Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Machine learning based Comprehensive Study for Stock Market Prediction of Pharmaceutical Industry Index on Covid 19

Arash Salehpour[1], Karim Samadzamini[2]

[1]Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Gilan, Iran

[2]Department of Computer Engineering, University College of Nabi Akram, Tabriz, Iran

Email: [1]arash.salehpour4@gmail.com, [2]samadzamini@ucna.ac.ir

Abstract

This research examines how COVID-19 vaccinations impact the accuracy of machine-learning models in forecasting the Tehran Stock Exchange's Pharmaceutical Companies Index. The study analyses daily vaccination and stock data during the pandemic using statistical and linear regression models. Results reveal a negative correlation between vaccinations and the stock index. Two regression models were developed, one with vaccination data and one without. Although both models fit the training data well, the latter performed significantly better on the test set with lower errors. This suggests that vaccination data does not enhance the predictive ability of the regression model for the stock index during the pandemic. In fact, excluding vaccination data leads to better predictive performance. Therefore, accelerating vaccination programs could aid in the stock market recovery. However, avoiding vaccination data as an input feature for machine learning models forecasting this pharmaceutical stock index is advisable.

Keywords: Covid 19, Stock market prediction, pharmaceutical industry index, Machine learning

1. Introduction

During volatile periods, investors and policymakers must predict stock market indices accurately. However, existing machine learning models for stock index prediction have limitations, such as overfitting and poor performance on out-

DOI: https://doi.org/10.36548/jaicn.2023.2.007
Received: 02.05.2023, received in revised form: 02.06.2023, accepted: 17.06.2023, published: 28.06.2023

of-sample test data. In order to improve these models, some have suggested incorporating additional explanatory variables, such as COVID-19 vaccination data. However, how vaccination data affect the predictive ability of machine learning algorithms for stock market prediction is still being determined. This study aims to investigate the impact of including COVID-19 vaccination data as an input feature on the performance of linear regression models for predicting the Pharmaceutical Companies Index of the Tehran Stock Exchange during the COVID-19 pandemic. The study has three objectives: 1) analysing the correlation between vaccination data and the index, 2) comparing the performance of regression models with and without vaccination data as a feature, and 3) determining if incorporating vaccination data improves the predictive ability and out-of-sample accuracy of the models. Data on daily stock market activity and COVID-19 vaccination counts during the pandemic were collected and analysed. Two linear regression models were trained, one with total vaccinations as a feature and one without it. The models were evaluated and compared based on their training and test sets' performance This study aims to investigate whether incorporating vaccination data into machine learning models can improve their accuracy in predicting stock market indices during the COVID-19 pandemic. The stock market plays a crucial role in the economy, allowing companies to obtain capital and investors through ownership distribution and income potential. Due to the spread of the Coronavirus,the markets worldwide, including the stock market, experienced fluctuations. Sectors such as transportation and tourism were heavily affected, resulting in losses for shareholders. However, controlling the spread of the Coronavirus through vaccination can increase production and sales activities, ultimately helping stabilize prices. The study focus on the Pharmaceutical Companies Index to study the impact of vaccination over the healthcare industry. The healthcare sector's importance is evident, and governments have always aimed to expand the healthcare network in their country. With the Coronavirus pandemic, many people have sought different medicines and insurance, resulting in increased demand. However, as the spread of the virus decreases, these industries may become less prosperous. The drug production industry is an example of progress in the healthcare field, separating leading countries from others and contributing to economic growth. Companies involved in drug production and distribution or investing in this area are classified

in this group. The study conducts an exploratory data analysis, statistical analysis, and machine learning regression analysis to study the effects of vaccination on this group daily [1-4]. Exploratory Data Analysis (EDA) involves collecting and summarizing available data using different analysis methods. By examining the data visually and through statistical reports, what is happening in the desired data set and what to expect is understood. This knowledge helps in choosing appropriate data mining and analysis methods and have reasonable expectations. EDA is any method that does not involve formal and inferential statistical modelling and can be used to manage noise and deviations in the data. It aims to help people look at data before making assumptions and find mistakes, patterns, outliers, strange events, and connections between variables. So the data can be more efficiently analysed due theadvancements in artificial intelligence and machine learning, particularly in evolutionary computing. This technology has enabled us to understand underlying patterns in financial markets. Researchers in economics, statistics, and finance have long been interested in developing and testing behavioural models of stock prices. Machine learning has become increasingly crucial in predicting stock prices. However, forecasting stock price changes in financial time series can be challenging. Correct predictions can lead to significant profits for investors. Nevertheless, the complexity of stock market data makes this task difficult [5-7]. Developing efficient forecasting models requires effort. Linear regression is a statistical technique to identify the connection between dependent and independent variables. This method is helpful in supervised machine learning and is typically applied to continuous data. Linear regression aims to determine the most appropriate line that fits the given data points. The objective is to find the optimal linear equation that predicts the desired output for new data [3, 5, 8].. In many fields of scientific investigation, the changes in experimental measurements of a variable are mainly due to other dependent variables whose values change during the experiment. Therefore, by entering these variables into the statistical analysis the nature of the relationship between these variables is learnt to obtain more accurate results. From another point of view, examining the relationship between variables is essential, which means that the value of one variable can be predicted or even controlled from the observations of other variables and optimized by manipulating the influencing factors. Regression analysis is the central part of statistical methods

for formulating mathematical models that determine the form of relationship between variables. Regression analysis deals with predictions and other statistical inferences using these modelled relationships. The study of relationships between variables is typical in many fields of scientific activity. Simple linear regression is a type of regression analysis in which the number of independent variables is one, and there is a linear relationship between the independent variable (x) and the dependent variable (y) [9, 10]. In addition to simple linear regression, there is multivariate linear regression, which involves more than one independent variable. Machine learning is a type of artificial intelligence that enables systems to learn and improve on their own without human programming. It involves creating computer programs that can access data and use it to learn. Learning typically begins with observations or data, such as examples or experiences, intending to identify patterns in the data and use those patterns to make better decisions. The ultimate goal is for the computer to learn independently and adjust its actions based on learning. There are numerous machine learning algorithms, with new ones being constantly developed. They are generally categorized based on how they learn, such as supervised, unsupervised, or semi-supervised, as well as by their shape and purpose. A supervised machine learning algorithm is capable of using previously learned data, as well as newly labelled data, to predict future outcomes. The learning algorithm creates an inference function that predicts output values by analysing the training data set. With enough training, this system can use new data to identify a target. Additionally, this learning algorithm can compare its results with predetermined, correct results to identify any errors and refine the model. Fig 1. demonstrate Flowchart for the stages involved in the study.

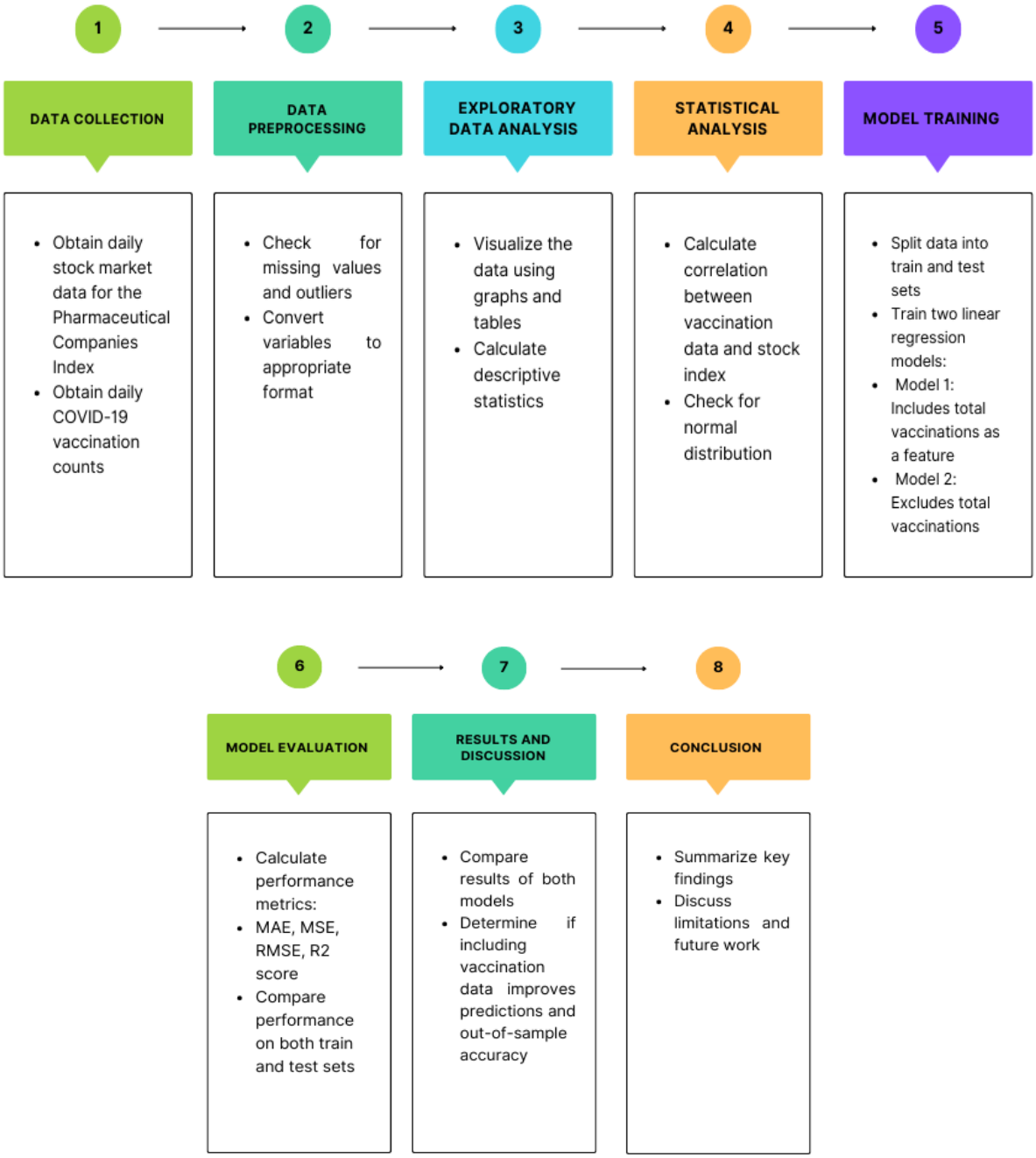

Figure 1. Flowchart for the Stages Involved in the Study

2. Data

This section describes the data that was utilized in the study. This study used daily data on the stock market's closing prices. The study period was from 2021-09-04 to 2022-08-09, a total of 339 days.

2.1 Data Description

Fig. 2 describes the data that was utilized in the study. The "Total" column is for the total number of vaccinations, which includes the stock market data and vaccination. Figure 2 depicts the total number of vaccinated people prior to that time.

	Date	Open	High	Low	Vol	Close	Total
0	2021-09-04 00:00:00	133741.000000	134622.000000	133741.000000	155420721	134601.000000	29002000
1	2021-09-05 00:00:00	134978.000000	135828.000000	134978.000000	149709387	135828.000000	29152527
2	2021-09-06 00:00:00	136969.000000	137773.000000	136969.000000	156763657	137309.000000	29885245
3	2021-09-07 00:00:00	137183.000000	137183.000000	135269.000000	155029577	135269.000000	30667969
4	2021-09-08 00:00:00	135247.000000	135247.000000	134017.000000	127098095	134018.000000	31450694

Figure 2. Sample of Pharmaceutical Companies Index

2.2 OHLC chart

The chart is shown in Figures 3 was created in Jupyter Notebook in Python with the Plotly library in Windows using the data set's open, high, low, close, and volume prices.

Figure 3. OHLC Chart

2.3 Heat Map

Data visualization is a graphical presentation of data, the primary purpose of which is to optimally convey information to users by displaying the relationships between data with the help of charts. Matplotlib is one of the famous Python libraries for drawing graphs. Seaborn is another Python library for a visualization based on Matplotlib and provides more possibilities

for users to draw graphs. These are among Python's most widely used libraries for drawing graphs and data visualization Fig.4 A heat map that shows the correlation matrix from the graph shows that there is not much of a link between the total number of vaccinations and the closing price of the stock market. Empirical results show the total correlation with close is -0.62; the effect of this correlation on the machine learning algorithm will be studied.

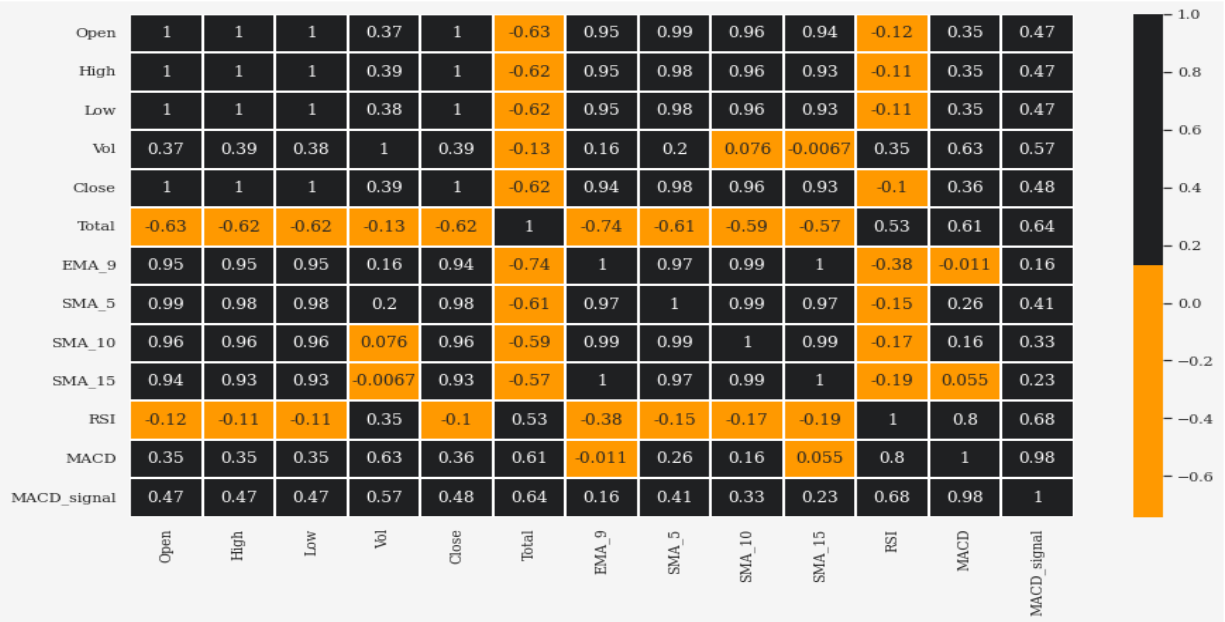

Figure 4. Heat Map Annotate True

2.4 Quantitative Description Statistics

Statistics describes the relationship between two numbers. As illustrated in Figure 5. Descriptive statistics is the set of methods used to organize, summarize, and describe information. Specific preliminary steps must be taken before analyzing the data. When confronted with a large volume of quantitative data prepared for research, the researcher must organize and summarize it in an actionable form to make secret points directly before proceeding to statistical tests. First, it deals with the exploratory examination of the data. Descriptive statistics is the arrangement and classification of data, graphic representation, and subject matter such as “family,” “middle,” etc., indicating a family’s characteristics in the society under discussion. Therefore, descriptive statistical methods are used for this purpose. In general, three methods are used in descriptive statistics to summarize data: using tables and graphs and calculating specific values that indicate essential characteristics of the data.

	Open	High	Low	Vol	Close	Total
count	225.000000	225.000000	225.000000	225.000000	225.000000	225.000000
mean	106182.306222	106492.722667	105708.387111	69887452.040000	106004.443556	122596611.893333
std	11761.137750	11788.843619	11595.924215	34818238.597463	11631.504671	40132125.901822
min	86467.400000	86559.800000	86344.300000	22409809.000000	86345.600000	29002000.000000
25%	98080.100000	98342.400000	97572.700000	45159439.000000	98003.500000	103582369.000000
50%	108051.000000	108349.000000	107594.000000	59158983.000000	107978.000000	137334302.000000
75%	112733.000000	113041.000000	112220.000000	88259662.000000	112772.000000	149515460.000000
max	137183.000000	137773.000000	136969.000000	262332788.000000	137309.000000	414526830.000000

Figure 5. Figure 8Quantitative Description Statistics

2.5 Pair Plot

To predict the stock's closing price; let us consider the "Close" feature as the target variable. Figure 6 depicts the pair plot.

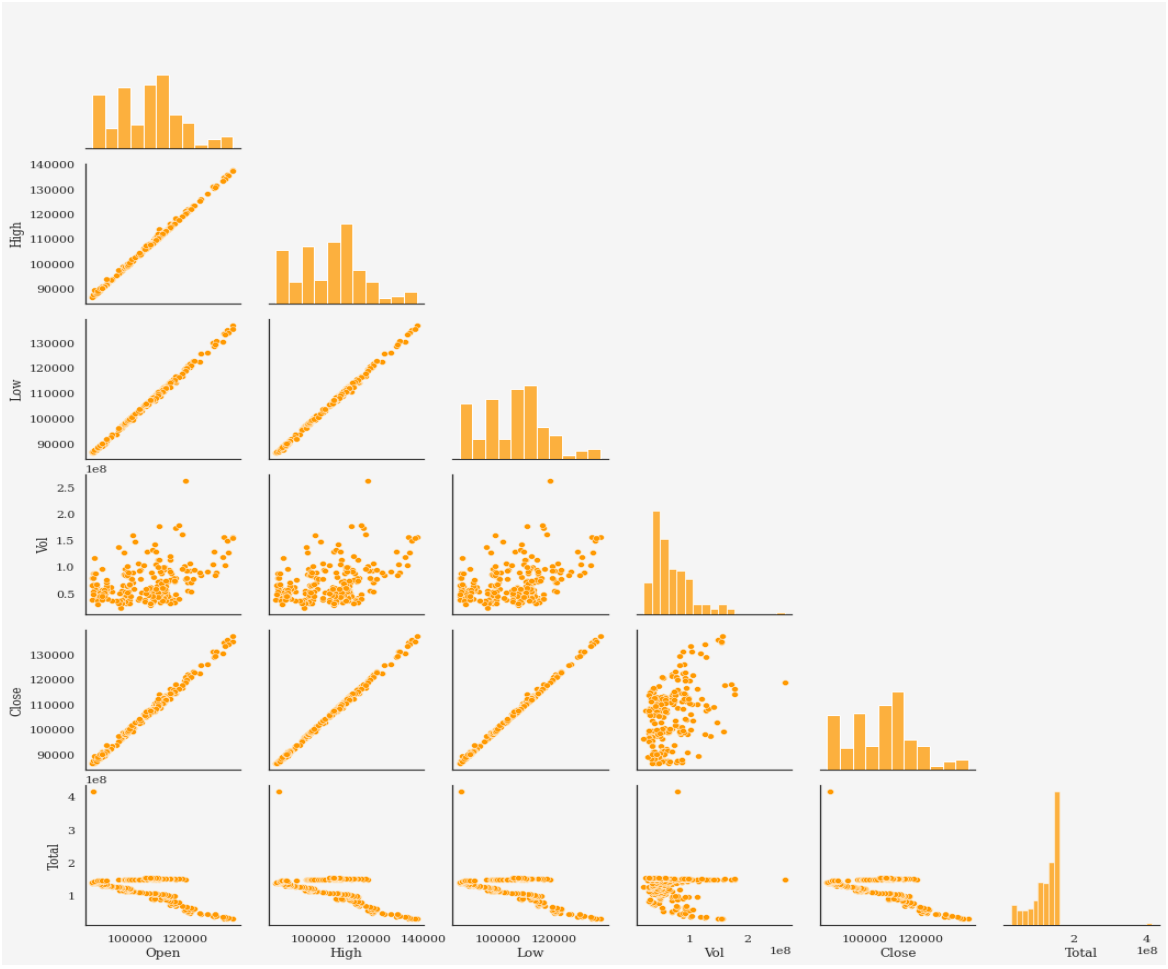

Figure 6. Pair Plot

2.6 Normal Distribution

After performing any measurement, in order to discover the relationship between the data's or categorize the data's are analyzed. First, the distribution of the data is learnt. In layperson's terms, data distribution tells us how widely the data collection is distributed the. The normal distribution is also one of these patterns. This continuous probability distributions is one of the most important probability theory. The reason for its name and the importance of this distribution is that many values obtained during natural and physical fluctuations around a fixed value matches with the values obtained from this distribution[11-13]. Fig 7. Depict Normal distribution on close.

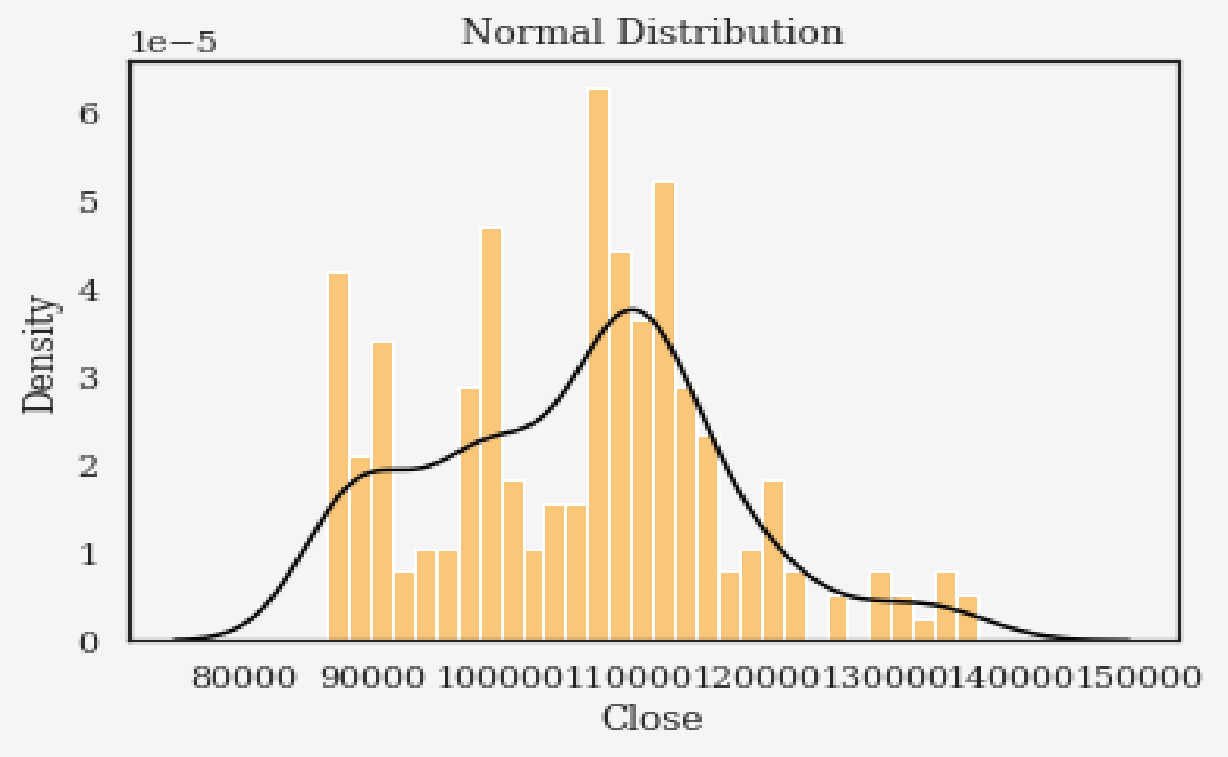

Figure 7. Normal Distribution on Close

2.7 Bootstrap Plot

One of the powerful computer methods in statistical inference is bootstrapping. This method tries to estimate the error of the estimators without considering the complex conditions. For example, the bootstrap method in machine learning can create confidence intervals and regression models and estimate model error. In statistical topics, the bootstrap technique is used to estimate the parameters of the statistical population. the Python programming language is used to implement the calculations related to this technique. As mentioned, the purpose of the bootstrap application method is to make inferences about the parameter estimator of the

statistical population. The bootstrap method tries to obtain the best estimate for the error of the estimators with the help of the sampling technique and replacement resampling according to the limited sample size. Of course, the samples obtained from resampling are independent of each other. The main reason for using the bootstrap technique is to take advantage of the speed of modern computers when performing statistical inference. "Efren" invented this method to determine the accuracy of estimators in statistical analysis problems. As a general principle, after estimating the parameter by the estimator, the amount of estimation error should also be determined, which is the same as the estimator's accuracy.

Table 1. Bootstrap Sample

Bootstrap Sample:	
37	124524.0
140	97991.5
72	98406.1
137	90180.4
203	108329.0

Name: Close, dtype: float64

Mean of the population: 106004.44355555554

Standard Deviation of the population: 11631.504670842316

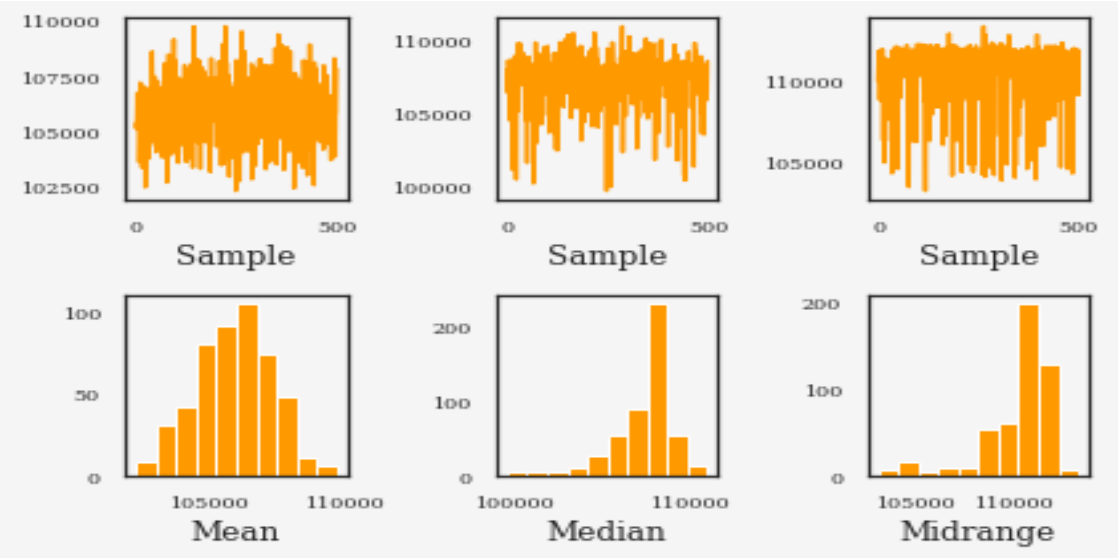

Figure 8. Bootstrap Plot

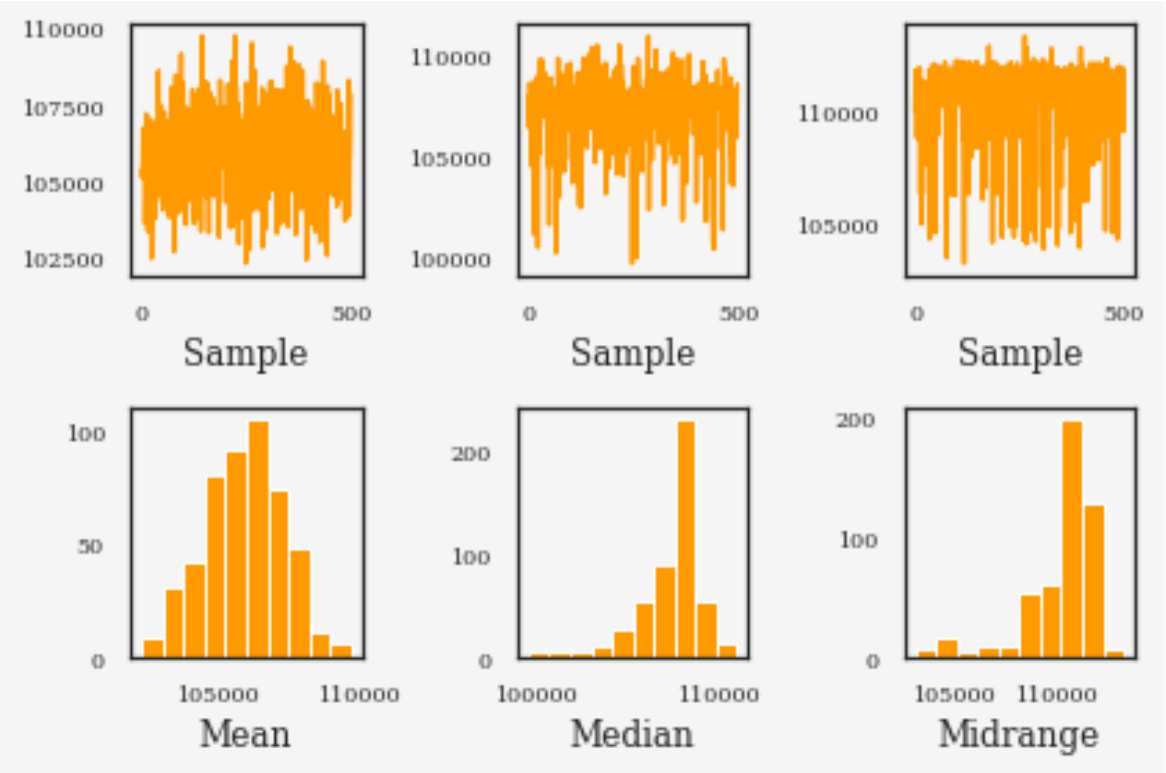

Figure 9. Bootstrap Plot

2.8 EDA, Month wise Comparison Between Pharmaceutical Companies Index Actual, Open and Close Price

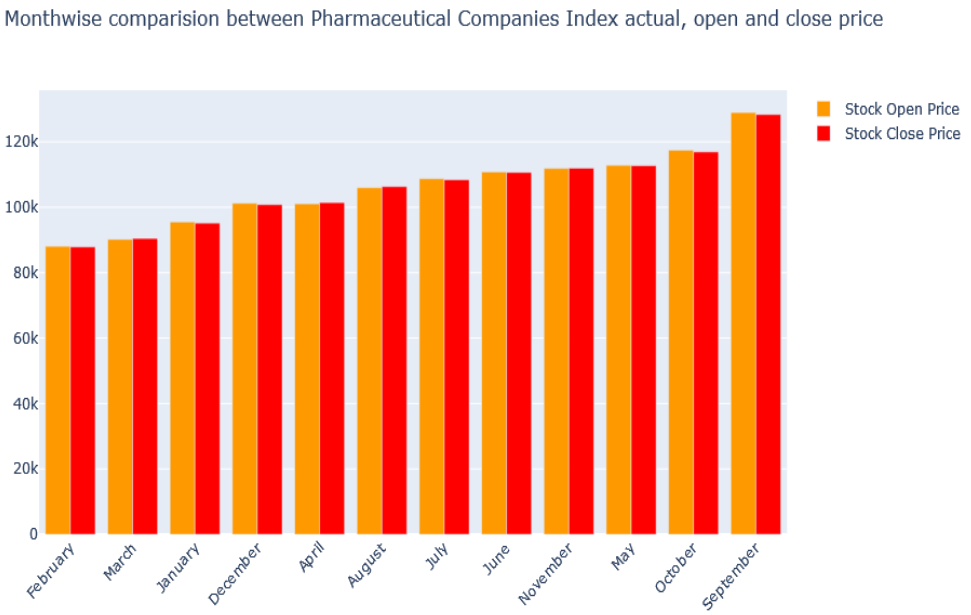

Figure 10. Month Wise Comparison Between Pharmaceutical Companies Index Actual, Open and Close Price

2.9 EDA, Month wise High and Low Stock Price

Figure 11. Month wise High and Low Stock Price

2.10 Statistics

Skewness and Kurtosis

Skewness In the probability theory and statistics, skewness indicates the degree of asymmetry of the probability distribution of data around its mean. The skewness value can be negative or positive. The degree of the tendency of the probability distribution curve of a data series is skewness, but this measure indicates asymmetry in the curve's tails. If the data have a symmetrical distribution, the elongation of the right and left tails is the same. Kurtosis, is a statistical parameter that describes the probability distribution of a random variable; Kurtosis reveals how much information is included in the tails. it shows the a measurement of the total weight of a distribution's tails in relation to the mean (or average). This is also known as the peak degree. Table 2 depicts skewness and kurtosis.

Table 2. Skewness and Kurtosis

Skew	0.22157761608718327
kurt	-0.2316526067981024

Measures of Spread

The spread and dispersion measures the similarity /variations in the data. Measures of centre tell us the average value, while measures of spread tell us how far the data tends to be from the average value. Distance between the maximum and minimum observations are stated as : **Max – min** = 50963 on Close.

Table 3. Measures of Spread

Quantile (0)	86345.6
quantile (0.25)	98003.5
Quantile (0.50)	107978
Quantile (0.75)	112772
Quantile (1)	137309

Data set Describe

```
count       225.000000
mean     106004.443556
std       11631.504671
min       86345.600000
25%       98003.500000
50%      107978.000000
75%      112772.000000
max      137309.000000
Name: Close, dtype: float64
```

Figure 12. Data Set Describe

3. Technical Indicators

Traders use various tools to find suitable trading opportunities; technical indicators are one of the most critical technical analysis tools. It is a mathematical calculation of price, time, and volume. By using them, the market trend can be understood better [14].

3.1 Moving Averages

In statistics, a moving average is a calculation made to examine data points using averages generated from subsets of the actual price data. In financial sciences, the moving average (MA) is a stock indicator usually used in technical analysis. Calculating the moving average of a share is to adjust the price information created by calculating an average price. On the close

price, we used an exponential moving average of 9, a simple moving average of 5, a moving average of 10, and a moving average of 15. Fig 13 shows moving averages.

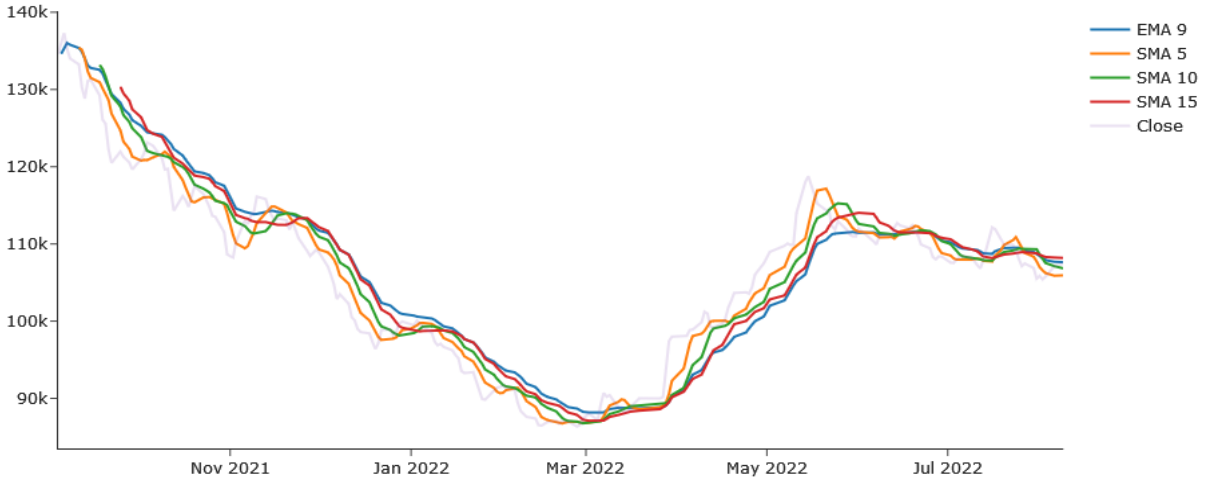

Figure 13. Moving Averages

3.2 Relative Strenght Index

Among the most powerful and effective methods for forecasting the market and profiting from digital asset price fluctuations Among the various tools that technical traders often use for market analysis, the Relative Strength Index (RSI) is one of the most popular indicators. The RSI indicator fluctuates between two specific intervals. It can show the trader the overbought and oversold areas, evaluate the acceleration of price changes, and show the trader the best points to enter and exit the market. As n = 14, we use rsi deficiencies. Figure 14 describe RSI chart in Data Set.

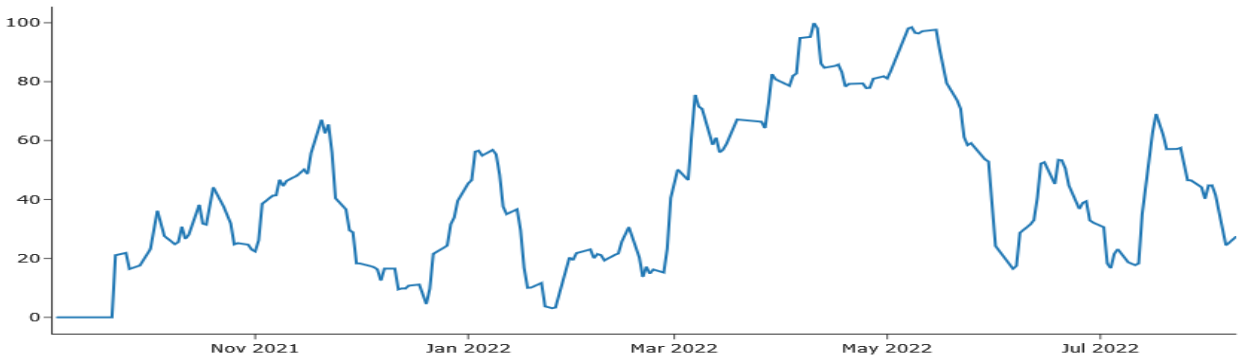

Figure 14. Relative Strength Index

3.3 Moving Average Convergence/Divergence

The MACD indicator is one of the most popular oscillators and is considered one of the leading indicators among technical analysis traders. MACD is an abbreviation for Moving Average Convergence/Divergence. This indicator includes three moving averages. By using it, investors can identify the trend's strength, direction, and acceleration of the downward or upward trend in the stock market. The moving averages used in the MACD indicator include the 9, 12, and 26-day moving averages. The MACD chart in the data set is depicted in Fig. 18. MACD=(12-Period EMA)-(12-PeriodEMA).Fig 15 depicts Moving average convergence/divergence

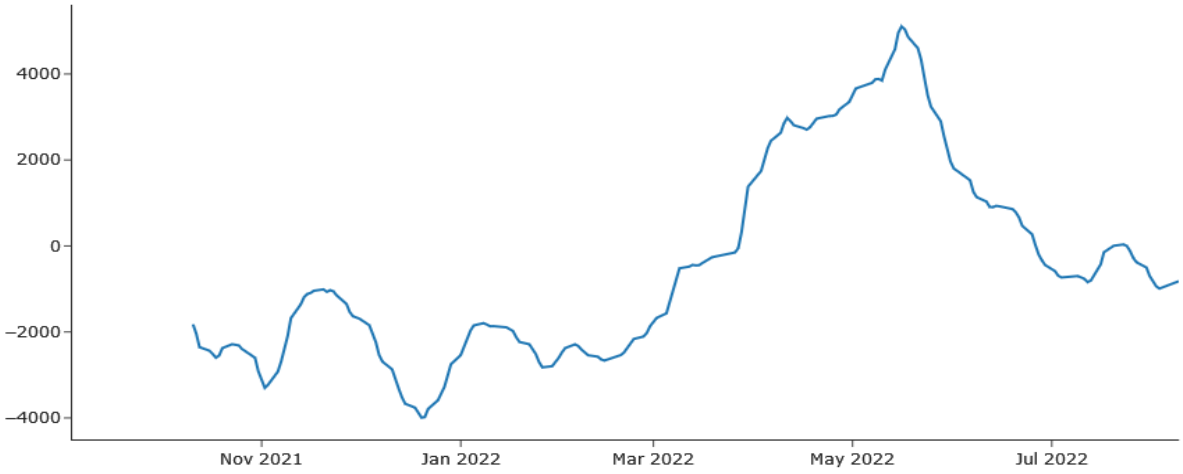

Figure 15. Moving Average Convergence/Divergence

4. Machine Learning Base

Linear regression is considered the first machine-learning algorithm. Linear regression in machine learning is considered a supervised learning model aiming to find the best-fit line between independent and dependent variables. In other words, this model creates a linear relationship between independent and dependent variables[15].

4.1 Linear Regression with Total Vaccination as One of the Features

In this section, we use the vaccination data as one of the features and check whether the use of vaccination data helps to improve the prediction.

X Train Features

We have used Train test split from the Sklearn library with test_size = 0.2 and shuffle = False As depicted in Fig. 18; the feature column is "Volume," "Total,"

"EMA_9," "SMA_5", "SMA_10", "SMA_15", "RSI," "MACD," "MACD_Signal." Fig. 11 depicts the x-train head based on Pandas' data frame.

	Vol	Total	EMA_9	SMA_5	SMA_10	SMA_15	RSI	MACD	MACD_signal
44	70452769	95327143	113888.107900	110510.600000	111468.000000	113002.133333	46.732206	-2392.439087	-2756.825635
45	93977063	95669900	113858.236392	111548.400000	111347.000000	112865.400000	44.532769	-2094.439973	-2623.115242
46	112861120	96752585	113882.604166	112719.200000	111393.400000	112819.933333	46.316489	-1679.584473	-2433.006332

Figure 16. X Train Features

Predictions from Model

Figure 17. shows predictions from the test set.

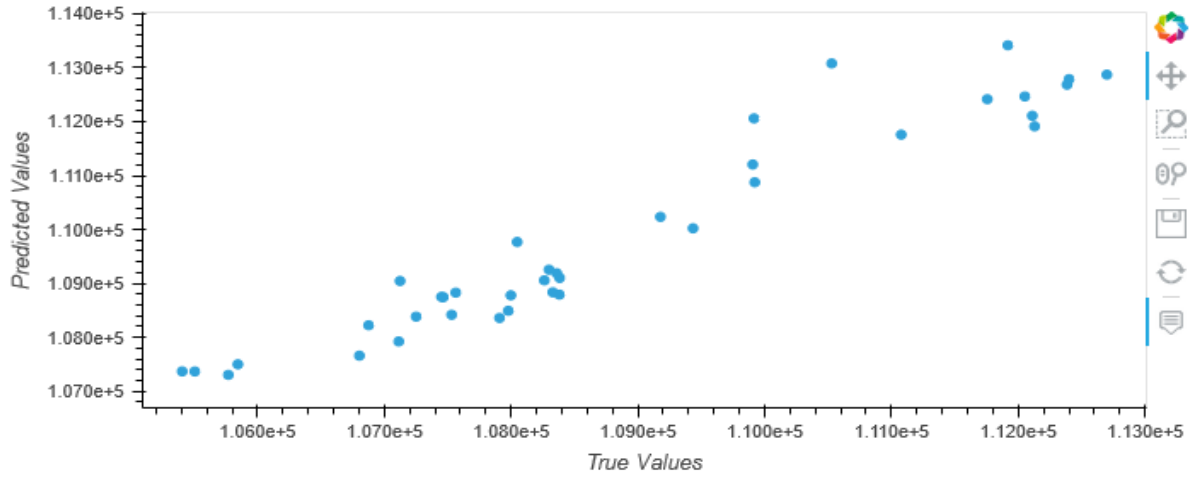

Figure 17. Predictions from Model

Actuals and Forecasts based on Linear Regression with Vaccine as A Feature

Figure 18. Actual and Predictions based on Linear Regression Chart

4.2 Linear Regression without Total Vaccination as a Feature

X Train Features

We used the Train test split from the Sklearn library with test_size = 0.2 and shuffle = False. The feature columns are "Volume," "EMA_9," "SMA_5", "SMA_10," "SMA_15," "RSI," "MACD," and "MACD_Signal," as shown in Fig. 22. Fig. 19 depicts the x-train head based on Pandas' data frame.

	Vol	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	SMA_50	RSI	MACD	MACD_signal
59	40197182	112342.862600	110593.200000	111788.200000	112801.133333	112901.633333	116428.260000	28.842916	-1637.567113	-1377.333673
60	54314987	111990.242679	109803.400000	111301.800000	112483.200000	112674.300000	115979.580000	18.324896	-1664.269612	-1434.739490
61	47353322	111714.973215	109272.200000	110922.100000	112159.333333	112489.633333	115580.060000	18.277793	-1693.977744	-1486.600604

Figure 19. X Train Features

Predictions from Model

Figure 20 shows Take predictions from the test set.

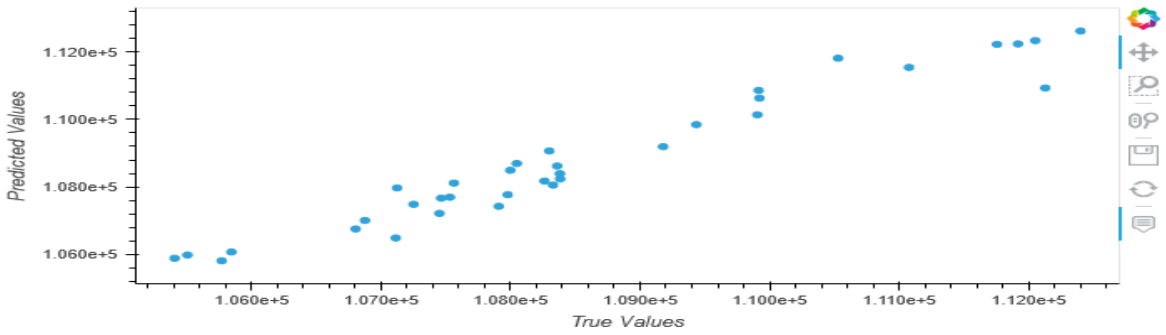

Figure 20. Predictions from Model

Actuals and Predictions Based on Linear Regression without Vaccine as one of the Features

Figure 21. Actual and Predictions based on Linear Regression Chart

5. Empirical Results

In both methods, the models performed well on the training data, and there is no significant difference in the result metric criteria. However, the vaccination data method needed to be more balanced in the test data.

5.1 Test set evaluation

Table 4. Test Set

Test set	with vaccination as a feature	Vaccination is not a feature.
MAE	1011.060751726878	405.98770803124313
MSE	1376848.6223770226	263176.3451282789

RMSE	1173.391930420958	513.0071589444722
R2 Square	**0.6942390590639962**	**0.9293604214373858**

Table 5. Train Set

Test set	**with vaccination as a feature**	**Vaccination is not a feature.**
MAE	386.77779174859324	330.2162791484553
MSE	250229.71795773035	183055.70426947915
RMSE	500.2296652116209	427.8500955585719
R2 Square	**0.9972196506381351**	**0.9978492699440519**

6. Conclusion and Discussion.

This study has found a negative correlation between the total number of vaccinations and the closing price of the Pharmaceutical Companies Index. Essentially, as vaccination counts increase, the stock index tends to decrease. The research also found that excluding vaccination data as an input feature in the linear regression model performed significantly better than including it, especially on the test set. This suggests that vaccination data does not improve the model's predictive ability and may hinder its performance. The model without vaccination data has lower errors and a higher R-squared score on the test set, indicating that excluding vaccination data leads to better out-of-sample predictions for the stock index. Overall, incorporating vaccination data as an input feature does not improve and may worsen the performance of machine learning models for predicting this pharmaceutical stock index during the COVID-19 pandemic. This study provides empirical evidence that excluding vaccination data from the models leads to better out-of-sample predictions.

Additionally, the findings suggest that while speeding up mass vaccination programs may help economic recovery, vaccination data does not provide helpful information for machine learning algorithms to predict stock market indices for pharmaceutical companies during the COVID-19 pandemic When it comes to machine learning models, it is crucial to assess whether adding new variables as features can enhance their predictive accuracy. Factors impacting active stock market companies fall into three categories: macro-environmental,

micro-internal, and industrial. Predicting stock prices is crucial for investors, researchers, financial professionals, and companies involved in the stock market. Accurate predictions can help decision-makers make informed choices. Traditional methods like ARIMA and newer techniques like neural networks are used for stock price forecasting. In this study, the focus was on the impact of vaccination on pharmaceutical stock index prices. Due to the limited availability of vaccination data, time series, and deep learning methods were employed. The findings suggest that vaccination data does not affect the prediction of pharmaceutical stock index prices.

Author contributions: All authors listed have made a substantial, direct, and intellectual contribution to the work and have approved it for publication.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: available on requests from corresponding author

Conflicts of Interest: The authors declare no conflict of interest.

Human and Animal Rights and Informed Consent: This article does not contain any studies with human or animal subjects performed by any of the authors.

Reference

[1] Arash Salehpour, Bibliometric Review of Applications of Deep Learning in Marketing: Advances in Resources and Top Trend Analysis. Journal of Artificial Intelligence and Capsule Networks, 2022. 4(4): p. 230-244.

[2] Ullah, K. and M. Qasim. Google Stock Prices Prediction Using Deep Learning. in 2020 IEEE 10th International Conference on System Engineering and Technology (ICSET). 2020.

[3] Xu, F., et al., Cost-sensitive regression learning on small dataset through intra-cluster product favoured feature selection. Connection Science, 2022. 34(1): p. 104-123.

[4] Wang, Y., Y. Wang, and J. Wang, Efficient self-adaptive access control for personal medical data in emergency setting. International Journal of Computational Science and Engineering, 2020. 23(4): p. 341-351.

[5] Takouk, D., R. Zeghdane, and B. Lakehali, A new approach based on generalised multiquadric and compactly supported radial basis functions for solving two-dimensional Volterra-Fredholm integral equations. International Journal of Computational Science and Engineering, 2022. 25(5): p. 532-547.

[6] Thakkar, A. and K. Chaudhari, A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. Expert Systems with Applications, 2021. 177: p. 114800.

[7] Salehpour, A.S., E., A Regression Analysis on the Car Index in the Tehran Stock Exchange. Journal of Soft Computing Paradigm, 2022. 4(4): p. 238-251.

[8] Sanjeevikumar, P., et al., Chapter 11 - Machine learning-based hybrid demand-side controller for renewable energy management, in Sustainable Developments by Artificial Intelligence and Machine Learning for Renewable Energies, K. Kumar, et al., Editors. 2022, Elsevier. p. 291-307.

[9] Jamil, S., et al. Machine Learning Price Prediction on Green Building Prices. in 2020 IEEE Symposium on Industrial Electronics & Applications (ISIEA). 2020.

[10] Islas-Cota, E., et al., A systematic review of intelligent assistants. Future Generation Computer Systems, 2022. 128: p. 45-62.

[11] Carrera-Rivera, A., F. Larrinaga, and G. Lasa, Context-awareness for the design of Smart-product service systems: Literature review. Computers in Industry, 2022. 142: p. 103730.

[12] Etemadi, M., et al., A systematic review of healthcare recommender systems: Open issues, challenges, and techniques. Expert Systems with Applications, 2023. 213: p. 118823.

[13] Carrera-Rivera, A., et al., How-to conduct a systematic literature review: A quick guide for computer science research. MethodsX, 2022. 9: p. 101895.

[14] RadukiÄ‡, S.a. and M. RadoviÄ‡, Long Term Trend Analysis in the Capital Market â€“ The Case of Serbia. Journal of Central Banking Theory and Practice, 2014. 3(3): p. 5-18.

[15] Song, Y.-G., Y.-L. Zhou, and R.-J. Han, Neural networks for stock price prediction. ArXiv, 2018. abs/1805.11317.