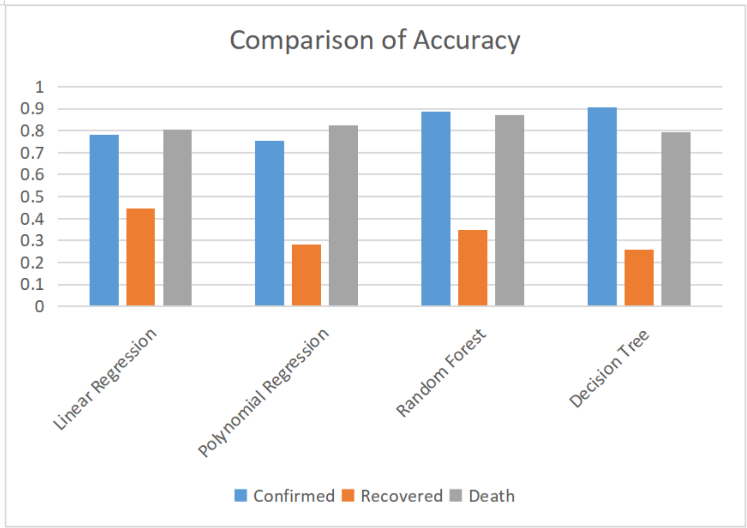


Figure 4-15: Comparison of R2 score

Tables 4.6 and 4.7 show the results of the Mean squared error and Mean absolute error based on the results of the experiment.

	Confirmed	**Recovered**	**Death**
Linear Regression	1032482.56066	1268209.4	23.6531
Polynomial Regression	1170918.34	1164299.09	25.6656
Random Forest	886487.66	958944.84	18.6492
Decision Tree	723209.36	1091758.69	31.8276

Table 4.6: Mean squared error of ML Algorithms

	Confirmed	**Recovered**	**Death**
Linear Regression	617.1686	700.8554	4.0446
Polynomial Regression	575.008	505.1396	4.3394
Random Forest	430.9185	506.2885	3.5287
Decision Tree	460.7992	566.894	4.5785

Table 4.7: Mean absolute error of ML Algorithms

4.3 Discussion

In this part, I will answer to the main Research questions of this thesis work. RQ 1 was answered in the section Systematic Literature Review of Result chapter. RQ 2 was "What kind of prediction methodologies will be used to make an analysis of coronavirus cases?" and the answer to this question is given in the section Results of Experiment of Result chapter. Briefly, the methodologies that were used for prediction are Linear and Polynomial Regression, Random Forest, and Decision Tree. From Table 4.5 and Figure 4-15, we can clearly see that the Random Forest model

outperformed other ML algorithms. However, the Decision Tree model also showed very good results that are almost similar to the Random Forest algorithm.
Research question 3 was "What problems may encounter when predicting the Covid-19 situation?" and here is the answer to this question. While writing this thesis work there were many problems related to the prediction of Covid-19.

- First of all, when collecting the dataset it was hard to find actual daily data with a number of cases.
- Data availability: Most of the datasets do not contain Covid-19 information in Kazakhstan
- Data quality: The recovered cases dataset does not contain any information after 8th May 2021, because of that while training and testing the models, they performed results that are not accurate.
- Because of the little amount of data the prediction model could not be very accurate.
- Model complexity: At the first stages it was hard to understand how some of the Machine Learning algorithms work and what type of evaluation metrics to use.

As you can see above, most of the problems were related to the dataset. It is because I decided to analyze and predict the cases of only Kazakhstan. However, I can clearly say that it was a really good experiment. Also, this research can be helpful for many other spheres, especially healthcare, for a better understanding of the situation in the future.

Chapter 5

Conclusion

The COVID-19 pandemic has been a significant challenge for countries around the world, with many struggling to control the spread of the virus and minimize its impact on public health and the economy. This thesis analyzed COVID-19 data and developed machine-learning algorithms to predict future cases of the coronavirus.

The thesis work has been conducted Systematic Literature Review to find the most suitable Machine Learning algorithm for the prediction of Covid-19 cases. After summarizing and analyzing scientific papers, I came to the decision that Random Forest and Decision Tree models are the most suitable algorithms for prediction. Also, most of the authors used regression models in their research. After understanding and analyzing the other papers related to prediction, Linear and Polynomial Regression, Random Forest, and Decision Tree algorithms were chosen for the experiment. These algorithms were trained and tested by the dataset of confirmed, recovered, and death cases dataset. Then, using performance evaluation metrics, which is R^2 score, mean absolute error, mean squared error, assessed the accuracy of each model. As a result, Random Forest and Decision Tree showed the highest accuracy among other algorithms.

The results suggest that machine learning algorithms can be a valuable tool in forecasting COVID-19 cases and can help policymakers make informed decisions about public health interventions and resource allocation. However, it is important to note that these algorithms are only as accurate as the data they are based on, and there-

fore, improving data quality and completeness is crucial for their effectiveness.
Overall, this thesis demonstrates the potential of machine learning algorithms in analyzing and predicting COVID-19 cases and provides valuable insights for future research and public health efforts.

For future work, there are several areas that could be explored to build upon the findings of this thesis and further improve the accuracy of COVID-19 predictions using machine learning algorithms.
Firstly, it may be beneficial to explore the use of more advanced machine learning techniques, such as deep learning and neural networks, which may be better suited to handling complex and nonlinear relationships in the data.
Secondly, it may be worthwhile to investigate the use of ensemble methods, which combine multiple machine learning models to improve prediction accuracy.
Finally, it may be valuable to incorporate data from other countries and regions to develop a more comprehensive understanding of COVID-19 transmission patterns and improve the generalizability of the machine learning models.

Bibliography

[1] A. Panthakkan, S. M. Anzar, S. A. Mansoori and H. A. Ahmad, "Accurate Prediction of COVID-19 (+) Using AI Deep VGG16 Model," 2020 3rd International Conference on Signal Processing and Information Security (ICSPIS), DUBAI, United Arab Emirates, 2020, pp. 1-4, doi: 10.1109/ICSPIS51252.2020.9340145.

[2] E. Elbasi, S. Mathew, A. E. Topcu and W. Abdelbaki, "A Survey on Machine Learning and Internet of Things for COVID-19," 2021 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2021, pp. 0115-0120, doi: 10.1109/AIIoT52608.2021.9454241.

[3] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 989-995, doi: 10.1109/Confluence51648.2021.9377137.

[4] A. Saini, A. S. Kumar, S. J. N. Kumar and M. U, "Analysis And Implementation of a Novel AI-Based Hybrid Model for Detecting, Predicting and Identification Of COVID-19 Spread," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 2059-2064, doi: 10.1109/ICAC3N53548.2021.9725778.

[5] A. Chitturi, R. J. Pandya and S. Iyer, "AI-Enabled Covid-19 Prediction Methods and Anti-Covid Strategies," 2022 IEEE International Conference on Distributed

Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2022, pp. 1-5, doi: 10.1109/ICDCECE53908.2022.9792761.

[6] S. Bodapati, H. Bandarupally and M. Trupthi, "COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks," 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 525-530, doi: 10.1109/ICCCA49541.2020.9250863.

[7] 1.K. S. Tejaswi, K. Meghavarshini and P. Nivedhitha, "Autism Prediction using ML Algorithms," 2022 1st International Conference on Computational Science and Technology (ICCST), CHENNAI, India, 2022, pp. 1-6, doi: 10.1109/ICCST55948.2022.10040300.

[8] 2.S. Sinha and S. R, "An Educational based Intelligent Student Stress Prediction using ML," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-7, doi: 10.1109/INCET54531.2022.9824636.

[9] 3.K. Kansal and S. Maitrey, "Predicting Covid-19 by Referring Three Supervised ML Algorithms: A Comparative Study using WEKA," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 649-655, doi: 10.1109/ICAAIC53929.2022.9792998.

[10] M. Omkar and K. Nimala, "Machine Learning based Diabetes Prediction using with AWS cloud," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-7, doi: 10.1109/ICSES55317.2022.9914160.

[11] 5.K. Boussaoud, M. Ayache and A. En-Nouaary, "Performance Evaluation of Supervised ML Algorithms for Elephant Flow Detection in SDN," 2022 8th International Conference on Optimization and Applications (ICOA), Genoa, Italy, 2022, pp. 1-6, doi: 10.1109/ICOA55659.2022.9934652.

[12] 7.A. Saini, A. S. Kumar, S. J. N. Kumar and M. U, "Analysis And Implementation of a Novel AI-Based Hybrid Model for Detecting, Predicting and Identification Of COVID-19 Spread," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 2059-2064, doi: 10.1109/ICAC3N53548.2021.9725778.

[13] 10.Nikhil, A. Saini, S. Panday and N. Gupta, "Polynomial Based Linear Regression Model to Predict COVID-19 Cases," 2021 International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), Bangalore, India, 2021, pp. 66-69, doi: 10.1109/RTEICT52294.2021.9574032.

[14] 11.Y. Dubey, P. Wankhede, T. Borkar, A. Borkar and K. Mitra, "Diabetes Prediction and Classification using Machine Learning Algorithms," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 60-63, doi: 10.1109/BECITHCON54710.2021.9893653.

[15] 12.X. Song, "Prediction of People's Abnormal Behaviors Based on Machine Learning Algorithms," 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Guangzhou, China, 2022, pp. 406-409, doi: 10.1109/MLISE57402.2022.00087.

[16] 13.S. Bhardwaj, H. Bhardwaj, J. Bhardwaj and P. Gupta, "Global Prediction of COVID-19 Cases and Deaths using Machine Learning," 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 422-426, doi: 10.1109/ICIIP53038.2021.9702560.

[17] 14.J. Varshitha, K. Jahnavi and C. Lakshmi, "Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-4, doi: 10.1109/ICCCI54379.2022.9740817.

[18] 15.N. Mohan, V. Jain and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," 2021 5th International Conference on

Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-3, doi: 10.1109/ISCON52037.2021.9702314.